PROBABILITY THEORY

with Applications in Science and Engineering

A Series of Informal Lectures

by

E. T. Jaynes

Professor of Physics

Washington University

St. Louis, Missouri 63130

Fragmentary Edition

of

February, 1974

*The following notes represent what is completed to date of a projected book manuscript. All Lectures after No. 9 are incomplete; Lectures 11 and 12 are missing entirely, although their content is already largely published in E. T. Jaynes, "Prior Probabilities," IEEE Trans. Syst. Sci. and Cybern. SSC-4, Sept. 1968, pp. 227-241; and "The Well-Posed Problem," in Foundations of Physics, 3, 477 (1973). The projected work will contain approximately 30 Lectures; in the meantime, comments are solicited on the present material.

SUMMARY OF BASIC RULES AND NOTATION

Deductive Logic (Boolean Algebra): Denote propositions by A, B, etc., their denials by a ≡ "A is false," etc. Define the logical product and logical sum by

$$AB \equiv \text{"Both A and B are true."}$$

$$A+B \equiv \text{"At least one of the propositions, A, B is true."}$$

Deductive reasoning then consists of applying relations such as $AA = A$; $A(B+C) = AB + AC$; $AB+a = ab+B$; if $D = ab$, then $d = A+B$, etc., in which the = sign denotes equal "truth value."

Inductive Logic (probability theory): This is an extension of deductive logic, describing the reasoning of an idealized being (our "robot"), who represents degrees of plausibility by real numbers:

$$(A|B) = \text{probability of A, given B.}$$

Elementary requirements of common sense and consistency, such as: (a) if a conclusion can be reasoned out in more than one way, every possible way must lead to the same result; and (b) in two problems where the robot has the same state of knowledge, he must assign the same probabilities, then uniquely determine these basic rules of reasoning (Lect. 3):

Rule 1: $(AB|C) = (A|BC)(B|C) = (B|AC)(A|C)$

Rule 2: $(A|B) + (a|B) = 1$

Rule 3: $(A+B|C) = (A|C) + (B|C) - (AB|C)$

Rule 4: If $\{A_1 \ldots A_n\}$ are mutually exclusive and exhaustive, and information B is indifferent to them; i.e., if B gives no preference to one over any other, then

$$(A_i|B) = 1/n, \quad i = 1, 2, \ldots, n$$

Corollaries: From Rule 1 we obtain Bayes' theorem:

$$(A|BC) = (A|C)\,\frac{(B|AC)}{(B|C)}$$

From Rule 3, if $\{A_1 \ldots A_n\}$ are mutually exclusive,

$$(A_1 + \ldots + A_n|B) = \sum_{i=1}^{n} (A_i|B)$$

If in addition the A_i are exhaustive, we obtain the chain rule:

$$(B|C) = \sum_{i=1}^{n} (BA_i|C) = \sum_{i=1}^{n} (B|A_iC)(A_i|C)$$

These are the relations most often used in practical calculations.

(continued on inside back cover)

PREFACE

This book has grown over several years from a nucleus consisting of transcripts of tape recordings of a series of lectures given at the Field Research Laboratories of the Socony-Mobil Oil Company in Dallas, Texas, during March, 1958 and June, 1963. The lectures were given also, with gradually increasing content, at Stanford University in 1958, at the University of Minnesota in 1959, at the University of California, Los Angeles in 1960 and 1961, at Purdue University in 1962, at Dartmouth College in 1962 and 1963, at the Standard Oil Company Research Laboratories, Tulsa, in 1963, at the University of Colorado in 1964, at the University of Maryland in 1968, and at Washington University in 1966, 1969, 1970 and 1972. The material of lectures 1-10 and 16-17 was issued by the Socony-Mobil Oil Company as Number 4 in their series, "Colloquium Lectures in Pure and Applied Science", and is reproduced here, with permission, in considerably expanded form.

In editing and adding new material, the informal style of the original presentation has been retained. This and the general format are intended to emphasize that the book is in no sense a textbook or complete treatise, but only a series of informal conversations (necessarily rather one-sided), concerning the foundations of probability theory and how to use it for current applications in physics, chemistry, and engineering. The speaker is simply sharing his views with the audience, and trying to give some more or less convincing arguments in support of them. Often, the trend of a lecture was determined by questions raised from the audience.

The material is addressed primarily to scientists and engineers who are already familiar with applied mathematics and perhaps with certain special uses of probability theory, such as statistical mechanics, communication theory, or data analysis, but who may not have had the time to make an extensive study of modern statistics. Such persons may be appalled, as I was when I commenced serious study of the field in 1950, by the enormous volume of literature dealing with statistical problems, and may despair of ever mastering it--not because it is too advanced, but simply because the field is too large. There is so much diverse and intricate detail that it is almost impossible to locate the underlying

principles; and one finally succeeds only to have them dissolve in confusion and controversy, no two authors being in agreement about them.

For such persons, we have good news. Recently, a great simplification and unification of this field has become possible. There is a single very simple set of principles, which can be stated in a few lines and which, when applied to specific problems, will be found to give automatically conventional probability theory, the formalism of equilibrium and nonequilibrium statistical mechanics, the results of communication theory, and the newest methods of statistical inference, which represent a great advance over the "orthodox" methods prevalent in the period 1930-1960.

These principles are summarized on the inside front and back covers of this book. Although at present we are able to give only heuristic (but nevertheless convincing) arguments for their uniqueness, there is no difficulty in demonstrating that they do include the aforementioned applications as special cases. Therefore, whatever changes in viewpoint may come in the future, these principles will retain at least their didactic value, as a concise summary of presently known statistical methods.

Current applications have advanced to the point where the perennial confusion surrounding foundations of probability theory now poses a serious threat to further progress. In particular, we have struggled for over two centuries with concentual problems involving the relation between abstract probability theory and the "real" world. Should we use probability only in the sense of describing frequencies in some "random experiment", or is it legitimate to interpret the mathematical rules in the broader Laplace sense of a "calculus of inductive reasoning?" In my opinion, the time has come when such questions must be settled. Until this is done we cannot hope to resolve the paradoxes of quantum theory and irreversible statistical mechanics, or even to justify the use of probability theory in describing time-dependent phenomena.

Because of conceptual difficulties with Laplace's viewpoint, many attempts have been made to evade the general problem of inductive reasoning, and to develop probability theory from more restrictive postulates concerning limiting frequencies (i.e. the von Mises "collective", etc.). This approach encountered such great mathematical and logical difficulties that it has been almost completely abandoned; but strangely enough, the intolerance of broader views of probability has survived. Thus, today most writers on probability and statistics take the curious position of admitting that a probability cannot be defined merely as a frequency, but still insisting that it must always be interpreted as a frequency in applications.

The theory developed here is more general than in conventional expositions because the rules are derived in a way that makes it clear that neither the notion of probability nor the mathematical rules of probability theory depend on such concepts as random variables, random experiments, or relative frequencies.

According to the viewpoint expounded here, consideration of random experiments is only one particular application (and not even the most important one) of probability theory; the rules apply equally well to general inductive inferences where no random experiment is involved in any way. Indeed, most applications of probability theory can be formulated and carried to completion without ever introducing the notion of a "random variable". This is demonstrated repeatedly in the following text, particularly in Lectures 5, 6, 8, 9, 11 and 18.

In our emphasis on the conceptual, rather than the purely mathematical, problems we are necessarily dealing almost constantly with controversial aspects of probability theory. One of the most difficult problems of principle confronting a person trying to apply the theory (treatment of prior information) has been debated vigorously on philosophical grounds for over a century, without being brought perceptibly nearer solution. In this book (particularly, Lectures 10 and 12) we are able to report some progress in reducing these vague philosophical questions to definite and answerable mathematical ones, and in sufficient generality to cover most current applications. However, we make no claim to have fully resolved the situation, about which L. J. Savage has remarked that "there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel." Indeed, we make no claim to have proved anything at all, in a sense which would be accepted as rigorous by modern mathematician. But the arguments given here are, I believe, sufficiently compelling to justify a claim that we have shifted the burden of proof back to those writers who persist in asserting that the Laplace viewpoint is nonsense, and only the strict frequency interpretation is respectable.

The idea of independent repetitions of a random experiment, which the "frequentist" usually considers essential to the very notion of probability, is from our standpoint only a very special case of an exchangeable sequence. We are able to give, in Lectures 16 and 17, a fairly complete discussion of connections between probability and frequency in this case, via straightforward mathematical deduction without any appeal to arbitrary postulates about the relation between them. Similar, and equally general, connections arise in nearly every other application. As a result, we will claim--not as a theorem, but as an inductive generalization to which no exception has been found--that there is never any need to postulate such relations. All the connections between probability and frequency that are actually used in applications, far from conflicting with the Laplace-Bayes "inductive reasoning" form of probability theory, are derivable as elementary mathematical consequences of that theory.

A word of explanation and apology to mathematicians who may happen on this book not written for them; you will find here no appeal to the notions of Borel fields and Radon-Nikodym derivatives, no use of sets or measure theory other than an occasional "almost everywhere" remark, and no Lebesgue-Stieltjes integrals. I am not opposed to these things, and will gladly use and teach them as soon as I find one specific real application where they are needed. Never having uncovered such a problem, either in my own work or in all the statistical literature known to me, but knowing that their introduction would discourage many from reading this book, I have decided to forego them. From the standpoint of probability theory as I see it, they add little rigor to the subject, but serve rather to generalize and extend it in a direction different from the one we are traveling. We get along happily and without impediment using Riemannian integrals with integrands interpreted, when convenient, in the sense of generalized functions. It is well established that, in fourier analysis, this procedure is actually more powerful and appropriate to the subject than the measure-theoretic approach. I think a good case can be made for the view that this holds also in probability theory.

No author can hope to give proper acknowledgement to all those who have influenced his thinking; the list would be as long as the book. In my own case, however, the greatest debts must be to Sir Harold Jeffreys, R. T. Cox, C. E. Shannon, and G. Polya, the last three for reasons which will be clear from the text. In the case of Jeffreys, I would like to recall the following anecdote.

When, as a student in 1946, I decided that I ought to learn some probability theory, it was pure chance which led me to take the book "Theory of Probability" by Jeffreys, from the library shelf. In reading it, I was puzzled by something which, I am afraid, will also puzzle many who read the present book. Why was he so much on the defensive? It seemed to me that Jeffreys' viewpoint and most of his statements were the most obvious common sense; I could not imagine any sane person disputing them. Why, then, did he feel it necessary to insert so many interludes of argumentation vigorously defending his viewpoint? Wasn't he belaboring a straw man?

This suspicion disappeared quickly a few years later when I consulted another well-known book on probability (Feller, 1950) and began to realize what a fantastic situation exists in this field. The whole approach of Jeffreys was summarily rejected as metaphysical nonsense, without even a description. The author assured us that Jeffreys' methods of estimation, which seemed to me so simple and satisfactory, were completely erroneous, and wrote in glowing terms about the success of a "modern theory", which had abolished all these mistakes.

Naturally, I was eager to learn what was wrong with Jeffreys' methods, why such glaring errors had escaped me, and what the new, improved methods were. But when I tried to find the new methods for handling estimation problems (which Jeffreys could formulate in two or three lines of the most elementary mathematics), I found that the new book did not contain them. On the contrary, the reader was told that these problems did not belong to probability theory at all, but to a new field, statistical inference, which was based on entirely different principles, was very advanced, and should be studied only after one had mastered probability theory and measure theory.

Throwing caution to the winds, I then took down an armful of advanced texts in statistics. Here I found an entirely new vocabulary, new mathematical demonstrations, the most meticulous attention to minutiae which I could not conceive of as being relevant to any application, but no underlying unity of method. I was particularly interested in problems of parameter estimation of the type which arise in extracting signals from noise; but instead of giving a single method applicable to all such problems, as Jeffreys did, the authors would give several different methods for treating each individual problem. They gave demonstrably different results, and the reader was left with nothing to choose between them. An "estimator" ought to be, if possible sufficient, unbiased, efficient, or asymptotically so; but the reader could find neither a clear statement of the relative importance of these, nor any general rule by which an estimator with such properties could be constructed. Instead, the procedure was merely to guess various functions of the sample values on intuitive grounds, and then test them for bias, efficiency, etc.

On the other hand, Jeffreys' method (which was, of course, application of Bayes' theorem in essentially the same way Laplace had used it) told us at once which estimator should be used. It was a revelation to me to read Jeffreys' beautifully clear explanation of why the sample mean is not the best estimate of the population mean except in the special case of Gaussian distribution; in all other cases one should use a weighted average of the sample values. All of a sudden I could see the justification for the physicist's common-sense practice of discarding measurements which show a wide deviation, about which mathematical writers still complain; and a refinement of this in which the theory tells us just how to assign less weight to more widely deviating measurements. But none of this was to be found in the books on statistical inference. Here, for example, authors quote the well-known proposition that for a Cauchy distribution the mean of an arbitrarily large sample is no better estimate of the population mean (by the criterion of efficiency) than a single observation, and that the sample median has a rather good asymptotic efficiency; but they do not offer us any reasonable estimator for the small-sample case. Jeffreys' method determines a definite weighted average estimator which is better than the median for any sample size, and much better in the case of small

samples; but I have yet to find an orthodox writer who uses it; or even acknowledges its existence.

Observing these things, I was completely mystified by every author's contemptuous dismissal of Jeffreys' method, which was done invariably on purely philosophical grounds, without letting the reader see how it works in even a single application.

If yousay that method A is better than method B, I think you ought to mean by this, at the very minimum, that there is at least one specific problem where it leads to a better answer; and to prove your point you need only exhibit that problem. But I could not, then or since, find any orthodox writer who had produced any such example, except for one case (noted in Lecture 7) where there was an error in the calculation, and a few others (Lecture 16) based on gross misapplication of the Bayesian method (i.e., taking the solution to one problem and complaining that it is not also the solution to an entirely different problem). Indeed, on working out a few cases for myself, the outcome was invariably the same; whenever there was any appreciable difference in the results, it was Jeffreys' result which clearly agreed best with common sense. Once one understood the mathematical situation, it was easy to invent problems where the orthodox statisticians' methods broke down entirely and gave absurd results; but I was unable to produce any case where Jeffreys' method, properly applied, failed to lead to an intuitively reasonable conclusion.

All this took place just at the time of appearance of Wald's book, "Statistical Decision Functions". It required several years for the full implications of this monumental work to be appreciated, but by now many workers in statistics have recognized the source of, and remedy for, all this confusion.

The details occupy much of Lectures 5, 6, 13, 14; but stated in the briefest terms, the mathematical situation uncovered by Wald showed that in all respects that matter in real applications, Jeffreys was right all along. The most important recent advance in statistics has taken us right back to the methods developed by Bayes, Laplace, and Daniel Bernoulli in the 18'th century, which generations of statisticians have held to be nonsense. For twenty years, the physicist who was fortunate enough to consult Jeffreys' book had at his fingertips statistical methods which were simpler, more general, and more powerful than anything the orthodox statistician had to offer.

Needless to say, the above assertions are not going to be accepted overnight in all quarters. If the reader is puzzled by my repeated lapses into argumentation, I ask him to realize that I am not only trying to be constructive and give a unified method, but I am also trying to answer in a

a single volume all the objections to this method which have filled the statistical literature for sixty years. In these sections, I am, in effect, supplying the reader with ammunition which he will need if he tries to discuss these issues with colleagues who have been trained only in the "orthodox" point of view. In this connection, I would like to express my gratitude to two anonymous reviewers of this book who gave valuable suggestions on how to strengthen these arguments, and most of all to a third reviewer, who by his objections reassured me more than any friendly reviewer could possibly have done, that in these sections I am not wasting time and space belaboring a straw man.

The Galileo Strategy. Recently, my attention was called to a remarkable article, "Linguistic Analysis of a Statistical Controversy", by Irwin D. J. Bross (1963),* which attacks Bayesian methods in a way that cries out for a constructive reply. I hope that this book may serve that purpose; and to make it "constructive" we need to recognize that further debate on the philosophical level would be not only fruitless, but it would miss the real issue facing us today. As already noted, we have been debating the philosophical issue for well over a century, and perhaps no great harm will be done if it goes on for another century. But there is a far more important issue which should, and I think can, be settled quickly.

The question of immediate importance is not whether Bayesian methods are 100 per cent perfect, or whether their underlying philosophy is opprobrious, but simply whether, at the present time, they are better or worse than orthodox methods from the standpoint of (a) the actual results they give in practice, (b) the range of problems where they can be applied, and (c) their ease of application.

For example, a large amount of reliability and quality-control testing is needed in modern technology. In some cases, particularly, in the aero-space field, acquisition of a single data point can cost more than the yearly salary of a statistician. Use of statistical methods that fail to extract all the relevant information from a sample, or fail to make use of relevant prior information, is therefore not only illogical; it can lead to staggering economic waste.

As another example, statistical methods are destined for an every-increasing role in helping make fundamental military and governmental policy decisions. In this case, use of methods that fail to make full use of the available information might lead to consequences whose magnitude cannot be measured in economic terms at all. In a very real sense, statistics has become too important to allow its methods to be determined merely by the relative numbers, or aggressiveness, of two parties in a philosophical dispute.

---

*Full references are given in Appendix A.

To assert the superiority of either approach on grounds of some philosophy about the "true meaning of probability" without examining the facts concerning performance in specific cases would be to cast out everything we have learned about scientific methodology and to return to the methods of that learned Doctor of the seventeenth century, who assured the world his theology had proved there could be no moons about Jupiter, and steadfastly refused to look through Galileo's telescope.

Since 1953, I have been making constant routine use of Bayesian methods in statistical problems of physics and engineering, and comparing their results with those obtained by orthodox methods. As a result, I believe that the practical issue can be removed entirely from the realm of ideology, and settled on the level of demonstrable fact. To indicate how this can be done, I offer this book as a small, but revealing glimpse through the Galileo telescope of statistics.

To define our field of view let us start, as did Bross, by quoting the words of J. W. Pratt (1963): "Now that I have ceased pretending to be impartial, I may point out that no connected argument leading to the orthodox methods has ever been advanced. Neyman and Pearson contributed vitally to our understanding by their formulation of statistical problems, but they have never claimed their methods were more than ad hoc procedures with some pleasant properties. Their methods, while extremely ingenious and useful, are not completely satisfactory, let alone uniquely objective and scientific."

Unlike Bross, I am unable to discern any "Neo-Bayesian jargon" or "incongruencies" here; only a clear and accurate statement of fact. But, since a lengthy attack on this statement has been published, it will be of interest to see how it can be defended. Bross particularly objects to the remark that "no connected argument" has been advanced for the orthodox methods, and he specifically brings up the matter of significance tests and confidence intervals. Therefore, we will give particular scrutiny to these topics [as previewed in Jaynes (1973)] when we come to them in the course of the lectures. In fact, we give a quite general proof that these methods, when improved to the maximum possible degree by taking into account all ancillary statistics, lead finally to just the results that could have been derived in three lines by the methods of Laplace.

I am indebted to S. R. Faris, John Heller, L. Massé, S. M. Foulks, and many other workers at the Socony-Mobil Field Research Laboratories in Dallas, for their kind hospitality during my two visits, and for undertaking the monumental task of preparing a typed copy of the lectures, complete with equations, from the tape recordings. Only a person who has done this kind of work can realize how much labor is involved.

Many of the details given in connection with applications in the last half of the book were worked out by, and appear in the doctoral theses of, my students, Perry Vartanian, Steve Heims,

Larry Davis, Ray Nelson, Douglas Scalapino, Baldwin Robertson, Joel Snow, Wm. C. Mitchell, and Charles Tyler.

I am indebted also to several hundred students and colleagues who have attended these lectures at various places. By their penetrating questions they have forced me to think much more carefully about many of the issues raised in these talks.

Finally, it should be emphasized again that in most places the text is a literal transcript of actual lectures, and that expressions used in speaking are not always those one would use in writing. In particular, the term "statistician" is often used as an abbreviation for "statistician of the extreme objectivist school of thought which has dominated the field for several decades." In actual fact, many statisticians are well aware of the points made here, and would find themselves in agreement with my arguments. We have noted the views of Pratt; as other examples one can cite the work of D. V. Lindley (1965) and I. J. Good (1959) who outlines a "neoclassical" theory in which "Bayes' theorem is restored to a primary position from which it had been deposed by the orthodox statisticians..." I hope that statisticians of the neoclassical persuasion (whose numbers are rapidly increasing) will understand, and not be offended by, my use of the term.

E. T. Jaynes
St. Louis, Missouri

June 1973

Lecture 1

INTRODUCTION AND BACKGROUND

Let's start out by putting our motto on the board:

"PROBABILITY THEORY IS NOTHING BUT COMMON SENSE
REDUCED TO CALCULATION" (Laplace).

This is the motto and this is the exact summary of everything I'm going to tell you in all these talks.

Our main concern is with applications of probability theory, but we're going to have to spend some time on foundations of probability theory for a very simple reason. Before you can apply any theory to any problem, you first have to make the decision that the theory applies to the problem. It turns out that this is not always an easy decision to make. In most of the problems in science and engineering where you might think of using probability theory, your decision as to whether its use is really justified can depend entirely on how you approach the fundamentals of probability theory itself. In other words, what do we mean by probability? Before we can discuss any applications, we'll have to make up our minds about that.

My purpose in these talks is to show that, with a little different approach to fundamentals than the one usually given nowadays, we can extend the range of practical problems where probability theory can be used, and in some known applications we can simplify the calculations.

1.1 Historical Remarks*

Before going into details a few historical remarks might be of interest, to show how it could happen that a person who is a rather strange mixture of theoretical physicist and electrical engineer could get really worried about the foundations of probability theory. The things that I'm going to talk about here arose from my attempts, over a period of ten years, to understand what statistical mechanics is all about and how it is related to communication theory. In 1948 I was very fortunate in being a graduate student in Princeton, and I took a course in statistical mechanics from Professor Eugene Wigner, who went very carefully into the various approaches to statistical mechanics and in particular, pointed out the unsolved problems that still existed. I was impressed by the fact that everyone who has written about the fundamentals has a very ready way of resolving all the famous paradoxes; but that no two people have done this in the same way.

It was just during this year that Shannon's papers (Shannon, 1948)**, announcing the birth of information theory, appeared. I discovered them accidentally in the Princeton library, took them back to my room, and disappeared from the face of the earth for about a week. When I finally came out, I ran through the halls of Princeton explaining to anybody who would listen to me (and a few who wouldn't) that this was the most important piece of work done by any scientist since the discovery of the Dirac equation. It's almost impossible to describe the psychological effect of seeing our old familiar expression for entropy derived in a completely new way, and

---

*This and the following Section describe the history and motivation of the work reported. The reader who does not care about this and wants to get on with the constructive development can turn immediately to Lecture 2.

**Insertions of this type refer to the General Bibliography in Appendix A.

then applied to problems of engineering which apparently have no relation to thermodynamics. But all of the inequalities, which are often associated with the second law of thermodynamics, turn out also to be statements of the greatest significance in an entirely different context. It seemed to me that there must be something pretty important that we could learn from this situation.

This feeling was shared by a number of physicists and there was quite a rush to exploit all these wonderful new things. But then something went wrong. Quite a few papers appeared in the physics journals inspired by Shannon's work, but there was a scarcity of new results useful to physics. This caused a psychological reaction, and by 1956 Information Theory had acquired a bad reputation among physicists.

I think the time has come now when physicists might find it worthwhile to take a sober second look at Information Theory and what it can do for them. And with the benefit of hindsight, we can see what went wrong in those first few years. The first efforts were based only on a mathematical analogy between statistical mechanics and communication theory, in which the appearance of the same mathematical expression was the dramatic thing. The essential link between them--the thing I want to try to show here--is not one of mathematics, but something more subtle. Until you see what the link is, you can't expect to get results out of this situation. Now let's see why this is so.

The mere fact that a mathematical expression like

$$\Sigma p_i \log p_i$$

shows up in two different fields, and that the same inequalities are used, doesn't in itself establish any connection between the fields. Because after all,

$$e^x, \qquad \cos\theta, \qquad J_o(z)$$

are expressions that show up in every part of physics and engineering. Every place they show up, the same equalities and the same inequalities turn out to be useful. Nobody interprets this as showing that there is some deep profound connection between, say, bridge building and meson theory. The reason for that is the underlying ideas are entirely different.

Now the essential content of both statistical mechanics and communication theory, of course, does not lie in the equations; it lies in the ideas that lead to those equations. And at first glance there doesn't seem to be any relation at all between the kind of reasoning that the physicists go through in statistical mechanics and the kind of reasoning that Shannon went through. We might describe this by paraphrasing a statement of Albert Einstein (Einstein, 1946) that I like very much: Science is fully justified in identifying these fields only after the equality of mathematical methods has been reduced to an equality of the real nature of the concepts. You recall that Einstein insisted on exactly this point in connection with gravitational and inertial mass. It had been known, for 200 years before Einstein was born, that gravitational mass and inertial mass were experimentally proportional to each other; by proper choice of units you can make them numerically equal. Einstein refused to identify them; i.e. to accept this empirical equality as a general principle of physics, until he could reduce inertial mass and gravitational mass to the same concept. He had to pay a rather high price to do this. Before he could find a viewpoint from which he saw them as special cases of the same idea, he had to invent General Relativity.

It is interesting to note that this principle was appreciated equally well by J. Willard Gibbs, many years earlier. In his response to the Ameri-

can Academy of Arts and Sciences of Boston, on the occasion of his being awarded the Rumford Medal (January 12, 1881), Gibbs remarked: "One of the principal objects of theoretical research in any department of knowledge is to find the point of view from which the subject appears in its greatest simplicity." Gibbs had shown in his famous work of 1878 that classical thermodynamics appears particularly simple if we regard entropy as the fundamental quantity; from its dependence on energy, volume, and mole numbers all thermodynamic properties of a system are determined.

These examples could be used with profit in all parts of science. We won't commit any serious error of methodology if we try to follow the examples of Gibbs and Einstein in our problem, because it's really a very similar sort of thing. So the job as I saw it was not to try to invent any new fancy mathematics. That would presumably come later if we were successful; but the immediate job was to try to find a viewpoint from which we could see that the reasoning behind communication theory and statistical mechanics was really the same. As it turns out, to do this requires a rather drastic reinterpretation of both fields; and this reinterpretation clears up several outstanding difficulties in each field.

1.2 The Gibbs Model.

Now to state the problem a little more specifically, I'd like to go very briefly into the version of statistical mechanics that Gibbs gave us (Gibbs, 1902), and try to show the sense in which my work is not only an attempt to generalize his theory, but also an attempt to make use of another lesson in methodology which he gave to science.

Most of the discussions about the foundations of statistical mechanics consist of Mr. A criticizing the basic assumptions of Mr. B and this process is always fruitless and inconclusive. It never leads to any useful results.

However, there is one person who has kept free of that, and his name is J. Willard Gibbs. I think of all people who have written on statistical mechanics, he is the only person who has stayed above this kind of criticism. He did this by a very clever trick. He avoided criticism of his assumptions by not making any assumptions, and by pointing this out to the reader in the preface to his book.

Gibbs simply constructed models in which he assigned certain probabilities for certain situations, and in introducing them he did not say a word about why he chose those particular probabilities. In the preface he tells us that the reason for this has something to do with difficulties which the theory faced in his day, and in particular he mentioned the fact that the experimental specific heat of diatomic gases comes out only 5/7 of what he expected it to be on the basis of his theory. There are a few other difficulties. The paradox about entropy of mixing, for example, and the fact that his theory failed to predict the actual values of equilibrium constants and vapor pressures until you added still more assumptions.

I like to think that there is another reason why Gibbs operated this way. It was maybe even more compelling that the temporary difficulties. Of course, all those difficulties we recognize today as signaling the first clues to the quantum theory. We all know that Gibbs was a very shrewd old gentleman who was a master of science as it existed in his day. I think he was equally well a master of psychology. He realized that the physics of his day and the probability theory in his day didn't provide any really convincing arguments to justify the probability assignment of his canonical ensemble in terms of more fundamental things. And yet, his work had shown that it had all the formal properties which convinced him that it must be right. It clearly was the neatest, most elegant, and simplest way of describing thermodynamics.

Suppose you were in a situation like that. Which is the best way to proceed? I think Gibbs said to himself, "If I try to say a single word to justify this canonical distribution, if I try to invent any argument to back it up, then almost everybody who reads this work will conclude, quite irrationally, that the validity of my equations depends on the validity of those arguments. But I know in my bones that this theory is right independently of any arguments I am now able to give, because it has formal properties which make it superior to any other. So I will say as much as possible about what I know, and as little as possible about what I don't know. The real justification will have to come later." So he simply introduced his canonical ensemble by entitling a chapter "On the Distribution in Phase called Canonical, in which the Index of Probability is a Linear Function of Energy," and that was it. He goes right on into the discussion.

So you can't say to Gibbs, "How do you know that this is the right probability distribution?" He'd be perfectly justified by answering something like this: "I didn't say it was the right probability distribution, and I'm not sure the question has any meaning. I'm simply constructing a model for my own amusement. My canonical probability assignment is not derived from anything, it's not an assumption about anything. It's a definition of which model I propose to study. After this model is set up, we can compare its predictions with experimental facts and see how far this model is able to reproduce thermodynamic properties of systems. If the model turns out to be successful, then it will be worthwhile to consider whether, and in what sense, we might consider it to be correct."

I think that's a very clever attitude to take - it avoids so much useless argumentation. It's a good example also of the methodology we really have to use in all theoretical physics. If we had to be sure we were right before starting a study, we would just never be able to do anything at all. We have

to start out by arbitrarily inventing something, some model, which we don't attempt to justify in terms of anything deeper at the time, and see where it leads us. Every once in a while we find that we can invent a model which has very great success in reproducing observed phenomena, and whenever this happens we get convinced that there must be some deeper reason why this model is correct. Then we repeat the process. We try to invent another model operating at some deeper level, from which we can deduce the features of our old model. The exciting thing about this is that when we finally succeed, we always find that the new model is much simpler than the old model, but at the same time is much more general.

There are all sorts of examples of this in the history of science which you all know about; for example, in electromagnetic theory, the experimentalists had produced a large number of separate equations and rules of thumb--the work of Coulomb, Ampere, Faraday, Henry, and so on. And then these were all summed up in Maxwell's equations. Maxwell's equations are much simpler than this series of models which they replaced; but still they are more general, and predicted new phenomena which the experimentalists hadn't found. In fact, Maxwell's equations proved to be so general that to this day, a century later, they still provide the theoretical basis for all of electrical engineering.

Perhaps the best example of all is the tremendous complication which spectroscopy got into by the early 1920's. All the rules of thumb that were developed in predicting what spectral lines would occur and which ones would not, estimating where they would be, and so on. These rules of thumb were quite successful, of course. You could use them for practical prediction. But then we have the Schrödinger equation, which suddenly in a single differential equation says everything that all these rules ever said, and much more; so much more that we are still finding new things from it.

How has the Gibbs model fared? We've had it for 70 years now. It has fared very well, except for these minor changes which have something to do with quantum theory. We find that in every case where you can work out the mathematics, the model has been successful in reproducing observed properties of matter in the limiting case of thermal equilibrium. There are some equilibrium cases where the mathematics is rather resistant to calculation, particularly the phenomenon of condensation; and we don't really know whether the Gibbs model exhibits condensation for general attractive forces, in the sense of being able to prove it rigorously. But I don't think anyone doubts that the Gibbs model would be successful here if we were just better mathematicians than we are. So for the sake of the argument, let's just grant that the Gibbs model has turned out to be completely successful in reproducing all features of equilibrium thermodynamics.

Because of its success, naturally, attempts would be made to justify the Gibbs model in terms of something deeper. Unfortunately, these attempts do not seem to have been successful; at least I don't think there is a single one of them which is so considered by any clear majority of the physicists who worry about these things.

It hasn't been easy to get rid of the idea that the ultimate justification of the Gibbs model must be found somehow in the laws of physics. By this we mean particularly, say, the Schrödinger equation or the Hamiltonian equations of motion on a microscopic level. For this reason you have this enormous amount of work that has been expended on "ergodic" approaches to statistical mechanics, in which we tried to prove that the time average of some quantity for a single system would, in consequence of the equations of motion, be equal to an average over the Gibbs ensemble. But the results of this approach have remained inconclusive, and it has done nothing to extend the Gibbs model to more general situations, as real

advances in understanding always do.

More specifically, while the ergodic arguments have led to a number of important theorems (such as reduction of the original problem to that of metric transitivity), they have led to no definite conclusions proved applicable to real physical systems even in the equilibrium case; and they have provided no clues as to how a general theory of irreversible processes might be set up.

I don't want to go at this point into any detailed criticism of past attempts to justify the Gibbs model, because that would take a lot of time and would again be one of those fruitless and inconclusive kinds of criticism which leads nowhere. But I'd like to indicate why it seems to me that any appeal to the laws of physics may miss the point. It is simply that the problem is not to justify any statement about physics. The problem is to justify a probability assignment, and you can't deduce probability from certainty. No matter how profound your mathematics is, if you hope to come out eventually with a probability distribution, then some place you have to put in a probability distribution; and nothing in the equations of motion tells you what distribution to put in. They can give you only relations between probabilities, at different times.

You might note that this argument has nothing to do with whether we're considering classical or quantum statistical mechanics. In classical theory we have our precisely defined states where we've specified the value of every coordinate and every momentum to arbitrary accuracy, and the equations of motion then determine uniquely what every coordinate and momentum must be at some other time. In quantum theory we don't use that method of description, but we still have our precisely defined states. They now are points in a linear vector space, or Hilbert space, whose motion is uniquely determined by the Schrödinger equation.

The analogy goes a good deal deeper; Liouville's theorem in the classical case finds its analog in the fact that in quantum theory the equations of motion induce a unitary transformation, which is therefore a measure-preserving transformation, in the Hilbert space. The fact that the total phase volume below a certain energy is finite in the classical case, has its analog in the fact that the linear manifold spanned by all eigenfunctions of the Hamiltonian with energies below a certain value, is a finite-dimensional vector space. These are about the only properties which are actually used in the ergodic arguments. Therefore practically everything that has been said about these problems in classical statistical mechanics carries over immediately to quantum theory.

One of our major objectives is to justify the Gibbs canonical probability distribution in terms of something more fundamental. The only thing we could accomplish by applying the laws of physics is that we could carry out transformations and express the same distribution in terms of some other parameters. But the distribution of Gibbs is already as simple as any we could hope to get in this way, and afterwards we would still be faced with exactly the same problem; to justify some probability assignment.

It seems to me that if we're ever going to justify the Gibbs model in any meaningful way, we'll have to justify it directly on its own merits, without considering the laws of physics at all. In other words, the problem is to find a viewpoint from which we can see that the Gibbs model, and Shannon's model of a communication process, are special cases of a general method of reasoning.

In the next two lectures, we're going to take what may seem like a rather long detour, and study the general problem of plausible reasoning--also known by the more highbrow, and more restrictive, name of inductive reasoning (I'm not going to bother to distinguish between these terms).

But if you'll bear with me, I think you'll find that we can give, not quite rigorous theorems, but very powerful heuristic arguments, which indicate what this more general viewpoint is.

Lecture 2

PLAUSIBLE REASONING

Suppose some dark night a policeman walks down the street, and the place is completely deserted apparently; but all of a sudden he hears a burglar alarm, he looks across the street, and sees a jewelry store with a broken window. Also, there's a gentleman wearing a mask, crawling out through the broken window, carrying a bag which turns out to be full of watches and diamond rings. The policeman doesn't hesitate at all in deciding this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion?

2.1 Deductive and Inductive 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. He noticed that a passing truck had thrown a stone through the window, and he was merely protecting his own property. You see, the conclusion which seems so easily made was certainly not an example of logical deduction.

Now while we agree that the policeman's reasoning process was not an example of logical deduction, we still will grant that it had a certain degree of validity. The evidence didn't make the gentleman's dishonesty certain, but it did make it extremely plausible. This is an example of the

kind of reasoning which we all have to use a hundred times a day. We're always faced with situations where we don't have enough information to permit deductive reasoning, but still we have to decide what to do.

The formation of plausible conclusions is a very subtle process and it's been discussed for centuries, and I don't think anyone has ever produced an analysis of it which anyone else finds completely satisfactory. These problems haven't been solved and they're certainly not going to be solved in these talks; but I do hope that we'll be able to say a few new things about them.

All discussions of these questions start out by giving examples of the contrast between deductive reasoning and plausible reasoning. The syllogism is the standard example of deductive reasoning:

If A is true, then B is true

A is true

Therefore, B is true

or, its inverse:

If A is true, then B is true

B is false

Therefore, A is false

This is the kind of reasoning we'd like to use all the time; but, unfortunately, in almost all the situations we're confronted with we don't have the right kind of information to allow this kind of reasoning. We fall back on weaker forms:

If A is true, then B is true

B is true

Therefore, A becomes more plausible

The evidence doesn't prove that A is true, but verification of one of its consequences does give us more confidence in A. Another weak syllogism, still using the same major premise, is:

If A is true, then B is true

A is false

Therefore, B becomes less plausible

In this case, the evidence doesn't 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 the policeman in this example was not even of the above types. It is best described by a still weaker form:

If A is true, then B becomes more plausible

B is true

Therefore, A becomes more plausible

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 had a very strong convincing power. There's something which makes us believe that in this particular case, his argument had almost the power of deductive reasoning.

This shows 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. And it does it in some way that makes use of our past experience as well as the specific data of the problem we're reasoning on. To illustrate, for example, that the policeman was making use of the past experience of policemen in general, we have only to

change that experience. Suppose that these events happened several times every night to every policeman, and in every case the gentleman turned out to be completely innocent. Well, very soon policemen would be ordered to ignore such trivial things. This shows that in our reasoning we depend very much on past experience--or as we will presently call it, on prior information--to help us in evaluating the degree of plausibility. This reasoning process goes on unconsciously, almost instantaneously, and we conceal how complicated it really is by calling it common sense.

Professor George Polya has written three books on plausible reasoning (Polya, 1945, 1954), pointing out all sorts of interesting examples, showing that there are fairly definite rules by which we do plausible reasoning (although in his work they remain in qualitative form). Evidently, the deductive reasoning described above has the property that you can go through arbitrarily long chains of reasoning of this type and the conclusions have just as much certainty as the premises. With the other kinds of reasoning, the reliability of the conclusion attenuates if you go through several stages. Polya showed that even a pure mathematician actually uses these weaker kinds of reasoning most of the time. Of course, when he publishes a new theorem, he'll be very careful to invent an argument which uses only the first kind of reasoning; and his professional reputation depends on his ability to do this. But the process which led him to the theorem in the first place almost always involves one of the weaker forms.

Now the problem I'm concerned with is this. Is it possible to reduce this process of plausible reasoning to quantitative terms? The idea of inventing a mathematical theory of reasoning, both deductive and inductive, is a very old one. Leibnitz had speculated on such a "Characteristica Universalis," almost 200 years before Boole's The Laws of Thought (1854) provided a calculus of deductive reasoning. When the theory of probability

was developed, culminating in Laplace's Theorie Analytique (1812), it was believed to be the long-awaited "calculus of inductive reasoning," fully developed. Throughout the 19th century this was the prevailing view, expounded by such people as Laplace, de Morgan, Maxwell, Poincaré, and many others. And yet, in the 20th century we find that probability theory has erupted into controversy, almost all of this fruitless, inconclusive kind, in which one person attacks the assumptions of another person.

This issue has been framed rather sharply by Ludwig von Mises (von Mises, 1957; 1963) who is really violent in denouncing any idea that probability theory has anything to do with inductive reasoning. He insists that it is, instead, "the exact science of mass phenomena and repetitive events." On the other hand, Sir Harold Jeffreys (Jeffreys, 1939; 1955) is equally vigorous in upholding the opposite view, and insists that probability theory is exactly what Laplace thought it was: the "calculus of inductive reasoning."

Well, which is it? I want to point out that it makes a big difference in applications. Science and engineering offer many problems where use of probability theory is entirely legitimate on one interpretation, and entirely unjustified on the other. Even in cases where both viewpoints would allow the use of probability theory, your decision as to which mathematical problems are important and worth working on, can still depend on which viewpoint you adopt. (As an example, whose meaning will become clear later: when approximations are necessary, is it the sampling distribution of a statistic or the posterior distribution of a parameter that should be approximated? The two different schools of thought will give opposite answers to this; and each regards the mathematical labors of the other as effort wasted on a false problem.)

Sooner or later, such an unsettled condition in probability theory couldn't fail to have pretty serious repercussions in theoretical physics

and engineering--both of which make more and more use of probability methods. And so now you see why any serious student of physics or engineering must become worried about this situation. I hope to show in these talks that some of the outstanding unsolved problems in both physics and communication theory have their origin in this state of utter confusion which exists in the foundations of probability theory.

2.2 Analogies with Physical Theories

In physics, we quickly learn 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, as I already said, we can invent a model which reproduces several features of one of these pieces, and whenever this happens we feel that great progress has been made. These mathematical 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. 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'll take a similar course. We won't try to understand it all at once, but we'll feel that progress has been made if we are able to construct idealized mathematical models which reproduce a few of its features; that is the methodology of Gibbs. We expect that any model we are now able to construct will be replaced by better ones in the future, and we don't know whether there is any natural end to this process.

The ultimate test of a physical theory is not, "Can you demonstrate it by logic?" but only; "Is it free of obvious inconsistencies and does it agree with experiment?" It has taken the human race thousands of years to comprehend this simple fact. It was utterly unknown to the ancient

philosophers, and Galileo was the first to demonstrate clearly the advantages of recognizing it.

It is exactly the same in our present problem. The test of any model of plausible reasoning is not "Can you prove that it is correct?" Real life, unfortunately, does not permit such a Utopian program. The only test which can actually be applied in practice is: "Is it free of inconsistencies and in agreement with common sense?" It has taken us a long time to realize this, and I'm sure that there are still many people who will dispute it vigorously.

The analogy with physical theories goes a lot deeper than a mere analogy of method. Often, the things which are most familiar to us turn out to be the hardest to understand. Our universities can train people to perform surgery on the living heart and measure the internal charge distribution of the proton; but nobody seems to know how to prevent the common cold, and all of modern science is practically helpless when faced with the complications of such a commonplace thing as 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 find that any advance in knowledge ultimately leads to consequences of the greatest practical value, but of a totally unpredictable nature. Roentgen'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 the possibility of practically instantaneous communication all over the earth.

Our mathematical models for common sense also exhibit, although on a more modest scale, this feature of practical usefulness. Any successful model, even though it may reproduce only a very 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 plausible reasoning which are so involved that we would never attempt to solve them without its help. Thus the problem of optimum design of an electrical filter or an antenna (which is just a particular kind of filter, operating in space instead of in time) can sometimes be solved by applying a model of common sense. Similarly, we will show that the prediction of the laws of thermodynamics, including all experimentally reproducible features of irreversible processes, can be viewed as an application of a single, formally very simple model of common sense.

Models may 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." One of the best answers to this attitude was given by J. von Neumann in a talk on computers given at the Institute for Advanced Study in Princeton in 1948, which I 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: "Look here. 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!"

The only operations which a machine cannot perform for us are those which we cannot describe in detail. The only limitations on making "machines which think" are our own limitations in not knowing exactly what "thinking" consists of. For further comments on this, see my recent Letter (Jaynes, 1963a). But in our study of common sense we will be led to some very explicit ideas about the detailed 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, it becomes a kind of blueprint showing us how

to build a machine which operates on incomplete data and does plausible reasoning instead of deductive reasoning. In science fiction, such machines have been an accomplished fact for many years. In fact, I want to turn this idea around and instead of asking, "How can we build a mathematical model of common sense?" I want to ask, "How could we build a machine which would do plausible reasoning?"

2.3 Introducing the Robot

Now the question of the process of plausible reasoning that actual human brains use is very charged with emotion and misunderstanding, to the extent that the only solution is to avoid it. Also, it is so complicated that we can make no pretense of explaining all its mysteries; and in any event we are not trying to explain all the abberations and inconsistencies of human brains. That is an interesting and important subject, but it is not the subject we are studying here. We are trying rather to understand some of the good features of human brains.

In order to direct attention to constructive things and away from controversial things which we can't answer at present, we will follow the methodology of Gibbs and invent an imaginary beast. His brain is to be designed by us, so that he reasons according to certain definite rules. The rules are suggested by properties of human brains which we think, or hope, exist; but by introducing the beast we accomplish the following. You can't object to the theory on the grounds that we have failed to prove the "correctness" of the rules, whatever that may mean. We are free to adopt any rules we please. That's our way of defining which beast we are going to study. After we've worked out the properties of this beast, we can then compare the results of his reasoning process with the results of ours. If you find no resemblance between the way the beast reasons and the way you

reason, then you're free to decide that the beast is nothing but an idle, useless toy. But if you find a very strong resemblance, which makes it almost impossible to avoid concluding "I am this beast," then that will be an accomplishment of the theory, not a premise.

Now, let's take a problem with maybe some science fiction overtones. We've been assigned the job of designing the brain case of a robot. This is supposed to be a very sophisticated robot. He doesn't just receive orders and carry them out. He also has to have the ability to learn, he has to be able to make judgments on his own, he has to decide on the best course of action even when we fail to give him full instructions. This means that his brain has got to contain some kind of computing machine which will carry out plausible reasoning whenever the information we give him is insufficient to permit deductive reasoning. How shall we design his brain case? This is a fairly definite engineering problem.

Well, our robot is going to reason about propositions. We denote various propositions by letters A, B, C, and so on, and for the time being we'll have to require that any proposition we use will have, at least to the robot, an unambiguous meaning. It must also be of such a "logical type" that it makes sense to say that the proposition must be either true or false. Of course, not all propositions are of that type at all. Later on we'll see whether there are any possibilities of relaxing that restriction.

Now to each proposition the robot is going to associate some plausibility, which represents his degree of belief in the truth of the proposition, based on all the evidence we have given him up to this time. In order that these plausibilities can be handled in the circuits of his brain, they must be associated with some physical quantity such as voltage or pulse duration or frequency, and so on, however you want to design him. This means that there will have to be some kind of association between degrees of plausibility

and real numbers. This assumption, you see, is practically forced on us by the requirement that the robot's brain must operate by the carrying out of some definite physical process.

Let me emphasize the contrast between such a robot and a human brain. We have decided that we will attempt to associate mental states with numbers which are to be manipulated according to definite rules. Now it is clear that our attitude toward any given proposition may have a very large number of different "coordinates." 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 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 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 an extremely large number of coordinates. A moment's introspection will show that, quite generally, the situations of everyday life are those involving many coordinates. It is just for this reason, I suggest, that the most familiar examples of mental activity are often the most difficult to reproduce by a model.

We might speculate further. 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.

I interject these remarks 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 "multi-dimensional" theories of mental activity, which would more and more resemble the behavior of actual human brains. Such a theory, if successful, might have an importance beyond our present ability to imagine.

For the present, however, we will have to be content with a much more modest undertaking. Is it possible to develop a consistent "one-dimensional" model of 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; and at the risk of belaboring it, let me stress again: 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 any such thing; it is to investigate whether it is possible, in our robot, to set up such a correspondence without contradictions. If the attempt to do this should fail, then we will have to consider more complicated kinds of association; but I propose to try out the simplest possibility first.

We'll adopt a convenient but nonessential convention; that this will be done in such a way that a greater plausibility always corresponds 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.

To state the above ideas more formally, we introduce some notation of the usual symbolic logic, or Boolean algebra. By the symbolic product

$$AB$$

we mean the proposition "both A and B are true." Obviously, AB and BA are

the same proposition. The expression

$$A+B$$

stands for the proposition: "at least one of the propositions A, B is true," and is the same as B+A. The plausibility that the robot associates with proposition A could, in general, depend on whether we told him that some other proposition B is true. And so we indicate this by the symbol

$$(A|B).$$

I'll call this the "conditional plausibility of A, given B;" or just, "A given B." It stands for some real number. Thus, for example,

$$(A|BC)$$

(I'll read this as "A given BC") represents the plausibility that A is true, given that B and C are true. Or,

$$(A+B|CD)$$

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. Now we've decided that we're going to associate greater plausibility with greater numbers, so

$$(A|B)>(C|B)$$

says that given B, A is more plausible than C.

You know that when a computing machine is asked to divide by zero, it develops a psychosis--the poor machine tries its best, but just can't solve the problem. On some old kinds of desk calculators the only thing you can do is to put the machine out of its misery by pulling the plug. In the interest of avoiding impossible problems, we are not going to ask our robot to undergo the agony of reasoning on the basis of mutually contradictory propositions. Thus, we make no attempt to define $(A|BC)$ when B and C are mutually contradictory. Whenever such a symbol appears, we will understand that B and C are compatible propositions.

Now we wouldn't want this robot to behave in a way that's very greatly different from human behavior, because that would make him very hard to live with and nobody would want to keep such a robot in his home. So, we'll want him to reason in a way that is at least qualitatively like the way you and I reason, as described by the above weak syllogisms. As a further example, if he gets new information which increases the plausibility $(A|BC)$ but does not affect the plausibility $(B|C)$, this of course can only produce an increase, never a decrease, in the plausibility $(AB|C)$ that both A and B are true. And it can only produce a decrease, not an increase, in the plausibility that A is false. This qualitative requirement simply gives us the sense of direction in which reasoning goes; it says nothing about how much the plausibilities change.

Also, it would be nice if we could give this robot a very desirable property which we don't have; namely, that he always reasons consistently. By "consistently" I mean three things:

(a) If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.

(b) If two problems are entirely equivalent; i.e., if the robot's state of knowledge is the same in both, then he must assign the same plausibilities in both.

(c) The robot is completely non-ideological; if he has several pieces of evidence relevant to a question, he does not arbitrarily throw out part of his evidence, basing his conclusions only on what remains; he always takes into account all of the evidence available to him.

All right. Now I claim something which may seem startling. The conditions that we have imposed are:

1. Representation of degrees of plausibility by real numbers.

2. Qualitative correspondence with common sense.

3. Consistency.

These requirements, I claim, uniquely determine the rules according to which this robot must reason; there is only one set of mathematical operations which has all these properties. In the next Lecture we commence the mathematical development by deducing these rules.

Lecture 3

LAPLACE'S MODEL OF COMMON SENSE

We have now formulated our problem, and it ought to be a matter of straightforward mathematics to work out the consequences of our three desiderata:

1. Representation of degrees of plausibility by real numbers.
2. Qualitative correspondence with common sense.
3. Consistency.

This seems in retrospect an obvious and natural thing to do; but historically, the rules we are about to deduce were first stated as arbitrary axioms, on intuitive grounds, without any attempt to demonstrate their uniqueness or consistency. This, of course, left room for practically endless controversy; if the rules are introduced in that way, what right have we to suppose that they are any better than a hundred other arbitrary ones we could invent? It was just this kind of doubt, strengthened by some ridiculous misapplications, that led many to reject Laplace's work and to deny that probability theory has any connection with inductive reasoning. As a result, the development of statistical theory was delayed for many years, and the very "latest" advances in this field amount only to a rediscovery of methods that had been described and used by Laplace and Daniel Bernoulli in the 18th century.

To the best of my knowledge, the first person to see that there is a better way of developing the theory was Professor R. T. Cox (Cox, 1946; 1961). Instead of stating the rules in a way that leaves their consistency and

uniqueness open to doubt, the requirement that they be consistent can be imposed from the start as one of the basic conditions of the theory; and then their uniqueness can be deduced mathematically. Cox's argument, which we follow here, therefore cuts the ground out from under more than a century of unjust criticisms of Laplace's methods.

3.1 Deduction of Rule 1.

We first seek a consistent rule for obtaining the plausibility of AB from the plausibilities of A and B separately. In particular, let us find the plausibility $(AB|C)$; on what others must it depend? Now in order for AB to be a true proposition, it is certainly necessary that B be 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 anything about A, so if we have $(B|C)$ and $(A|BC)$ we will not need $(A|C)$. It would tell us nothing about AB that we didn't already have. Similarly, $(A|B)$ and $(B|A)$ are not needed; whatever plausibility A or B might have in the absence of data C could not be relevant to judgments of a case in which we know from the start that C is true.

We could, of course, interchange A and B in the above paragraph, so the knowledge of $(A|C)$ and $(B|AC)$ would also suffice to determine $(BA|C) \equiv (AB|C)$. The fact that we must obtain the same value for $(AB|C)$ no matter which procedure we choose will be one of our conditions of consistency.

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

$$(AB|C) = F[(B|C),\ (A|BC)] \tag{3-1}$$

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

$$(AB|C) = F[(A|C), (B|C)]$$

might be a permissible form. But we can show easily that no relation of this form could satisfy the conditions that we've imposed on our robot. 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, if I'm told that Mr. Jones lives in Dallas, it might be quite plausible that his eyes are blue, and it might be quite plausible that his hair is black; and it's reasonably plausible that both are true. But, if I'm told that Mr. Smith lives in St. Louis, it is quite plausible that his left eye is blue, and it's quite plausible that his right eye is brown; but it's extremely implausable 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 human beings do, even qualitatively, with that kind of functional relation.

You might try further a relation of the form

$$(AB|C) = F[(A|C), (A|B), (B|A), (B|C)]$$

in which you try to take the above cases into account by allowing all four of these simple plausibilities to determine $(AB|C)$. But even here you can produce counter-examples which show that a function of this form could not reproduce plausible reasoning even qualitatively like ours.

You can blow this up into a whole research project, if you like. Thus, introduce the real numbers

$$u = (AB|C), \quad v = (A|C), \quad y = (A|BC), \quad x = (B|C), \quad w = (B|AC).$$

If u is to be expressed as a function of two or more of v, w, x, and 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 implies A false, etc. If you do this, Myron

Tribus has shown (Tribus, 1969) 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 $u = F(w,v)$, which are just the two possibilities already suggested.

Another way of looking at this, suggested by Mr. Alfred S. Gilman, may seem more attractive than this laborious elimination of alternatives, one by one. We may regard the process of deciding that AB is true as a sequence of two "mental transitions" in which there are only two possible routes, illustrated by the decision tree diagram, Fig. 3.1. In order to decide that AB is true, we

(1) decide that B is true,

(2) having accepted B as true, decide that A is true.

or, we can

(1') decide that A is true,

(2') having accepted A as true, decide that B is true.

Along either route, the state of knowledge in which we decide to make the next transition is indicated by the plausibility symbols on the arrows.

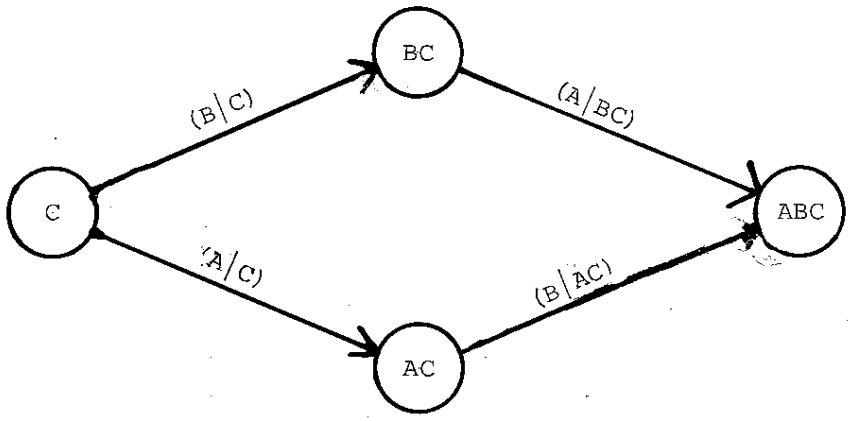

Fig. 3.1. The possible "mental transitions" in deciding that A and B are true, given that C is true.

However you like to view this, I don't think you'll be able to produce any situation where equation (3-1) does not reproduce qualitatively the way you would reason about the situation. (If you can, then all I can say is that your common sense is qualitatively different from mine--and Laplace's--and you are free to design your own robot!)

Now let's start imposing our conditions on the form of this function and see if we can nail down what function it has to be. If anything increases the plausibility $(B|C)$, then that must produce only an increase, never a decrease, in the plausibility $(AB|C)$. Similarly, if anything increases $(A|BC)$, this must also produce an increase, not a decrease, in $(AB|C)$. The only case where it would not produce an increase is where the other independent variable happened to represent impossibility; if we know that A is impossible given C, then, of course, the plausibility of B could increase without affecting $(AB|C)$. Also, the function $F(x,y)$ must be continuous; for otherwise we could produce a situation where an arbitrarily small increase in one of the plausibilities on the right side still results in the same big increase in $(AB|C)$.

In summary, $F(x,y)$ must be a continuous monotonic increasing function of both x and y. I will assume that it's a differentiable function. The derivatives cannot be negative, and they can be zero only in the case where AB is impossible. Now for the condition that it should be consistent.

Suppose that I try to find the plausibility $(ABC|D)$ that three propositions would be true simultaneously. I can do this in two different ways. If the rule is going to be consistent, we've got to get the same result for either order of carrying out the operations. I can first say that BC will be considered a single proposition, and then apply our rule. This plausibility would then be

$$(ABC|D) = F[(BC|D), (A|BCD)]$$

and now in this plausibility of $(BC|D)$ we can again apply the rule to give us

$$(ABC|D) = F\{F[(C|D), (B|CD)], (A|BCD)\}$$

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:

$$(ABC|D) = F[(C|D), (AB|CD)]$$
$$= F\{(C|D), F[(B|CD), (A|BCD)]\}.$$

So by doing it in the other order, we come out with a different expression. If this rule is to represent a consistent way of reasoning, these two expressions must always be the same. The condition that our robot will reason consistently in this case takes the form of a functional equation,

$$F[F(x,y),z] = F[x,F(y,z)]. \qquad (3\text{-}2)$$

Conversely, if this functional equation is satisfied, then our original rule is automatically consistent for all possible ways of finding the joint plausibility of any number of propositions; $(ABCDE|F)$, for example. You can see that there are an enormous number of different ways you can work this out by successive applications of Equation (3-1). And you can show by induction that if the functional Equation (3-2) is satisfied, then you're guaranteed to get the same answer for every possible way of doing it.

This functional equation is one which has quite a long history in mathematics. The earliest reference to it that I know about goes back to 1826, and is a paper by N. H. Abel in the first issue of Crelle's journal. Abel considered equation (3-2) merely as an amusing exercise, and found the general solution by reducing it to a differential equation. The solution has been rediscovered probably dozens of times since 1826. In particular, this is done in a paper by R. T. Cox (Cox, 1946) which I rate as one of the most important ever written on the foundations of probability theory. Cox established the conditions for consistency of this theory in the sense (a) given above, and

my only contribution was to add the qualitative requirements and the other conditions of consistency, which are needed to make the result unique. In a later book (Cox, 1961) Cox's work is given more fully, with some improvements in the derivations. For an appreciation of the importance of Cox's contributions to probability theory, see my review of his book (Jaynes, 1963b).

A particularly neat mathematical treatment of our functional equation (3-2) has been given by J. Aczel in a paper (Aczel, 1948) and in his monumental book on functional equations (Aczel, 1966; Sec. 6.2). He calls it, "The associativity equation." Let me just quote you the theorem that Aczel gives. He says, "Let's let

$$z = x \circ y$$

where

$$x \circ y$$

represents any operation which maps z into the same interval with x and y. In other words, if x is in the interval from a to b, and y is in the interval from a to b, then this operation is one which will always put z into the same interval." He gives a theorem which is exactly backwards from the way we would want it for our application. He considered a formula for the design of the most general slide rule. The general condition that z could be calculated without ambiguity on a slide rule calibrated with numbers x and y is, of course, that there is some monotonic function $f(z) = f(x) + f(y)$. If this is true then you can make a slide rule which gives z in terms of x and y. Aczel shows that a necessary and sufficient condition for that is that the operation x o y must have the following properties:

(1) It must be monotonic: if $x' > x$, then $x' \circ y > x \circ y$, and similarly for y.

(2) It must be continuous: $\lim (x \circ y) = (\lim x) \circ (\lim y)$.

(3) It must be associative: $(x \circ y) \circ z = x \circ (y \circ z)$.

You see that these are precisely the conditions that we have imposed on our

function $z = F(x,y)$. It had to be a monotonic, continuous operation in order to agree qualitatively with common sense. The condition that it should represent a consistent kind of reasoning was just the condition that it be associative. We conclude that the general relation between x, y, z, implied by $z = F(x,y)$ must be expressible in the form

$$F(x,y) = f^{-1}[f(x) + f(y)], \text{ or}$$

$$f(z) = f(x) + f(y).$$

Now, of course, we can write this equally well as a product,

$$p(z) = p(x)\ p(y),$$

where $p(x) \equiv \exp[f(x)]$ is still an arbitrary continuous monotonic function. It makes no difference which form we choose, but the second choice will prove more convenient later on.

So our rule for finding the plausibility of both A and B takes the form

$$p(AB|C) = p(A|BC)\ p(B|C). \tag{3-3}$$

The condition that this shall represent reasoning qualitatively like ours can tell us something more about this function p(x). For example, let's imagine first that A is certain, given C. What would happen then? Well, if A is certain given C, then in the "environment" produced by knowledge of C, AB and B are the same proposition, in the sense that one is true if and only if the other is true. So, the plausibility that AB is true must be just the plausibility that B is true:

$$(AB|C) = (B|C).$$

And also we would have:

$$(A|BC) = (A|C),$$

because if A is already certain given C, the fact that we may also have B given would not be relevant; it's still certain. To what is our equation (3-3) reduced in this case? It then says

$$p(B|C) = p(A|C)\ p(B|C),$$

and this would have to hold no matter how plausible or implausible B might be. So our function p(x) has to have the property that certainty must always be represented by p = 1.

Now suppose that A is impossible, given C. In this case, the proposition AB is also impossible given C:

$$(AB|C) = (A|C)$$

and if A is already impossible given C, then if we had been given B also, A would still be impossible:

$$(A|BC) = (A|C).$$

In this case, equation (3-3) reduces to

$$p(A|C) = p(B|C)\ p(A|C) \qquad (3\text{-}4)$$

and again this equation would have to hold no matter what plausibility B might have. Well, there are two possible values of $p(A|C)$ that might satisfy this condition. It could be zero or plus infinity. The choice minus infinity can be ruled out [see what happens in (3-4) if B also becomes impossible], but at present there's nothing to tell us to choose zero rather than plus infinity; either one is equally good.

All right, let's sum up what we know about p(x) so far. It is a continuous monotonic function. It may be either increasing or decreasing. If it's an increasing function, it must range from zero for impossibility up to one for certainty; if it's a decreasing function, it must range from one for certainty up to infinity for impossibility. The way in which it varies between these limits, of course, our rule says nothing at all about.

3.2 Deduction of Rules 2 and 3.

Now there are still other conditions of consistency which these rules must satisfy. Let me introduce another notation. By a small letter I'll mean the denial of the big letter. In other words, proposition a stands

for the proposition "A is false." Conversely, A stands for the proposition "a is false." Most of the literature follows the notation of Boole, who indicated denial by placing a bar over the letter. This is fine except that it's a little hard to do reproducibly on a typewriter, so I've taken the liberty of changing it in a way that makes typed notes easier to produce, and less ambiguous to the reader. Actually, we will have little use for this notation beyond the present derivation; so it hardly matters.

Because of the fact that these propositions are of the type which must be either true or false, we see that the logical product aA is always false, and the logical sum a+A will always be true. Now the plausibility of a, given some data B, depends in some reciprocal way on the plausibility of A; if we define $x \equiv p(A|B)$, $y \equiv p(a|B)$, then

$$y = S(x). \tag{3-5}$$

Evidently, if this is going to agree qualitatively with common sense, the function S(x) must be some continuous monotonic decreasing function. But the relation between propositions a and A is a symmetrical one; it doesn't matter which I choose to call a capital letter and which the small letter. I can equally well say that

$$x = S(y). \tag{3-6}$$

It would have to be the same function. So S(x) must satisfy a functional equation that when we apply it twice we get back to where we started:

$$S[S(x)] = x \tag{3-7}$$

Now this alone is not enough to tell us much about this function. It says only that the graph of y = S(x) has mirror reflection symmetry about the line y = x. So now I'd like to give you another argument. There's another condition which S(x) will have to satisfy in order to represent a consistent way of reasoning, and for this we already have one rule of calculation worked out:

$$p(AB|C) = p(B|C)\ p(A|BC) \tag{3-8}$$

We'll call this Rule 1 from now on. Now we can make this step:

$$p(AB|C) = p(B|C)\ S[p(a|BC)]$$

but Rule 1 also says that $p(aB|C) = p(B|C)\ p(a|BC)$, and so

$$p(AB|C) = p(B|C)\ S\left\{\frac{p(aB|C)}{p(B|C)}\right\} \tag{3-9}$$

This looks like a very strange thing to do. But notice that the quantity we started with involved A and B in a symmetric way. If I interchange A and B, I don't change $p(AB|C)$. Therefore, although it doesn't look like it at all, this final expression must also be symmetric in A and B. In other words,

$$p(A|C)\ S\left\{\frac{p(bA|C)}{p(A|C)}\right\} = p(B|C)\ S\left\{\frac{p(aB|C)}{p(B|C)}\right\} \tag{3-10}$$

These two expressions must be equal no matter what propositions A, B, and C are. In particular, they must be equal when the denial of B is the same as the proposition "both A and D are true," that is, when $b = AD$, or

$$B = a + d.$$

But in that particular case, equation (3-10) simplifies. If B has this meaning, then what is $p(bA|C)$? Well, b is the statement that A is true and also that D is true. But this means that $bA = ADA = AD = b$; the propositions bA and b are the same, in the sense that they have the same "truth value." One of them is true if and only if the other is true. Therefore, they must have the same plausibility:

$$p(bA|C) = p(b|C) = S[p(B|C)].$$

Likewise, $aB = a(a+d) = a + ad = a$; in other words, aB and a are the same proposition in the sense that they have the same truth value, and so

$$p(aB|C) = p(a|C) = S[p(A|C)]$$

Substituting these into (3-10), we get a rather awful looking functional equation:

$$x\, S\left[\frac{S(y)}{x}\right] = y\, S\left[\frac{S(x)}{y}\right] \tag{3-11}$$

Here is another functional equation which has to be satisfied in order to have a consistent set of rules for reasoning.

At this point, we will simply turn again to the paper by R. T. Cox (Cox, 1946), or to his later book (Cox, 1961), which solves this problem. He shows that the only twice differentiable function which satisfies all of our conditions is

$$S(x) = (1 - x^m)^{1/m}.$$

and you easily verify that this does satisfy (3-7) and (3-11). This means that our reciprocal relation between the proposition and its denial would then have to take the form

$$p^m(a|B) + p^m(A|B) = 1. \tag{3-12}$$

m can be any constant except zero. I might say that I'm not entirely satisfied with the argument that we went through to get this; not because I think it's wrong, but because I think it's too long. The final result we get is so simple that there must be a simpler way of deriving it; but I haven't found it.

Now suppose that we make the choice that p = 0 is going to represent impossibility. In that case, we'll have to choose m as a positive number in order that (3-12) can be satisfied; but notice that choosing different values of m is really idle, because the only condition on this function p is that it is a continuous monotonic function which increases from zero to one as we go from impossibility to certainty. But if $p_1(x)$ satisfies these conditions, then $p_2(x) \equiv [p_1(x)]^m$ also satisfies them. So the statement that we could use different values of m doesn't give us any freedom that we didn't already have in the fact that p(x) was an arbitrary monotonic function. This means that if I choose to write equation (3-12) in the form

$$p(a|B) + p(A|B) = 1 \qquad (3\text{-}13)$$

this is just as general.

On the other hand, we could represent impossibility by $p = \infty$. In that case, we would have to choose m negative. Once again, to say that we can use different values of m wouldn't say anything that wasn't already implied by the fact that p was an arbitrary monotonic function which increased from one to infinity as we went from certainty to impossibility. So I could equally well write this reciprocal law in the form

$$\frac{1}{p(a|B)} + \frac{1}{p(A|B)} = 1.$$

Now we could go through our entire theory of the design of this robot's brain with the choice of $p = \infty$ to represent impossibility, and we would not get stopped any place. Everything would go through just fine. We would end up with equations which don't look quite so familiar to you as the ones that the other choice will give us. But notice that they're not different theories, because if $p_1(x)$ is a possible choice which goes to plus infinity to represent impossibility, then

$$p_3(x) = \frac{1}{p_1(x)}$$

is a function which represents impossibility by zero, and has all the properties that we needed. So regardless of which choice I make to represent impossibility, it makes the form of equations look different but their content will be exactly the same. You can go from one to the other simply by replacing all p's by the reciprocals of the p's. So if we agree not to use this choice of $p = \infty$ and always to use the choice $p = 0$ to represent impossibility, we're not throwing away any possibility of representation as far as content is concerned. We're just removing a redundancy in how you could have stated the theory. Let us agree, then, to use the choice:

$$0 \le p \le 1.$$

(for impossibility) (for certainty)

You recognize, of course, that this equation (3-13)

$$p(a|B) + p(A|B) = 1$$

which we henceforth call Rule 2, plus our Rule 1

$$p(AB|C) = p(B|C)\ p(A|BC)$$

are actually the fundamental equations of probability theory. Everything in probability theory follows from those by sufficiently complicated arguments. For example, I'd like to get the formula for

$$p(A + B|C),$$

the plausibility that at least one of the propositions A or B would be true, given C. This follows from the rules we already have; we just apply Rule 1 and Rule 2 over and over again:

$$\begin{aligned} p(A + B|C) &= 1 - p(ab|C) \\ &= 1 - p(a|bC)\ p(b|C) \\ &= 1 - [1 - p(A|bC)]\ p(b|C) \\ &= p(B|C) + p(Ab|C) \\ &= p(B|C) + p(b|AC)\ p(A|C) \\ &= p(B|C) + p(A|C)\ [1 - p(B|AC)]. \end{aligned}$$

Finally, we get

$$p(A + B|C) = p(A|C) + p(B|C) - p(AB|C). \qquad (3\text{-}14)$$

At long last we come out with the above form. And it's this result that I will take as our Rule 3.

We can summarize what we have learned up to this point by writing down our fundamental rules:

Rule 1: $p(AB|C) = p(A|BC)\ p(B|C) = p(B|AC)\ p(A|C)$ (3-15)

Rule 2: $p(A|B) + p(a|B) = 1$ (3-16)

Rule 3: $p(A + B|C) = p(A|C) + p(B|C) - p(AB|C)$ (3-17)

Rule 1, of course, involves A and B in a symmetric way and we could have interchanged A and B in all the argument leading up to it, so we have the liberty of writing it with A and B interchanged, as shown.

3.3 Deduction of Rule 4.

We've found so far the most general consistent rules by which our robot can manipulate plausibilities, granted that he must associate them with real numbers in some way so that his brain can operate by the carrying out of a definite physical process, and we are encouraged by the familiar appearance of these rules. But there are two evident circumstances which show that our job isn't yet finished. In the first place, while Rules 1, 2, and 3 show 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'd be a different set of rules.

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 his calculations. How is the robot to make his initial encoding of the given information, into definite numerical values of plausibilities?

The following analysis answers both of these questions, in a way that I think you will find both interesting and unexpected. Let's 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 Rule 3, as follows. The first application gives

$$p(A_1+A_2+A_3|B) = p(A_1+A_2|B) + p(A_3|B) - p(A_1A_3 + A_2A_3|B)$$

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

logical relation $(A_1+A_2)A_3 = A_1A_3 + A_2A_3$. Applying Rule 3 again to the first and third of these expressions, we obtain seven terms which can be grouped as follows:

$$\begin{aligned} p(A_1+A_2+A_3|B) &= p(A_1|B) + p(A_2|B) + p(A_3|B) \\ &\quad - p(A_1A_2|B) - p(A_2A_3|B) - p(A_3A_1|B) \\ &\quad + p(A_1A_2A_3|B) \end{aligned} \tag{3-18}$$

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

$$p(A_iA_j|B) = p(A_i|B)\delta_{ij} \tag{3-19}$$

where δ_{ij} is the Kronecker delta

$$\delta_{ij} \equiv \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} .$$

If the A_i are mutually exclusive, then the last four terms of (3-18) vanish, and we have

$$p(A_1+A_2+A_3|B) = p(A_1|B) + p(A_2|B) + p(A_3|B) \tag{3-20}$$

Adding more propositions A_4, A_5, etc., it is easy to show by induction that if we have n mutually exclusive propositions $\{A_1 \ . \ . \ . \ A_n\}$, (3-20) generalizes to

$$p(A_1+ \ . \ . \ . \ +A_m|B) = \sum_{i=1}^{m} p(A_i|B), \quad m \leq n \tag{3-21}$$

a rule which we will be using constantly from now on. In conventional expositions, Eq. (3-21) is usually introduced directly as one of the basic axioms of the theory, without any attempt to demonstrate its uniqueness or consistency. The present approach shows that this rule is deducible from simpler relations, which in essence represent the conditions for this theory to be consistent in the sense (a) given in Sec. 2.3.

Now suppose that the propositions $\{A_1 \ . \ . \ . \ A_n\}$ are not only mutually exclusive but also exhaustive; i.e., on data B one and only one of them

must be true. In that case the sum (3-21) over all of them must be unity:

$$\sum_{i=1}^{n} p(A_i|B) = 1 \tag{3-22}$$

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 $p(A_i|B)$ can be a difficult problem.

There is, however, one case in which the answer is particularly simple, requiring only direct application of principles already given. But we are now entering a very delicate and crucial area which has caused trouble and controversy for over a century; so I ask that you suppress all intuitive feelings that you may have, and contemplate the following logical analysis minutely. The point we are about to make cannot be developed too carefully; and unless it is clearly understood, you 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 $\{A_1 \ . \ . \ . \ A_n\}$ and we seek to evaluate $p(A_i|B)$. 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 $\{A_1' \ . \ . \ . \ A_n'\}$, given by

$$\begin{aligned} A_1' &= A_2 \\ A_2' &= A_1 \\ A_k' &= A_k, \quad k \geq 3 \end{aligned} \tag{3-23}$$

and we seek to evaluate the quantities $p(A_i'|B)$, $i = 1, 2, \ . \ . \ ., n$.

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, he must have the same state of knowledge about A_2' in problem II, for they are the same proposition, his given information B is the same in both problems, and he is contemplating the same totality of propositions $\{A_1 \ . \ . \ . \ A_n\}$ in both problems. Therefore we must have

$$p(A_1|B)_I = p(A_2'|B)_{II} \tag{3-24}$$

and similarly

$$p(A_2|B)_I = p(A_1'|B)_{II} \tag{3-25}$$

We will call these the transformation equations. What we have just done may appear utterly trivial to you, but bear with me; this line of reasoning, as Professor Eugene Wigner has aptly remarked (Wigner, 1959), consists of a number of steps each of which appears trivial in itself, but which in their totality are far from trivial. At this point, note that the transformation equations (3-24), (3-25) 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., it gives the robot no reason to prefer either over the other. In this case, problems I and II are entirely equivalent; i.e., he is in exactly the same state of knowledge about the set of propositions $\{A_1' \ . \ . \ . \ A_n'\}$ in problem II, including their labeling, as he is about the set $\{A_1 \ . \ . \ . \ A_n\}$ in problem I.

Now we invoke our requirement of consistency in the sense (b) as given above (Sec. 2.3). This stated that, in two equivalent problems, where the robot has the same state of knowledge, he must assign the same plausibilities. In equations, this statement is

$$p(A_i|B)_I = p(A_i'|B)_{II}, \quad i = 1,2,\ldots,n \tag{3-26}$$

which we will call the equivalence equations. But now, combining equations

(3-24), (3-25), (3-26), we obtain

$$p(A_1|B)_I = p(A_2|B)_I \tag{3-27}$$

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 (3-27). You recall that I asked you to suppress whatever intuitive feelings you may have, and allow yourself to be guided solely by the logical analysis. We will discuss the reasons for this presently; but first let us extend the result. More generally, let $\{A_1" \ . \ . \ . \ A_n"\}$ be any permutation of $\{A_1 \ . \ . \ . \ A_n\}$ and let Problem III be that of determining the $p(A_i"|B)$. If the permutation is such that $A_i = A_k"$, there will be n transformation equations of the form

$$p(A_i|B)_I = p(A_k"|B)_{III} \tag{3-28}$$

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 $\{A_1" \ . \ . \ . \ A_n"\}$ in problem III as he was about the set $\{A_1 \ . \ . \ . \ A_n\}$ in problem I; and again our desideratum of consistency demands that he assign equivalent distributions in equivalent problems, leading to the n equivalence equations

$$p(A_k|B)_I = p(A_k"|B)_{III}, \quad k = 1, 2, \ldots, n \tag{3-29}$$

From (3-28) and (3-29) we obtain n equations of the form

$$p(A_i|B)_I = p(A_k|B)_I \tag{3-30}$$

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 in which, for given i, the index k will range over all of the (n-1) others in (3-30). 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). Since the $\{A_1 \ldots A_n\}$ are exhaustive, Eq. (3-22) will hold, and the only possibility is therefore

$$p(A_i|B)_I = \frac{1}{n}, \quad i = 1, 2, \ldots, n \tag{3-31}$$

and we have finally arrived at a set of definite numerical values. We will call this result Rule 4.

Perhaps you intuition had already led you to just this conclusion, without any need for the rather tortuous reasoning we have been through. If so, fine; then your intuition is consistent with our axioms. But merely writing down (3-31) intuitively does not give one a full appreciation of the importance and uniqueness of this result.

To see this importance, note that Eq. (3-31) 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 $p(A_i|B)$, and <u>not</u> the numerical values of the plausibilities $(A_i|B)$ that we started with. This, also, will be found to be true in general. But recognizing this gives us a beautiful answer to the first question posed at the beginning of this Section; after having found Rules 1, 2, and 3 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.

But now we see that no matter what function $p(x)$ we choose, we would still be led to the same result (3-31), 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 Rules 1, 2, and 3 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 you could design the robot's memory circuits. For each proposition A_i about which he is to reason, he will need a storage register in which he enters some number representing the degree of plausibility of A_i, on the basis of all the data he has been given. Of course, instead of storing the number p he could equally well store any monotonic function of p. But no matter what function he used internally, the externally observable behavior of the robot would be exactly 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 is an arbitrary monotonic function of p, defined in the interval $0 \leq p \leq 1$; it is p that is rigidly fixed by the data of a problem. The question of uniqueness is therefore disposed of automatically by the result (3-31); 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 $x \equiv (A|B)$ that we started with 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 name; from now on, we will call these quantities probabilities. I have studiously avoided using the word "probability" in our derivations up to this point, because while the word does have a colloquial meaning to the "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 such unique 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.

This situation is analogous to that in thermodynamics, where out of all possible temperature scales, which are monotonic functions of each other, we finally decide to use the Kelvin scale; not because it is more "correct" than others but because it is more convenient; i.e., the laws of thermodynamics and statistical mechanics take the simplest form in terms of this particular temperature scale.

3.4 Philosophical Digression.

For historical reasons, we still need quite a long discussion of Rule 4, Eq. (3-31). There seem to be only two kinds of people working in probability theory: those who consider Rule 4 to be so utterly trivial and obvious as to be in no need of any proof; and those who regard it as such a foolish and unjustified piece of metaphysical nonsense as to discredit anyone who uses it.

As far as I have been able to determine, there is no middle ground between these opinions; in the past, every writer on probability theory has been an extremist on one side or the other. I myself was an extremist of the first genre for some twenty years, and it was only recently that more mature reflection finally made me realize that Rule 4 is in need of logical demonstration. More important, it now appears to me that the method

of reasoning we have used to find Rule 4 is fundamental to all of probability theory, almost every present application requiring it to give a full logical justification of the result.

The reasoning we have just used is the most rudimentary example of the general group-theoretical approach which has been used with great success in theoretical physics for some forty years (Wigner, 1959). I had been teaching the use of group-theoretical methods for finding solutions of differential equations and boundary-value problems for sixteen years, without realizing that this same technique is the key to several deep unresolved issues in probability theory.

Rule 4 is itself fundamental to all of probability theory; although some will deny it, I don't think I am exaggerating when I assert that there is no known application of probability theory in which Rule 4 is not needed at one place or another. Those who profess to dislike it merely find some way of disguising the fact that they are using it; I will cite some specific examples in a later lecture. To understand this, we have to study the history of probability theory.

Rule 4 appears to have been first stated explicitly by James Bernoulli at the end of the seventeenth century (although it was, of course, implicit in the still earlier work of Cardano and Pascal). In the old literature it is often called the "Principle of Insufficient Reason," and it was used and defended by Laplace on the grounds that, on the given information, there was "no reason to think otherwise." This terminology and reasoning have been most unfortunate--I am tempted to say tragic--for the development of probability theory, because it has created a psychological block which has prevented many from seeing the real point of Rule 4.

But note that, in view of our derivation, we are asserting the validity of Rule 4, not for the weak and negative reason given by Laplace, but for

the strong and positive reason that it is uniquely determined by elementary requirements of consistency. In the state of knowledge defined by B in (3-31), if the robot were to assign any probability distribution other than the uniform one, then by a mere permutation of labels we could exhibit a second problem in which the robot has exactly the same state of knowledge, but in which he is assigning a different probability distribution. It just would not make sense, then, to say that the distribution described the robot's state of knowledge, or to claim that he is behaving in a consistent way.

But there is still a mystery here. For, no matter what method of reasoning we use, how is it possible that otherwise rational and mathematically competent people could be in violent disagreement on such an apparently simple matter as Equation (3-31)? I think that we have been caught in a semantic trap of our own making; to explain this, let me try to state the position of both extremists.

The extremist of the first camp says, "If the information B gives the robot no reason to prefer any of the propositions A_i over any other, then these propositions must appear equally likely to him; there is obviously no other thing he can possibly do but to assign them equal probabilities by Eq. (3-31). To do anything else would be to jump to conclusions not warranted by the data."

The extremist of the second camp says, "If the information B merely gives the robot no reason to prefer any proposition over another, this provides absolutely no justification for supposing them to be equally likely; they might not be equally likely at all. Unless the information B contains positive evidence that they <u>are</u> equally likely, the problem is simply not well-posed; and to write Eq. (3-31) is to jump to conclusions not warranted by the data."

Perhaps I have not, in spite of some effort, managed to verbalize these two positions in the most felicitous way; but I think you will grant that a more expert verbalizer could make either of these positions seem highly convincing, so at least from a psychological standpoint we can understand how there can be two diametrically opposing camps on this issue.

But, to be more constructive, what is the source of the difference? If you study these two statements, I think you'll agree that it is semantic; the phrase "equally likely" has two entirely different meanings in the two camps. In camp 2, the statement, "A_1 and A_2 are equally likely" is taken to describe a property <u>of the propositions</u> which is either true or false in an objective sense independently of the state of knowledge you or I--or the robot--might have about them. With that interpretation, of course, we have no justification for assuming this property to exist unless there is positive evidence for it.

In camp 1, the statement, "A_1 and A_2 are equally likely" is not regarded as describing any property of A_1 and A_2. In fact, each proposition is, in an objective sense, either true or false; and the only reason for using probability theory is that we are not in a position to say which. In writing Eq. (3-31), we are asserting nothing whatever about the propositions; we are describing only the <u>state of knowledge of the robot</u>.

Now you can, if you like, make value judgments as to which of these interpretations is the more desirable. But this has already been done quite enough to show that arguments on that level are futile. Debate on this issue has been going on more or less furiously in the literature of probability theory since the time of Laplace, one camp and then the other gaining a momentary ascendancy in numbers. But I think you will agree that we have here an issue that can never be settled by philosophical arguments about the meaning of words; much less by taking votes. We are in a situation very

much like the scientist who must decide between two rival theories of physics; and it has taken the human race thousands of years to realize that the only real, objective criterion for deciding such matters is the pragmatic one: casting aside all philosophical or ideological considerations, which viewpoint leads to a theory with the widest range of useful applications?

Therefore, I don't intend to waste any more time on the issue at this point; it is a major objective of these lectures to examine the problem on just the above pragmatic grounds. We are going to study a wide range of problems, covering almost all present applications of probability theory; and whenever possible we will exhibit the actual calculations, and final results, that the two viewpoints lead to.

It is perhaps already clear that viewpoint 1 is more widely applicable; there are many problems which our robot can undertake at once starting from Rule 4, but which on viewpoint 2 are ill-posed, offering no basis for applying probability theory. Now of course, a human statistician belonging to camp 2 may simply refuse to work on a problem (possibly at the cost of his job) if the information available is not as complete as he would like; but our robot is not free to do this, because the whole point of designing him is that he is to do the best he can whatever the information at hand. The issue will then be: in such problems, does the robot arrive at useful and defensible conclusions?

Of course, if the given information is too vague to justify any definite conclusions, we will want the robot to recognize this and tell us that more data are needed. His way of doing this will be to give us a final probability distribution that is very broad, indicating no strong preference for one conclusion over another. If the data do justify definite conclusions, he will find very sharply peaked final distributions, and report, "The data you gave me point to conclusion C as overwhelmingly the most likely to be

correct." And, of course, the robot should have some way of interpolating between these extremes, where most of the really interesting problems of the theory lie.

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. But it is just the function of our consistency requirements to make these probability assignments completely "objective" in the sense that they are independent of the personality of the user; i.e., they are a means of describing (or if you like, of encoding) the given information, independently of whatever personal feelings you or I might have. It is "objectivity" in this sense that is needed for a scientifically respectable theory of plausible reasoning.

The job before us now is, therefore, not to engage in philosophical disputation, but to put our robot to the test by examining just what he will do if he reasons by applying Rules 1 - 4 and their generalizations that we will develop as needed.

Lecture 4

BAYES' THEOREM AND MAXIMUM LIKELIHOOD

From now on, instead of writing $p(A|B)$, I will often leave off the p, and write it simply as $(A|B)$. You can interpret this two ways. You can say I'm changing my notation; since it's always the function p that we're concerned with, I'll simply understand that it's always that function that is meant. Or, since it was an arbitrary function anyway, you can say that I've now adopted the convention that

$$p(x) \equiv x$$

by definition. It will make no difference at all which way you interpret this. Our fundamental rules of reasoning will then take the form:

Rule 1: $$(AB|C) = (A|BC)(B|C) = (B|AC)(A|C) \tag{4-1}$$

Rule 2: $$(A|B) + (a|B) = 1 \tag{4-2}$$

Rule 3: $$(A+B|C) = (A|C) + (B|C) - (AB|C) \tag{4-3}$$

Rule 4: If $\{A_1 \ldots A_n\}$ are mutually exclusive and exhaustive, and B does not favor any over any other, then

$$(A_i|B) = \frac{1}{n} \quad , \; i=1, 2, \ldots n. \tag{4-4}$$

4.1 Prior Probabilities.

Now out of all the propositions that this robot has to think about, there is one which is always in his mind. By X I mean all of his past experience since the day he left the factory to the time he started reasoning on the problem he's thinking about now. That is always part of the information

which is available to him, and obviously it would not be consistent for him to throw away what he knew yesterday in reasoning about his problems today. If human beings did that, education and civilization would be impossible. So for this robot there is no such thing as an "absolute" probability. All probabilities are conditional on X at least. X might be irrelevant to some problem and in that case this postulate would be unnecessary, but at least harmless. If it's irrelevant, it will cancel out mathematically. Any probabilities which are conditional on X alone we will call prior probabilities. If there is any additional evidence in addition to X, which the robot is now reasoning on, we will sometimes leave off the X. We'll understand that even when we don't write X explicitly, it's always built into all expressions:

$$(A|B) \equiv (A|BX) \quad .$$

But in a prior probability, I'll always put in X explicitly:

$$(A|X) \quad .$$

Because of some strange things that have been written about prior probabilities in the past, we have to point out that it would be a big mistake to think of X as some sort of hidden major premise, some universally valid proposition about nature, or anything of that sort. X is simply whatever initial information the robot had available up to the time we gave him his current problem. When we consider applications, you can think also that X stands for some set of hypotheses whose consequences we want to find out, plus the general conditions specified or implied in the statement of the problem.

4.2 Bayes' Theorem.

By far the most important rule which this robot uses in his everyday tasks is the one we get by dividing through the second equality of Rule 1

by, say, $(B|C)$:

$$(A|BC) = (A|C)\,\frac{(B|AC)}{(B|C)} \tag{4-5}$$

This is called Bayes' theorem, or the principle of inverse probability. You see it represents the process by which the robot learns from experience. He starts out with the probability of A, on the basis of evidence C; he is given new evidence B in addition, and this theorem tells how the probability of A changes as a result of this new evidence. Bayes' theorem comes from the fact that Rule 1 was symmetric in propositions A and B, which of course it had to be in order to be consistent. To this robot it is quite clear that if he wants to make any judgments about the truth of proposition A, the only correct way to do this is to calculate the probability of A, conditional on all the evidence he has. This will almost always mean that he will have to use Bayes' theorem.

Now let's imagine we let this robot examine some procedures that are used in statistical inference. A very large part of statistical inference is taken up with problems in which we are given certain evidence, which is typically the result of some experiment, and from this evidence we are supposed to do the best job we can of estimating some unknown parameter, or testing one hypothesis against another. All of these represent plausible reasoning on the basis of new evidence; the evidence of the experiment. Therefore, to our robot it's perfectly obvious that any such example of parameter estimation or hypothesis testing must be a special case of the application of Bayes' theorem. You see, his brain has been built so that this is the only possible way he can reason. To him, the fact that all these procedures must derive from Bayes' theorem is just as much a necessity of thought as the validity of a strong syllogism is to us.

Although this conclusion about Bayes' theorem is obvious to our robot,

it has not been at all obvious to most human statisticians. They largely regard Bayes' theorem as not having any logical basis except in the case where every probability in it can be interpreted as a relative frequency in some "random experiment." In that case, Bayes' theorem can be interpreted as selecting out of an original population of events some subpopulation in which the frequency of event A might be different from the frequency that it has in the population as a whole. But to the robot this is the only possible way of reasoning regardless of whether you can give the probabilities a frequency interpretation.

To a statistician of the "orthodox" school of thought, to be defined more completely later, the first thing he must do in solving a problem is to decide which quantities are "random," and which are not; the procedures he will use, and the whole way he will set up the problem, depend on which decision he makes. But our derivation of the rules for plausible reasoning in the last Lecture made no reference whatsoever to any random experiment. To the robot, therefore, whether any random experiment is or is not involved in the problem is totally irrelevant to the question of how he should reason.

Since this is perhaps <u>the</u> crucial issue in the controversies about probability theory, and the central point in most of the applications that I want to talk about later, we have got to meet it squarely right now. So let's ask the robot to make a strong, definite, and constructive statement about it. Here's what he has to say:

"Consider any procedure in statistical inference in which we reason on the basis of new information. If this procedure is fully consistent and in full qualitative agreement with common sense, then it is necessarily <u>exactly</u> derivable from Bayes' theorem. Conversely, if it is found to represent only some approximation to Bayes' theorem, then it follows that

(1) It is either inconsistent or it does qualitative violence to common sense, or both;

(2) These shortcomings can be exhibited by producing special cases; and

(3) Bayes' theorem will then represent a superior (and often simpler) way of handling the problem."

That is what the robot says. We've designed him in just such a way that it's the only thing he can say. It doesn't mean at all that what he says is right. We've got to put him to the test. For each particular procedure, this is a definite issue of fact; and not a vague matter of personal taste. Either the robot is right or he's wrong in the above statement, and it's in our power to find out whether he's right or wrong. So we'll browse through the statistical literature, and every time we see an example where the man says, "I'm not using Bayes' theorem," then we can look at it a little more carefully and see whether what he actually does can be derived from Bayes' theorem; and if not, whether we can exhibit the defects in his procedure.

4.3 Maximum Likelihood.

The first example is Sir Ronald A. Fisher's method of maximum likelihood. This is a way of estimating an unknown parameter, and I'll illustrate it by the problem of estimating the magnitude of a signal which is obscured by noise. You might be interested in some quotations from Fisher's book (Fisher, 1959). On page 9, he refers to "...my personal conviction which I have sustained elsewhere, that the theory of inverse probability is founded upon an error, and must be wholly rejected" (inverse probability and Bayes' theorem are the same thing as far as we're concerned). And later on he says on page 20 that "maximum likelihood has no real connection with

inverse probability." Well, let's illustrate the method. Suppose we have observed a voltage just at one instant, which is the sum of an unknown signal plus an unknown noise:

$$V = S + N \tag{4-6}$$

Our prior knowledge about the nature of the noise can be described by some probability distribution; the probability that the noise amplitude is in the range dN is

$$(dN|X) = W(N)dN \tag{4-7}$$

Now if we knew that the signal had a certain value S, then the probability of observing a voltage in the range dV would be given by some relation of the form

$$(dV|SX) = L(V,S)dV \tag{4-8}$$

where L(V,S) is called the likelihood function. In the present case, from the linearity of Eq. (4-6), this must be just the probability that the noise would have made up the difference; and so

$$L(V,S) = W(V-S). \tag{4-9}$$

But in the given problem, it's the voltage that's known and the signal that's unknown. The maximum likelihood estimate of the signal magnitude would then be the value of S which renders this likelihood function L an absolute maximum for the observed value of V:

$$\frac{\partial L}{\partial S} = 0 \quad , \qquad \frac{\partial^2 L}{\partial S^2} < 0. \tag{4-10}$$

Stated intuitively, the maximum likelihood estimate is the value according to which the observed voltage would appear as the least remarkable coincidence.

How would our robot go about handling this problem? To him the way of reasoning about the unknown signal is, of course, to calculate the probability that the signal has a certain amplitude, on the basis of all the avail-

able evidence. In other words, the robot says we should calculate $(dS|VX)$ by Bayes' theorem:

$$(dS|VX) = (dS|X)\,\frac{(dV|SX)}{(dV|X)}$$

$$= A\ (dS|X)\ L(V,S) \qquad (4\text{-}11)$$

where A is independent of S. So if we ask the robot what is the most probable value of the signal [more precisely, for what value of S is it most probable that the signal lies in the interval (S, S+dS) for a fixed dS], he will maximize not L but the product of L with the prior probability. So you see that if the robot's prior information didn't give him any reason to expect one signal magnitude more than another [i.e. if the prior probability $(dS|X)$ is independent of S in the range of interest], then the robot's estimate would be the same as the maximum likelihood estimate. If the robot has prior information about the signal, then of course he may easily get a very different value.

Now I think it's obvious not only to the robot, but also to us, that if we do have any prior information about the signal, then it would be screamingly inconsistent for us to refuse to take that information into account in estimating the magnitude of the signal. You see, we could describe the maximum likelihood estimate in another way as the value which we would obtain by <u>throwing away</u> all the prior information we had about the signal, and basing our estimate <u>only</u> on our prior information about the noise.

Suppose you went to a doctor and described your symptoms, and you wanted him to diagnose what was wrong. You tell him that when you raise your left arm you feel a pain in your right side and a few things like this, and the doctor is supposed to do some plausible reasoning to figure out what could be causing it. Suppose that after consultation had been underway for some time you notice that the doctor is not showing any interest in your

previous medical history. You ask him, "Well, aren't you going to look up my medical history?" And suppose the doctor said, "Why, no, I must not look at your medical history, because that would introduce a bias into my conclusions." What would you say? You'd say that the man is crazy. He shouldn't be allowed to practice medicine. To refuse to take the prior information you have into account in plausible reasoning, is not a consistent way of doing things.

Now, of course, a human statistician who uses maximum likelihood has just as much common sense as anybody else; and in a case where we do have prior information which is clearly relevant to the problem, common sense will tell all but the most pedantic not to use the method of maximum likelihood. In practice, he will avoid the bad errors of reasoning by inventing a different method when a different kind of problem comes up. In other words, he will use his prior information to tell him how to formulate the problem,* and he prefers to formulate it so this information no longer appears explicitly in his equations. The robot, however, doesn't need to invent a new procedure for every new kind of problem. To him, Bayes' theorem is always the only way of doing it.

I don't want to go into more details now because this is close to a problem which we are going to talk about a great deal later on; but for the present we'll just note that the robot's prediction was correct. Except in the case where it's clearly inconsistent, the method of maximum likelihood is exactly derivable from Bayes' theorem. After all polemics, there remains the simple fact that, mathematically, it is nothing but Bayes' theorem with uniform prior probability.

---

*An example of such a reformulation suggested by prior information is given in Lecture 9, Equations (9-18)-(9-22).

Lecture 5

SEQUENTIAL HYPOTHESIS TESTING

Our second example of statements made about Bayes' theorem in the literature has been provided by Professor Wm. Feller. On page 85 of his book (Feller, 1950) he writes: "Unfortunately Bayes' rule has been somewhat discredited by metaphysical applications of the type described above.* In routine practice, this kind of argument can be dangerous. A quality control engineer is concerned with one particular machine and not with an infinite population of machines from which one was chosen at random. He has been advised to use Bayes' rule on the grounds that it is logically acceptable and corresponds to our way of thinking. Plato used this type of argument to prove the existence of Atlantis, and philosophers used it to prove the absurdity of Newton's mechanics. In our case it overlooks the circumstance that the engineer desires success and that he will do better by estimating and minimizing the sources of various types of errors in predicting and guessing. The modern method of statistical tests and estimation is less intuitive but more realistic. It may be not only defended but also applied."

Well, that gives us a pretty clear idea of one common attitude toward Bayes' theorem, at least for problems of quality control. Now what are the procedures referred to as the "modern method of statistical tests?" I can't tell of course from reading, but ever since the early days of World War II

---

*The reference is to Laplace's law of succession, about which we will have a lot to say later on in Lecture 16.

when he invented it, Wald's sequential testing procedure (Wald, 1947) has been generally considered the optimum one available, optimum according to several different criteria.

Let's illustrate the problem by considering manufacture of some small item. Suppose we take crystal diodes. One of the important things about a crystal diode is the maximum inverse peak voltage it can stand without damage. Clearly, the way to find out just how good our diodes are is to test each one and measure the voltage at which damage occurs. The trouble is that once we've done this the diode is ruined, so we can't test every one this way. We can test only some fraction of the batch and we would not want to test a very large fraction. So the problem of quality control in this case is to find some method of plausible reasoning which lets us do the best possible job of deciding whether we have a good batch or not, with the smallest number of diodes ruined in testing. I think all statisticians agree that Wald's method is the optimum one in this sense of requiring, on the average, fewer tests than any other for a given probability of error. Wald, in a footnote in his book, says that he conjectures that it's an optimum test in this sense but didn't succeed in proving it. We'll come back to that statement a little later.

Just for variety, let's go first into the way the robot would handle this problem. We will simply ignore Feller's warning, and see for ourselves whether Bayes' theorem can be "applied." After the final comparisons are at hand, we will also see whether it can be "defended."

5.1 Logarithmic Form of Bayes' Theorem.

First, let's manipulate Bayes' theorem a little bit in a manner suggested by I. J. Good (Good, 1950). Instead of calculating the probability, it would be just as good if we'd calculate any monotonic function of the

probability, if we know what function we've got. So, let's do a little rebuilding on Bayes' theorem. I'll use E to stand for new evidence.

$$(A|EX) = (A|X)\,\frac{(E|AX)}{(E|X)}\quad . \tag{5-1}$$

Now we could have written Bayes' theorem for the probability that A is false given the same evidence,

$$(a|EX) = (a|X)\,\frac{(E|aX)}{(E|X)}\quad , \tag{5-2}$$

and we can take the ratio of the two equations:

$$\frac{(A|EX)}{(a|EX)} = \frac{(A|X)\,(E|AX)}{(a|X)\,(E|aX)}\quad . \tag{5-3}$$

In this case, one of our terms will drop out. This doesn't look like any particular advantage. But the quantity that we have here, the ratio of the probability that A is true to the probability that it's false, has a technical name. We call it the "odds" on the proposition A. So if I write the "odds of A, given E and X," as the symbol

$$O(A|EX) \equiv \frac{(A|EX)}{(a|EX)} \tag{5-4}$$

then I can write Bayes' theorem in the following form:

$$O(A|EX) = O(A|X)\,\frac{(E|AX)}{(E|aX)} \tag{5-5}$$

The odds on A are equal to the prior odds multiplied by the ratio of the probability that E would be seen if A was true, to the probability that E would be observed if A was false. The odds are, of course, a monotonic function of the probability, so we could equally well calculate these quantities.

In some applications it is even more convenient to take the logarithm of the odds because of the fact that we can then add up terms--the same reason the logarithm was invented in the first place. Now we could take logarithms to any base we want. What I'm after here is something which is handy for numerical work, and the base 10 turns out to be easier to use

than the base e for that purpose, even though it makes our equations look less elegant. And so I'm going to define a new function which I'll call the evidence for A given E:

$$e(A|EX) \equiv 10 \log_{10} O(A|EX) \ . \tag{5-6}$$

This is still a monotonic function of the probability. By using the base 10 and putting the factor 10 in front, we've now reached the condition where we're measuring evidence in decibels! Now what does Bayes' theorem look like? The evidence for A, given E, is equal to the prior evidence plus the number of db provided by working out the probability ratio in the second term below:

$$e(A|E) = e(A|X) + 10 \log_{10}\left[\frac{(E|A)}{(E|a)}\right] . \tag{5-7}$$

Now let's suppose that this new information that we got actually consisted of several different propositions:

$$E = E_1 E_2 E_3 \ldots$$

In that case, we could expand this a little more by successive applications of Rule 1:

$$e(A|E) = e(A|X) + 10 \log_{10}\left[\frac{(E_1|A)}{(E_1|a)}\right] + 10 \log_{10}\left[\frac{(E_2|E_1A}{(E_2|E_1a}\right] + \ldots \tag{5-8}$$

In a lot of cases, it turns out that the probability of E_2 is not influenced by knowledge of E_1. For example, in the case where one says technically the probability is a chance; say the tossing of a coin, where knowing the result of one toss (if you know the coin is honest) doesn't influence the probability you would assign for the next toss. In case these several pieces of evidence are independent, the above equation becomes:

$$e(A|E) = e(A|X) + 10 \sum_i \log_{10}\left[\frac{(E_i|A)}{(E_i|a)}\right] \ , \tag{5-9}$$

where the sum is over all the extra pieces of information we get.

Now it would be a good idea for us to get some feeling for numerical values here. So, I'd like to give a table and a graph. We have here three different ways we can measure plausibility; evidence, odds, or probability; they're all monotonic functions of each other. Zero db of evidence corresponds to odds of 1 or to a probability of 1/2. Now every electrical engineer knows that 3 db means a factor of 2 and 10 db is a factor of 10, and so if we just go up in steps of 3 db, or 10, why we can write down this table pretty fast.

e	O	p
0	1:1	1/2
3	2:1	2/3
6	4:1	4/5
10	10:1	10/11
20	100:1	100/101
30	1000:1	0.999
40	10^4:1	0.9999
-e	1/O	1-p

You see here why giving evidence in db is nice. When probabilities get very close to one or very close to zero, our intuition doesn't work very well. Does the difference between the probability of 0.999 and 0.9999 mean a great deal to you? It certainly doesn't to me. But after living with this for a while, the difference between evidence of plus 30 db and plus 40 db <u>does</u> mean something to me. It's now in a scale which my mind can comprehend. This is just another example of the Weber-Fechner law. Now let's draw a graph showing reasonably well the numerical values of evidence versus probability. This graph is shown in Figure (5.1). The graph is symmetric about

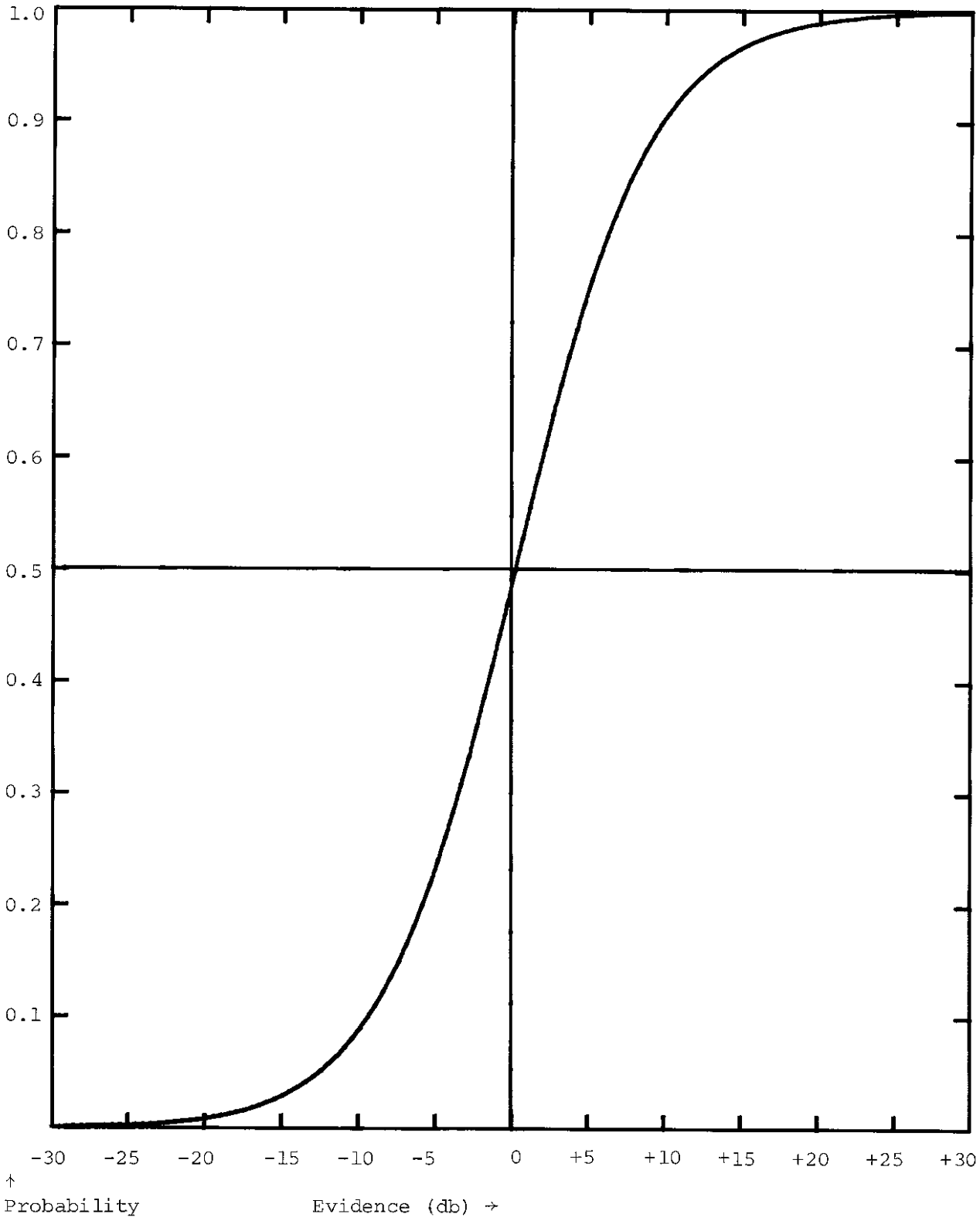

Figure 5.1. Probability vs. Evidence.

the center.

Now let's take our specific example of quality control. I'll assume numbers which are not at all realistic in order to bring out some points a little bit better. We have eleven automatic machines which are turning out crystal diodes. This example corresponds to a very early stage in the development of crystal diodes, because ten of the machines produce, on the average, one in six defective. The eleventh machine is even worse; it makes one in three defective. The output of each machine is collected in an unlabeled box and stored in the warehouse. We choose one of the boxes and we test a few of the diodes. Our job is to decide whether we got a box from the bad machine or not; that is, whether we're going to accept this batch or reject it. Now we're going to turn this job over to our robot and see how he handles it.

He says: "If we want to make judgments about whether we have the box of defective diodes, the way to do this is to calculate the probability that we have the box of defective diodes, conditional on all the evidence available." Let's say the proposition A shall stand for the statement "we chose the bad box." All right, what is the initial evidence for proposition A? The only initial evidence is that there are eleven machines and we don't know which one we got; so by Rule 4 $(A|X) = 1/11$, and by Rule 2 $(a|X) = 1 - (A|X) = 10/11$. Therefore,

$$e(A|X) = 10 \log_{10} \frac{(A|X)}{(a|X)} = 10 \log_{10} \frac{1/11}{10/11}$$

$$= -10 \text{ db} \qquad (5\text{-}10)$$

Evidently, the only property of X that's going to be relevant to this problem is just this number, - 10 db. Any other kind of prior evidence which led to the same initial probability assignment would give us exactly the same mathematical problem from this point on. So, it isn't really necessary

to say we're talking only about a problem where there are eleven machines, and so on. There might be only one machine, and the prior evidence consists of our previous experience with it. My reason for stating the problem in terms of eleven machines was just that we have, so far, only one principle, Rule 4, by which we can convert raw information into numerical values of probability. I mention this here only because of Professor Feller's remark about a single machine. To our robot, it doesn't make any difference how many machines there are; the only thing that counts is the prior probability, however arrived at.

Now from this box we take out a diode and test it to see where it breaks down. Every time we pull out a bad one, what will that do to the evidence? That will add to this the number

$$10 \log_{10} \frac{(\text{bad}|A)}{(\text{bad}|a)} \tag{5-11}$$

where $(\text{bad}|A)$ represents the probability of getting a bad diode, given A, etc. We have, then, to determine these probabilities.

If we have the box in which one in three are bad, what is the probability that we will draw a bad one? The final answer is obvious to all of us without any calculation, and the argument showing this from the principles of probability theory is almost trivial. Nevertheless, I want to give that argument in full because there is a very important general principle lurking here, which will apply in countless other applications of probability theory.

5.2. Sampling With and Without Replacement.

Consider first the traditional "urn" of probability theory, in which we have placed N balls, all of the same size, weight, surface texture, etc., labeled 1, 2, ..., N. Balls 1, 2, ..., n are black, and the remaining (N-n) are white. What is the probability of drawing blindfolded any parti-

cular ball, say the i'th? Rule 4 answers this, for there are N mutually exclusive possibilities, and the information given provides no justification for expecting any one of them in preference to any other. In this state of knowledge, therefore, the probability sought must be $p_i = 1/N$.

Let us recall clearly just what this means. The probability assignment $p_i = 1/N$ is not an assertion of any physical property of the balls; it is merely a means of describing the state of knowledge of the robot prior to the drawing. It is, therefore, utterly meaningless to speak of "verifying" this probability assignment by performing any experiment on the balls; that would be exactly like trying to verify a boy's love for his dog by performing experiments on the dog. What it does mean was explained in our derivation of Rule 4; the assignment $p_i = 1/N$ is uniquely determined by the requirement that the robot's reasoning be consistent in the sense that, in two problems where he has the same state of knowledge, he must assign the same probabilities. If he were to assign anything different from the uniform distribution, then merely by a permutation of labels we could exhibit a second problem in which the robot's state of knowledge is exactly the same; but in which he is assigning a different probability distribution. I have repeated this argument for emphasis, because to the best of my knowledge, this point is not recognized in any other work on probability theory.

Now, what is the probability that we shall draw a black ball? Since different balls are mutually exclusive possibilities, Rule 3 as extended to Eq. (3-21) applies, and the probability of drawing a black one is the sum

$$(\text{black}|X) = \sum_{i=1}^{n} p_i = n/N \tag{5-12}$$

i.e., it is just the fraction of black balls in the urn. It is, therefore, also equal to the relative frequency with which we would draw black balls, if we took them all out; or as it is usually stated, if we "sampled the entire population."

We have here one of the many different connections between probability and frequency. In spite of the triviality of its derivation, I ask you to note carefully just how it came about; because today most writers on probability and statistics deny that probability theory has anything to do with plausible reasoning, and insist that the only proper meaning of probability is that of relative frequency in some "random experiment." According to this school of thought, if a probability is not a frequency, then it is not "objective," and its use is just not scientifically respectable.

On the other hand, I maintain that, as its derivation shows, the relation (5-12) has absolutely nothing to do with the definition of probability; on the contrary, it is an almost trivial mathematical consequence of probability theory interpreted as the "calculus of inductive reasoning." In fact, by this broader interpretation of the theory, we lose none of the usual connections between probability and frequency; as will become clear gradually in the remaining lectures, every connection between probability and frequency that is actually used in applications, is deducible in a similar way as a consequence of our "inductive reasoning" form of the theory.

At this point, you might ask, "Aren't you making a tempest in a teapot? Since on either viewpoint we end up writing down the same equation (5-12), which was obvious intuitively without any derivation at all, what difference do these philosophical questions make? It seems like pedantic nit-picking." Well, it is true that in many problems the connection between probability and frequency is so close that the notions are easily confused, and this confusion does no harm in the pragmatic sense that we end up writing down the same equations. Usually, the importance of my nit-picking does not lie at all in the actual equations used; it lies in our judgment about the range of validity of those equations.

The point is that many of the most important problems of current science and engineering are just problems of inductive reasoning, in which no "random experiment" is involved in any way. If you insist that a probability is not respectable unless it is also a frequency, then you will have to conclude that probability theory is just not applicable to these problems. But I am going to insist in these lectures that the relations of probability theory are perfectly valid when used in the Laplace sense of the "calculus of inductive reasoning," whether or not there is any connection between probability and frequency. By using the theory in just this sort of problem, where the "frequentist" would deny the validity of probability theory, I hope to show that we can not only obtain important, useful, and nontrivial results; we can also clear up some of the paradoxes surrounding present communication theory, statistical mechanics, and quantum mechanics.

In fact, the problem of quality control, which led us into this little excursion, provides one of the most striking examples of the value of this nit-picking. However, I want to postpone discussion of the history of this problem until we have the full comparisons at hand; then we will be able to see how much statistical practice has suffered from the other kind of nit-picking, which restricts the apparent range of validity of the theory.

Before returning to the quality-control problem, let's extend the result (5-12) to get the general relations in sampling from a finite population. For this, we need a little more notation; let B_k stand for the proposition, "black ball at the k'th draw," whereupon $b_k \equiv W_k$ will stand for, "white ball at the k'th draw." And, let's indicate the prior information more explicitly. What I called X in (5-12) contained the statement that we have a total of N balls, of which n are black, and (N-n) white; to remind us of this, I will now write Eq. (5-12) in the form

$$(B_1|N,n) = n/N. \tag{5-13}$$

Now, what is the probability of drawing two black balls in two draws? This is, by Rule 1,

$$(B_1B_2|N,n) = (B_1|N,n)(B_2|B_1,N,n) \tag{5-14}$$

First, we suppose that a ball drawn is not replaced before drawing the next one. So, in evaluating the last factor, the fact that one black one has already been drawn means that at the second draw we are sampling from a population of (N-1) balls, of which (n-1) are black; and so

$$(B_1B_2|N,n) = \frac{n(n-1)}{N(N-1)} = \frac{n!\ (N-2)!}{(n-2)!\ N!} \tag{5-15}$$

Continuing in this way, we see that the probability of drawing r black balls in succession without replacement, is

$$(B_1\ldots B_r|N,n) = \frac{n!\ (N-r)!}{(n-r)!\ N!}\quad ,\ r \le n \tag{5-16}$$

The restriction $r \le n$ isn't necessary if we understand that we define factorials by the gamma function relation: $n! \equiv \Gamma(n+1)$; for then the factorial of a negative integer is infinite, and (5-16) automatically gives zero when $r > n$.

Likewise, the probability of drawing s white balls in succession without replacement is given by a relation of the same form, except that the roles of n and (N-n) are interchanged:

$$(W_1\ldots W_s|N,n) = \frac{(N-n)!\ (N-s)!}{(N-n-s)!\ N!} \tag{5-17}$$

Next, we ask for the probability that in m draws without replacement we shall obtain r black balls and (m-r) white ones, in a specified order. Suppose first that black balls are drawn on the first r trials, and white ones on the remaining (m-r) trials. Then Rule 1 gives

$$(B_1\ldots B_rW_{r+1}\ldots W_m|N,n) = (B_1\ldots B_r|N,n)(W_{r+1}\ldots W_m|B_1\ldots B_r,N,n) \tag{5-18}$$

of which the first factor is given by (5-16), and the second by (5-17), if we note that after r black balls have been drawn, we are then sampling

from a population of (N-r) balls (instead of N), of which (n-r) are black (instead of n). Also, the quantity denoted by s in (5-17) is equal to (m-r). So, we have

$$(B_1 \cdots B_r W_{r+1} \cdots W_m | N,n) = \frac{n!\ (N-r)!}{(n-r)!\ N!} \frac{(N-n)!\ (N-m)!}{(N-n-m+r)!\ (N-r)!} \tag{5-19}$$

Although this result was derived for a particular order of drawing black and white balls, the probability actually depends only on the numbers r, (m-r) drawn; and not on the particular order in which black and white appeared. To see this, write out the expression (5-19) more fully, in the manner

$$\frac{n!}{(n-r)!} = n(n-1)(n-2)\cdots(n-r+1) \tag{5-20}$$

and similarly for the two other ratios of factorials in (5-19). It then becomes

$$\frac{n(n-1)\cdots(n-r+1)\,(N-n)(N-n-1)\cdots(N-n-m+r+1)}{N(N-1)\cdots(N-m+1)} \tag{5-21}$$

Now suppose that r black balls and (m-r) white ones are drawn, in any other order. The probability of this is the product of m factors; every time a black one is drawn there appears a factor: (number of black balls in urn)/(total number of balls); and similarly for drawing a white one. The total number of balls in the urn decreases by one at each drawing; therefore, for the k'th drawing a factor (N-k+1) appears in the denominator, whatever the colors of the first k draws. Just before the k'th black ball is drawn, whether this occurs on the k'th trial or any later one, there are (n-k+1) black balls in the urn; so drawing the k'th black one places a factor (n-k+1) in the numerator. Just before the k'th white ball is drawn, there are (N-n-k+1) white balls in the urn; and so drawing the k'th white one places a factor (N-n-k+1) in the numerator regardless of whether this occurs on the k'th trial or any later one. Therefore, by the time all m balls have been drawn, one has accumulated exactly the same factors in numerator and

and denominator as in (5-21); different orders of black and white correspond only to different permutations of the order of factors in the numerator. The probability of drawing r black balls in <u>any</u> specified order in m trials, without replacement, is therefore given by (5-19).

Finally, we ask: what is the probability of drawing exactly r black balls in m trials without replacement, regardless of their order? Different orders of drawing are mutually exclusive events, so we must sum over all possible orders. But since all orders have the same probability (5-19), this means that we must multiply (5-19) by the binomial coefficient

$$\binom{m}{r} \equiv \frac{m!}{r!\ (m-r)!} \tag{5-22}$$

which represents the number of different possible orders of drawing r black balls in m trials. [Question for you to ponder: why isn't this factor just m!? After all, we started this discussion by saying that all the balls, in addition to being either black or white, also carried individual labels i = 1, 2, ..., N, so permutations of black balls among themselves are distinguishable events. A little private thought will enable you to answer this, unless you have had the misfortune of studying Bose and Fermi statistics in quantum theory from the usual textbook discussions; in that case you may have some unlearning to do first. <u>Hint</u>: In (5-19) we are not specifying <u>which</u> black balls and <u>which</u> white ones are to be drawn; if we did, (5-19) would collapse to (N-m)!/N!].

Taking the product of (5-22) and (5-19), the many factorials appearing can be reorganized into three binomial coefficients, and the probability of r black balls in m trials without replacement becomes

$$(r|m,N,n) = \frac{\binom{n}{r}\binom{N-n}{m-r}}{\binom{N}{m}} \tag{5-23}$$

This is our main result, and it is called the hypergeometric distribution, because the right-hand side of (5-23) is closely related to the coefficients in the power series representation of the hypergeometric function. As an aid to memory, we can put this into a more symmetrical form by adopting a new notation; the probability of drawing b black and w white balls, without replacement, from a population of B black and W white ones, is

$$(bw|BW) = \frac{\binom{B}{b}\binom{W}{w}}{\binom{B+W}{b+w}} \tag{5-24}$$

and in this form we can generalize still further. We have been considering an urn with only two kinds of balls: black and white. Suppose there are also red, green, brown, etc. balls present; in all, m different colors. I leave it for you to verify that the probability of drawing n_1 balls of type 1, n_2 of type 2, etc., without replacement, from a population of N_1 of type 1, N_2 of type 2, etc., is

$$(n_1 \ldots n_m | N_1 \ldots N_m) = \frac{\binom{N_1}{n_1} \cdots \binom{N_m}{n_m}}{\binom{\Sigma N_i}{\Sigma n_i}} . \tag{5-25}$$

The hypergeometric distribution (5-23) is rather complicated in its most general form, but it goes into a simpler distribution in the limit where the numbers n, (N-n) become very large compared to the number m sampled. Intuitively, this is clear; since then the proportions of black and white balls in the urn change only negligibly due to the small number drawn, so the probability of getting a black ball is essentially the same at each drawing. To see this mathematically, note that (5-21) can be written as

$$\frac{n^r (N-n)^{m-r}}{N^m} \left\{ \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{r-1}{n}\right)\left(1-\frac{1}{N-n}\right)\left(1-\frac{2}{N-n}\right)\cdots\left(1-\frac{m-r-1}{N-n}\right)}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{m-1}{N}\right)} \right\} \tag{5-26}$$

Now let $N \to \infty$, $(N-n) \to \infty$ in such a way that the ratio $p \equiv n/N$ remains constant.

All the factors in curly brackets in (5-26) tend to unity, and so (5-26) reduces in the limit to

$$\frac{n^r (N-n)^{m-r}}{N^m} = p^r \, (1-p)^{m-r} \tag{5-27}$$

This is the probability of drawing r black, (m-r) white balls in a specified order, and you see that it corresponds to a constant probability p of getting a black ball, (1-p) of getting a white one, at each trial. The probability of getting r black in m draws regardless of the order, again requires the combinatorial factor (5-22); and so in the limit the hypergeometric distribution goes into

$$(r|m,p) = \lim_{\substack{N\to\infty \\ N-n\to\infty \\ n/N\to p}} (r|m,N,n) = \binom{m}{r} p^r \, (1-p)^{m-r} \tag{5-28}$$

This is the binomial distribution, so called because the function

$$f(s) \equiv \sum_{r=0}^{m} s^r \, (r|m,p) = \sum_{r=0}^{\infty} \binom{m}{r} (sp)^r \, (1-p)^{m-r}$$

$$= (sp + 1 - p)^m \tag{5-29}$$

is just a representation of Newton's binomial theorem. F(s) is called the generating function of the binomial distribution; we will see later that generating functions provide a powerful tool for carrying out certain advanced calculations, as was first shown in Laplace's "Theorie Analytique." Note that the evident relation f(1) = 1 is just a verification that the probabilities in (5-28) are correctly normalized; i.e.

$$\sum_{r=0}^{m} (r|m,p) = 1 \tag{5-30}$$

We can carry out a similar limiting process on the generalized hypergeometric distribution (5-25). Again, I leave it for you to verify that in the limit where all the $N_i \to \infty$ in such a way that the fractions

$$p_i \equiv \frac{N_i}{\Sigma N_i}$$

tend to constants, (5-25) goes into the multinomial distribution

$$(n_1 \ldots n_m | p_1 \ldots p_m) = \lim_{N_i \to \infty} (n_1 \ldots n_m | N_1 \ldots N_m)$$

$$= \frac{n!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m} \qquad (5\text{-}32)$$

where $n \equiv \Sigma n_i$. And, as in (5-29), you can define a generating function of (m-1) variables, from which you can prove that (5-32) is correctly normalized.

Up to now, we have considered only the case where we sample without replacement; and that is obviously appropriate to our quality-control problem, where each diode drawn is tested to destruction. But suppose now that we sample balls, and after noting the color 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's go back to the probability of drawing two black balls in succession:

$$(B_1 B_2 | N,n) = (B_1 | N,n)(B_2 | B_1,N,n) \qquad (5\text{-}33)$$

Evidently, we still have (n/N) for the first factor; but what is the second one? Answering this would be, in general, an enormously difficult problem, requiring a vast amount of additional data before it could be solved. Because, what happens to that black 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 of drawing any particular (i'th) ball is no longer given by Rule 4, which led to (5-12).

Evidently, the probability of 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. Even if all these data were at hand, I don't think that a team of the 1,000 best mathematicians in the world, backed up by all the computing facilities in the world, would be able to solve the problem; or would even know how to get started on it. Still, I don't think it would be quite right to say that the problem is unsolvable in principle; only so complicated that it just isn't worth anybody's time even to think about it.

So, what do we do? Well, there's a very clever trick for handling problems that become too difficult. As far as I know, it originated in probability theory; but it produces such euphoria that it has already spread to physics, and there is some danger that it may spread also to other fields.

In probability theory, when a problem becomes too hard to solve, we 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) assert that the shaking has somehow made all these details irrelevant,

so that the problem reverts back to the simple one where Rule 4 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.

I 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 a Priest to convert ordinary water into Holy Water; i.e., it sanctifies their subsequent calculations and renders them immune to criticism. We agnostics often envy the sense of security that the True Believer thus acquires so easily; but 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., that the complicated details, while undeniably relevant, might nevertheless have little numerical effect on the answers to certain particularly simple questions, such as the probability of drawing r black balls in m trials when m is sufficiently small.

From the standpoint of principle, however, an element of vagueness necessarily enters at this point; for while we may feel intuitively that this leads to a good approximation, nobody has ever produced a proof of this, much less a reliable estimate of the accuracy of the approximation, which presumably improves with more shaking. The vagueness is particularly evident in the fact that different people have widely divergent views about exactly how much shaking is required to justify step (2). [Witness the minor furor surrounding a recent Government-sponsored and nationally televised game of chance, 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!]

Again, you may accuse me of nit-picking, because you know that after all these polemics, I am just going to go ahead and use the randomized solution like everybody else does. Note, however, that my objection is not to the procedure itself, provided that we frankly acknowledge 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 of unknown accuracy. 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 in this respect.

My objection is to this mystical belief that by "randomization" we have somehow washed away all our sins, and from that point on we proceed with exact relations--so exact that we can then subject our solution to all kinds of extreme conditions and believe the results. 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 $m \to \infty$ and obtaining the usual "laws of large numbers"). If we don't recognize the approximate nature of our starting equations, we delude ourselves into believing that we have "proved" things (such as the rigorous identity of probability and limiting frequency) that are just not true in real random experiments.

Returning to the equations, what answer can we now give to the question

posed after Eq. (5-33)? The probability $(B_2|B_1,N,n)$ of drawing a black ball on the second draw, clearly depends not only on N and n, but also on the fact that a black one has already been drawn and replaced. But this latter dependence is just 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 B_1 to be irrelevant: $(B_2|B_1,N,n) \simeq (B_2|N,n) = n/N$. After drawing and replacing the second ball, we again shake the urn, declare it "randomized", and set $(B_3|B_2,B_1,N,n) \simeq (B_3|N,n) = n/N$, etc. In this approximation, the probability of drawing a black one at <u>any</u> trial, is (n/N), and (N-n)/N is the probability, at every trial, of drawing a white ball. This leads us to write the probability of drawing exactly r black balls in m trials regardless of order, as

$$(r|m,N,n) = \binom{m}{r}\left(\frac{n}{N}\right)^r\left(\frac{N-n}{N}\right)^{m-r} \tag{5-34}$$

which is just the binomial distribution (5-28) with p = n/N.

Evidently, for small m, this approximation will be quite good; but for large m these small errors can accumulate (depending on exactly how we shake the urn, etc.) to the point where (5-34) is utterly useless. However, I think that some workers in probability theory would deny this; so let's demonstrate it explicitly by a simple, but realistic, extension of the problem.

Suppose that drawing and replacing a black ball actually increases the probability of a black one at the next draw by some small amount $\varepsilon>0$, while drawing and replacing a white one decreases the probability of a black 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. Then

$$\begin{aligned} (B_k|B_{k-1},N,n) &= p + \varepsilon \quad , & (B_k|W_{k-1},N,n) &= p - \delta \\ (W_k|B_{k-1},N,n) &= 1 - p - \varepsilon, & (W_k|W_{k-1},N,n) &= 1 - p + \delta \end{aligned} \tag{5-35}$$

where $p \equiv n/N$. The probability of drawing r black, (m-r) white balls in any specified order, is easily seen to be:

$$p(p+\varepsilon)^{b}(p-\delta)^{b'}(1-p+\delta)^{w}(1-p-\varepsilon)^{w'} \tag{5-36}$$

if the first draw is black, while if the first is white, the first factor in (5-36) should be (1-p). Here b is the number of black draws preceded by black ones, b' the number of black preceded by white, w the number of white draws preceded by white, and w' the number of white preceded by black. Evidently,

$$b + b' = \begin{Bmatrix} r-1 \\ r \end{Bmatrix} , \qquad w + w' = \begin{Bmatrix} m-r \\ m-r-1 \end{Bmatrix} \tag{5-37}$$

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

Now it is clear that, when r and (m-r) are small, the presence of ε and δ in (5-36) makes little difference, and it reduces for all practical purposes to

$$p^{r}(1-p)^{m-r} \tag{5-38}$$

as in the binomial distribution (5-34). But as these numbers increase, we can use relations of the form

$$\left(1 + \frac{\varepsilon}{p}\right)^{b} \simeq \exp\left(\frac{\varepsilon b}{p}\right) \tag{5-39}$$

and (5-36) goes into

$$p^{r}\ (1 - p)^{m-r}\ \exp\left\{\frac{\varepsilon b - \delta b'}{p} + \frac{\delta w - \varepsilon w'}{1-p}\right\} \tag{5-40}$$

The probability of drawing r black, (m-r) white balls now depends on the order in which black and white appear, and for a given ε, when the numbers b, b', w, w' become sufficiently large, the probability can become arbitrarily large (or small) compared to (5-38).

We see this effect most clearly if we suppose that $N = 2n$, $p = 1/2$, in which case we will surely have $\varepsilon = \delta$. The exponential factor in (5-40) then

reduces to:

$$\exp\{2\varepsilon[(b-b') + (w-w')]\} \tag{5-41}$$

This shows that, (1) as the number m of draws tends to infinity, the probability of results containing "long runs"; i.e. long strings of black (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 (εb), etc., become of order unity. Thus, if $\varepsilon = 10^{-3}$, the "randomized" approximation can be trusted up to about $m \sim 1000$; beyond that, you are deluding yourself by using it. In the limit $m \to \infty$, it cannot be trusted for <u>any</u> $\varepsilon > 0$.

All right, we've had a first glimpse at some of the principles and pitfalls of standard sampling theory, so let's turn back to the quality-control problem in which the question came up.

5.3. The Robot's Procedure

You recall, we were trying to use Bayes' theorem in the form of the evidence function:

$$e(A|E) = e(A|X) + 10 \log_{10} \frac{(E|A)}{(E|a)} \tag{5-42}$$

to test hypothesis A ≡ "we have a batch in which 1/3 are bad" against a single alternative B ≡ "we have a batch in which 1/6 are bad." The prior evidence for A was, by (5-10), $e(A|X) = -10$ db, and we had reached the "problem" of evaluating the other terms $(E|A)$, $(E|a)$ in (5-9) for the case that the experimental result was E ≡ "we draw a bad one on the first draw." What is the probability of this happening if A is true? Well, if 1/3 of them are bad, then we are sampling from a population of unknown total N, in which $n = N/3$ are bad, $(N-n) = 2N/3$ good. By (5-12), the probability of drawing a bad one on the first draw, given A, is of course $(\text{bad}|A) = n/N = 1/3$, as was obvious to all from the start. To evaluate $(E|a) = (\text{bad}|a)$, note

that in this problem it is part of the prior information X that either proposition A or B must be true; no other hypothesis about the batch is to be considered (we will see in Lecture 6 what happens if we change this condition). So, in this problem, a = B; if A is false, then B must be true; i.e. there are 1/6 bad, and $(E|a) = 1/6$. Thus, if we draw a bad one on the first draw, this will increase the evidence for A by

$$10 \log_{10} \frac{(E|A)}{(E|a)} = 10 \log_{10} \frac{(1/3)}{(1/6)} = 10 \log_{10} 2 = 3 \text{ db} \qquad (5\text{-}43)$$

What happens now if we draw a second bad one? We are sampling without replacement, so in the notation of (5-14), this contributes further evidence of

$$10 \log_{10} \frac{(B_2|B_1A)}{(B_2|B_1a)} \qquad (5\text{-}44)$$

But $(B_2|B_1A) = (n-1)/(N-1)$ now depends on the number N in a batch. To avoid this complication, let's suppose that N, while unknown, is at least known to be very much larger than any number that we contemplate testing; i.e. we are going to test such a negligible fraction that the proportion of bad and good ones in the batch is not changed appreciably by the drawing. Then the limiting form of the hypergeometric distribution (5-23) will apply, namely the binomial distribution (5-28). Or, you can say equally well that in this case sampling without replacement is practically the same thing as sampling with replacement, leading again to the binomial distribution (5-34). In any event, the result is that the probability of drawing a bad one is the same at every draw, regardless of what has been drawn previously; so Eq. (5-43) now applies for every draw in which we get a bad one. Every bad one we draw will provide +3 db of evidence in favor of hypothesis A, the proposition that we had a bad batch. Now suppose we find a good diode. We'll get evidence for A of

$$10 \log_{10} \frac{(\text{good}|A)}{(\text{good}|a)} = 10 \log_{10} \frac{2/3}{5/6} = -0.97 \text{ db}, \tag{5-45}$$

but let's call it -1 db. Again, this will hold for any draw, if the number in the batch is sufficiently large. If we have inspected n diodes, of which we found n_b bad ones and n_g good ones, the evidence that we have the bad batch will be

$$e(A|E) = -10 + 3n_b - n_g \ . \tag{5-46}$$

You see how easy this is to do once we've set up the machinery. For example, if the first twelve we test show up five bad ones, then we'd end up with

$$e(A|E) = -10 + 15 - 7 = -2 \text{ db} \tag{5-47}$$

or, from Figure (5-1), the probability of a bad batch is brought up to

$$(A|E) \simeq 0.4 \ . \tag{5-48}$$

In order to get at least 20 db worth of evidence for proposition A, how many bad ones would we have to find in a certain sequence of tests? Well, that's not a hard question to answer. If the number of bad ones satisfies

$$n_b \geq 5 + \frac{n}{4} \tag{5-49}$$

then we have at least 20 db of evidence for the bad batch above where we started. Which shows that if we make enough tests, if just slightly more than a quarter of the ones tested turn out to be bad, that will give us 20 db of evidence that we have the batch in which 1 in 3 are bad.

Now all we have here is the probability or plausibility or evidence, whatever you wish to call it, of the proposition that we got the bad batch. Eventually, we have to make a decision. We're going to accept it or we're going to reject it. How are we going to do that? Well, evidently we have to decide beforehand: if the probability of proposition A reaches a certain level than we'll decide that A is true. If it gets down to a certain value, then we'll decide that A is false. There's nothing in probability theory

which can tell us where to put these threshold levels at which we make our decision. This has to be based on our judgment as to what are the consequences of making wrong decisions, and what are the costs of making further tests. For example, making one kind of error (accepting a bad batch) might be very much more serious than making the other kind of error (rejecting a good batch). That would have an obvious effect on where we place our threshold. So we have to give the robot some instructions such as "if the evidence for A gets greater than +0 db, then we'll reject this batch. If it goes down as low as - 15, then we'll accept it."

Let's say that we'd set some threshold limits: we arbitrarily decided that we will reject the batch if the evidence reaches the upper level, and we will accept it if the plausibility goes down to the lower one. We start doing the tests, and every time we find a bad one the evidence for the bad batch goes up 3 db; every time we find a good one, it goes down 1 db. The tests terminate as soon as we get into either the accept or reject region for the first time. This would be the way our robot would do it if we told him to reject or accept on the basis that the posterior probability of proposition A reaches a certain level.

We could describe this in terms of a "control chart," where we start at -10 db at zero number of tests, and plot the result of each test (Fig. 5.2).

5.4. Wald's Probability-Ratio Test.

Now, how does Wald do this? He (Wald, 1947) does not mention Bayes' theorem. But what he actually does is just the same with the one characteristic difference which we find in all these comparisons. Like Fisher in the case of maximum likelihood, he always starts out by throwing away his prior information. His graphs always start out at 0 db.

Wald's probability ratio test involves the calculation of just the last term of Equation (5-9), except that he uses natural logarithms. The

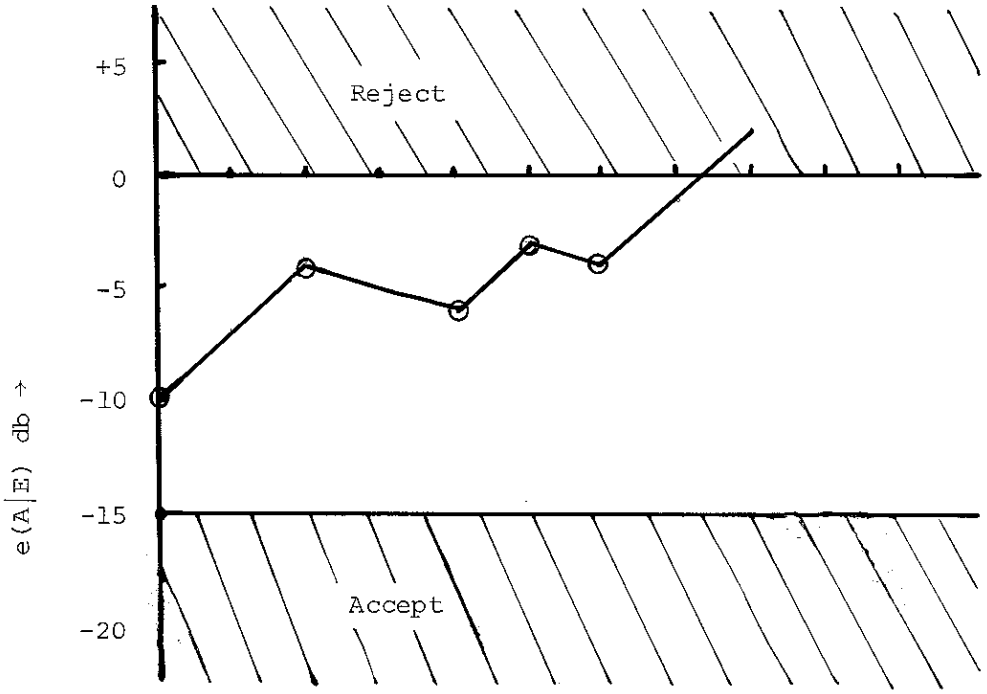

Figure 5.2. A control chart for sequential testing.

name "sequential" describes the fact that the number of tests is not determined in advance, but depends on what is observed. Thus, at each step of the sequence of tests we choose one of three alternatives: (1) accept; (2) reject; (3) make another test. This is the procedure which he conjectured represents an optimum procedure in the sense of requiring on the average fewer tests than any other, but he didn't succeed in proving it. Several years later, such a proof was offered, by Wald and Wolfowitz. We can well imagine how much mathematical effort has been expended on this problem. But how does it look to our robot? Well, to the robot this problem doesn't exist at all; it is only a "Scheinproblem." To him the fact that we have derived it from Bayes' theorem is already the proof that the probability ratio test is the optimum calculation to do, by any sensible criterion of "optimum." Any criterion which required us to reason in a manner not reducible to Bayes' theorem would also require us to be inconsistent in the

sense discussed earlier, or to violate qualitative common sense. Our robot would say this: "When you have calculated the probability of proposition A, conditional on all the available evidence, then you have got everything bearing on the truth of A that is to be had from the evidence. No method of analyzing the data can give you more than this, and there is nothing more to be said."

Does anyone incur any serious error by starting out at zero db? In principle, this is bad in the sense that it is inconsistent if we do have prior information. But, of course, in practice the person using the test still has his common sense; and if he has prior information he will use that information in deciding where to put the boundaries of the accept and reject regions. We cannot remove all the arbitrariness in location of these boundaries, but we can remove some of it, by taking into account prior probability. In practice, the orthodox statistician would use his common sense to take account of his prior information, without ever having to admit that there is any such thing as a "prior probability."

A particularly frank admission of the relevance of prior information is given by Lehman (1959; p. 62) in his well-known work on hypothesis testing according to the "orthodox" viewpoint. He writes: "Another consideration that frequently enters into the specification of a significance level [this is something essentially equivalent to choosing the threshold values in our problem] is the attitude toward the hypothesis before the experiment is performed. If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low." Exactly so! But it is just the prior probability that shows quantitatively how this is to be done.

Of course, there is a great deal more to sequential testing theory than just applying the probability ratio test. There are many questions about the procedure that the manufacturer and customer would ask, and would want the statistician to answer. For example, if all batches have a certain fraction f defective, and we use a sequential test with specified threshold levels, α, β what is the expected number of diodes tested per batch? How does this average sample number depend on $\{f,\alpha,\beta\}$? Or if a fraction g of the batches is in fact bad, what fraction do we expect to be rejected on the average if certain threshold levels are used?

Questions of this type can be answered by straightforward extensions of this analysis and there is an extensive literature on them. In these talks we are concerned only with showing that the rules for plausible reasoning which we have built into the robot's brain will, if applied to this problem, lead to the same actual procedures as the newest methods developed by statisticians. Their conceptual basis is, however, entirely different. To the orthodox statistician, the justification of the sequential probability-ratio test would probably lie in considerations of average sample numbers for given probability of errors. To the robot, this is only an incidental consequence of the fact that this procedure is the one that makes full use of the available data, <u>because</u> it is derivable from Bayes' theorem.

We see that the robot's prediction has been borne out in one more example. We are warned not to use Bayes' theorem for quality-control tests, because it was associated with some metaphysical nonsense 150 years ago. But so was everything else in science. It is even insinuated that Bayes' theorem cannot be "applied." But the simple fact is that the most powerful known method of quality control, only recently discovered by statisticians, is nothing but an application of Bayes' theorem, in exactly the way Laplace would have handled this problem.

5.5. The Value of Nit-Picking.

So now we are in a position to discuss the value of my "nit-picking" about the meaning of Eq. (5-12), and see why the problem of quality control provides a good example of the situation. Basically, of course, it is a problem of testing one hypothesis (we got a bad batch) against a single alternative (we got a good batch); and the mathematics we have developed applies just as well to any such problem of hypothesis testing, such as testing two rival theories in physics, or biology, or economics, against each other.

Now the procedures we are developing in this and the next three lectures, were used by Laplace in just such problems (although not in the logarithmic form, which is only a convenient mathematical detail) from about 1774, and they have been available to anyone who had the sense to use them since the appearance of Laplace's Theorie Analytique in 1812. Yet generations of statisticians were taught that these methods were wrong, and it was only in the early 1940's--130 years later--that statisticians rediscovered the procedure in this lecture from an entirely different view point without at first recognizing it. It was then hailed as a major new advance in statistical practice, and several more years elapsed before it was generally realized (Good, 1950; Wald, 1950) that it was mathematically identical with application of Bayes' theorem in exactly the manner that had long been rejected as wrong.

What caused this procedure to be lost to science for 130 years? Just the point about which I was nit-picking earlier in this lecture; stubborn adherence to a belief, for which there is no supporting evidence, that the notion of probability can be used only in the sense of "frequency in a random experiment". From this one concluded that it is meaningless to speak of the probability that an hypothesis is true, because that is not a "random variable." On such grounds statistical workers denied themselves use of

the proper statistical methods, and worked instead with a great variety of ad hoc approximate methods.

In his later book, "Statistical Decision Functions" (Wald, 1950), Wald developed this theory very much further, and here we have one of those ironical situations where years of the most careful and painstaking work leads right back to the very thing one had been trying to refute. Wald sought to develop a general theory of decision making in the face of uncertainty in a way which avoids the supposed mistakes of Laplace and Bayes, who with Daniel Bernoulli had already developed such a theory in the 18'th century. In order to keep the theory completely "objective," the notion of inductive reasoning, which to Laplace was the central problem of the theory, was suppressed, and attention was concentrated on the decision itself. After long mathematical arguments to impose various conditions of consistency, it finally developed that a class of "admissible" decision rules, which consists, roughly speaking, of all those any sane person would ever consider adopting, is identical with the class derivable by the methods of Bayes and Laplace, and the only basis for a choice among them lay in the prior probabilities! Wald called this class of rules, very properly, "Bayes strategies." As a final irony it was shown (Chernoff & Moses, 1959; Chap. 6), that in practical applications it is only the fact that these decision rules can be found by repeated application of Bayes' theorem that makes it feasible to use this theory at all in nontrivial problems, where the number of conceivable strategies is astronomical. We will come back to these topics when we take up Decision Theory in Lectures 13, 14.

Lecture 6

MULTIPLE HYPOTHESIS TESTING

Let's suppose something very remarkable happens in the sequential test just discussed. Suppose we tested fifty diodes and every one turned out to be bad. According to our equations, that would give us 150 db of evidence for the proposition that we had the bad batch. $e(A|E)$ would end up at +140 db, which is a probability which differs from 1 by one part in 10^{14}. Now our common sense rejects this conclusion. If you test 50 of them and you find that all 50 are bad, you are not willing to believe that you have a batch in which only 1 in 3 are really bad. What is it that went wrong here? Why doesn't our robot work in this case?

Our robot is still immature. He is reasoning like a 4-year-old child does. We've probably all had experience in talking to 4-year-old children. They have enough vocabulary so that you can carry out quite extended conversations with them; they understand the meanings of words. But the really remarkable thing about them is that you can say the most ridiculous things and they'll accept it all with wide open eyes, open mouth, and it never occurs to them to question you. They will believe anything you tell them. The information which our robot should have put into his brain case was not that we had either 1/3 bad or 1/6 bad. The information he should have put in was that Mr. Jaynes <u>said</u> we had either 1/3 bad or 1/6 bad. Those are entirely different propositions.

6.1. Admitting an Unlikely Hypothesis.

The robot should take into account the fact that the information he had may not be perfectly reliable to begin with. There is always a small chance that the whole set of initial data that we've fed into the problem was all wrong. In every problem of plausible reasoning this possibility exists. We could say that generally every situation of actual practice is infinitely complicated. There are always an infinite number of possibilities, and if you start out with dogmatic initial statements which say that there are only two possibilities, then of course you mustn't expect your equations to make sense in every case. So let's see whether we can, in a rather ad hoc way, build this fact into our robot just for this particular example.

Let's provide the robot with one more possible hypothesis, although initially a very unlikely one. Let's say proposition A means as before that we have a box with 1/3 defective, and proposition B stands for the statement that we have a box with 1/6 bad. We add a third proposition, D, which will be the hypothesis that something went entirely wrong with the machine and it's turning out 99 per cent defective. Now, we have to adjust our prior probabilities to take this new possibility into account. I'm going to give hypothesis D a prior probability $(D|X)$ of 10^{-6} (-60 db). I could write out X as a verbal statement which would imply this, but I find that when I try to write a proposition as a verbal statement, there's always someone in the audience who manages to interpret it in a way which I didn't intend. I seem to be unable to write verbal statements which are unambiguous. However, I can tell you what proposition X is, with no ambiguity at all for purposes of this problem, simply by giving the probabilities conditional on X, of all the propositions that we're going to use in this problem. In that way I don't state everything about X, I state everything about X that is relevant to our particular problem. So suppose we start out with these initial probabilities:

$$(A|X) = \frac{1}{11}(1 - 10^{-6})$$

$$(B|X) = \frac{10}{11}(1 - 10^{-6}) \qquad (6\text{-}1)$$

$$(D|X) = 10^{-6}$$

where

A means "we have box which has 1/3 defectives"

B means "we have box which has 1/6 defectives" (this one was formerly called simply a)

D means "machine's putting out 99 per cent defectives."

The factors $(1 - 10^{-6})$ are practically negligible, and for all practical purposes, we will start out with the initial values of evidence:

- 10 db for A

+ 10 db for B

- 60 db for D

Proposition E stands for the statement that "m diodes were tested and every one was defective." Now, according to Bayes' theorem the evidence for proposition D, given E, is equal to the prior evidence plus 10 times the logarithm of this probability ratio:

$$e(D|E) = e(D|X) + 10 \log_{10} \frac{(E|DX)}{(E|dX)} \qquad (6\text{-}2)$$

(In this problem, we're saying that these are the only three hypotheses that are to be considered and, therefore, as far as this problem is concerned, the denial of D is equivalent to the statement that at least one of the propositions A and B must be true.) What are these numbers now? From our discussion of sampling with and without replacement in Lecture 5,

$$(E|DX) = \left(\frac{99}{100}\right)^m \qquad (6\text{-}3)$$

is the probability that the first m are all bad, given that 99 per cent of the machine's output is bad. This is the limiting form of the hyper-

geometric distribution, under our assumption that the total number in the box is very large compared to the number m tested.

We also need the probability $(E|dX)$, which we can evaluate by two applications of Bayes' theorem:

$$(E|dX) = (E|X)\,\frac{(d|EX)}{(d|X)} \tag{6-4}$$

But in this problem it is dogmatically stated that there are only three possibilities, and so the statement $d \equiv$ "D is false" implies that either A or B must be true:

$$\begin{aligned}(d|EX) &= (A+B|EX)\\ &= (A|EX) + (B|EX)\end{aligned} \tag{6-5}$$

where we used Rule 3, the negative term dropping out because A and B are mutually exclusive. Similarly,

$$(d|X) = (A|X) + (B|X) \tag{6-6}$$

Now if we substitute (6-5) into (6-4), Bayes' theorem will be applicable again in the forms

$$\begin{aligned}(E|X)(A|EX) &= (A|X)(E|AX)\\ (E|X)(B|EX) &= (B|X)(E|BX)\end{aligned} \tag{6-7}$$

and so finally we arrive at

$$(E|dX) = \frac{(E|AX)(A|X) + (E|BX)(B|X)}{(A|X) + (B|X)} \tag{6-8}$$

in which all probabilities are known from the statement of the problem.

Although we have the desired result (6-8), let's take time to note that there is another way of deriving it, which is often easier than direct application of Bayes' theorem. The principle is to resolve the proposition whose probability is desired (in this case E) into a set of mutually exclusive propositions, and calculate the sum of their probabilities. We can carry out this resolution in many different ways by, as Professor Myron Tribus has called it, "introducing into the conversation" any new set of mutually

exclusive propositions {P,Q,R,...}. But the success of the method depends on our cleverness at choosing a particular set for which we can complete the calculation. This means that the propositions introduced have to have a known kind of relevance to the question being asked.

In the present case, in evaluation of $(E|dX)$, it appears that propositions A and B have this kind of relevance. Again, we note that proposition d implies (A+B); and so

$$(E|dX) = (E(A+B)|dX) = (EA + EB|dX)$$
$$= (EA|dX) + (EB|dX) \tag{6-9}$$

These probabilities can be factored by Rule 1:

$$(E|dX) = (E|AdX)(A|dX) + (E|BdX)(B|dX) \tag{6-10}$$

But we can abbreviate $(E|AdX) \equiv (E|AX)$, $(E|BdX) \equiv (E|BX)$ because in the way we set up this problem, the statement that either A or B is true implies that D must be false, and so the "d" was redundant. For this same reason, $(d|AX) = 1$, and so by Bayes' theorem,

$$(A|dX) = (A|X)\frac{(d|AX)}{(d|X)} = \frac{(A|X)}{(d|X)} \tag{6-11}$$

Substituting these results into (6-10) and using (6-6), we again arrive at (6-8).

I wanted to exhibit these two ways of doing the calculation because you recall it was one of the conditions of consistency that we imposed on our robot back in Lecture 3, that if there is more than one way of calculating some probability, every such way must lead to the same result. If these two avenues had <u>not</u> led to the same result (6-8), we would have found an inconsistency in our rules, of exactly the sort we sought to guard against by the functional equation arguments of Lecture 3. Needless to say, no case of such an inconsistency has ever been found.

Returning to (6-8), we have the numerical values

$$(E|dX) = \left(\frac{1}{3}\right)^m \frac{1}{11} + \left(\frac{1}{6}\right)^m \frac{10}{11} \qquad (6\text{-}12)$$

and everything in (6-2) is now at hand. If we put all these things together, we come out with this expression for the evidence for proposition D:

$$e(D|E) = -60 + 10 \log_{10} \frac{\left(\frac{99}{100}\right)^m}{\frac{1}{11}\left(\frac{1}{3}\right)^m + \frac{10}{11}\left(\frac{1}{6}\right)^m} \qquad (6\text{-}13)$$

There are some good approximations we can make to this. If m is larger than 5, it's extremely accurate to replace the above by:

$$e(D|E) \simeq -49.6 + 4.73\,m \quad \text{for } m > 5. \qquad (6\text{-}14)$$

And if m is less than 3, there's another approximation which is pretty good:

$$e(D|E) \simeq -59.6 + 7.73\,m \quad \text{for } m < 3. \qquad (6\text{-}15)$$

Let's get some picture of what this looks like. We start out at minus 60 db for the proposition D. The first few bad ones we find will each give us about 7 3/4 db of evidence for the proposition, so the graph of $e(D|E)$ vs. m starts coming up at a slope of 7.7 but then the slope drops, when m gets greater than five, to 4.7. This curve crosses the axis at 10 1/2 and continues on up forever at that same slope. So, ten consecutive bad diodes would be enough to raise this initially very improbable hypothesis up out of the mud, up 58 db, up to the place where the robot is ready to consider it very seriously.

In the meantime, what is happening to our propositions A and B? Well, A starts off at - 10, B starts off at + 10. The plausibility of A starts going up 3 db per defective diode just like it did in the first problem. But after we've gotten too many bad diodes in a row, we'll begin to doubt whether the evidence really supports proposition A after all; proposition D is becoming a much easier way to explain what's observed. So at a certain value of m, the curve for A will stop going up and turn around and go back down.

When I gave these talks at Stanford, I asked the audience to make guesses and test your own plausible reasoning against our robot before you know the answer. Under these conditions, how many consecutive bad diodes would you have to get before you will begin to be very troubled about proposition A, and change your mind about whether the evidence really supports it? Do we have any volunteers? At Stanford I got only one answer, and the answer was eight. The student who gave this is either a mathematical genius or our robot in the flesh, because the turning point according to our equations, to the nearest integer, is just eight. After m diodes have been tested, and all proved to be bad, the evidence for propositions A and B, and the approximate forms, are as follows:

$$e(A|E) = -10 + 10 \log_{10} \frac{\left(\frac{1}{3}\right)^m}{\left(\frac{1}{6}\right)^m + \frac{11}{10} \times 10^{-6}\left(\frac{99}{100}\right)^m}$$

$$\simeq \begin{Bmatrix} -10 + 3\,m & \text{for } m < 7 \\ 49.6 - 4.73\,m & \text{for } m > 8 \end{Bmatrix}, \qquad (6\text{-}16)$$

$$e(B|E) = +10 + 10 \log_{10} \frac{\left(\frac{1}{6}\right)^m}{\left(\frac{1}{3}\right)^m + 11 \times 10^{-6}\left(\frac{99}{100}\right)^m}$$

$$\simeq \begin{Bmatrix} 10 - 3\,m & \text{for } m < 10 \\ 59.6 - 7.33\,m & \text{for } m > 11 \end{Bmatrix}. \qquad (6\text{-}17)$$

These results are summarized in Figure (6.1). We can learn quite a bit about multiple hypothesis testing from studying it. The initial straight line part represents the solution as we found it before we had introduced this proposition D, and both lines A and B would be straight indefinitely on the first solution. When we have introduced D, starting down here at minus 60 db, the plausibility of D will increase, with a change in slope between m = 3 and m = 4, and it continues to increase linearly from then on. The change in plausibility of propositions B and A starts off just

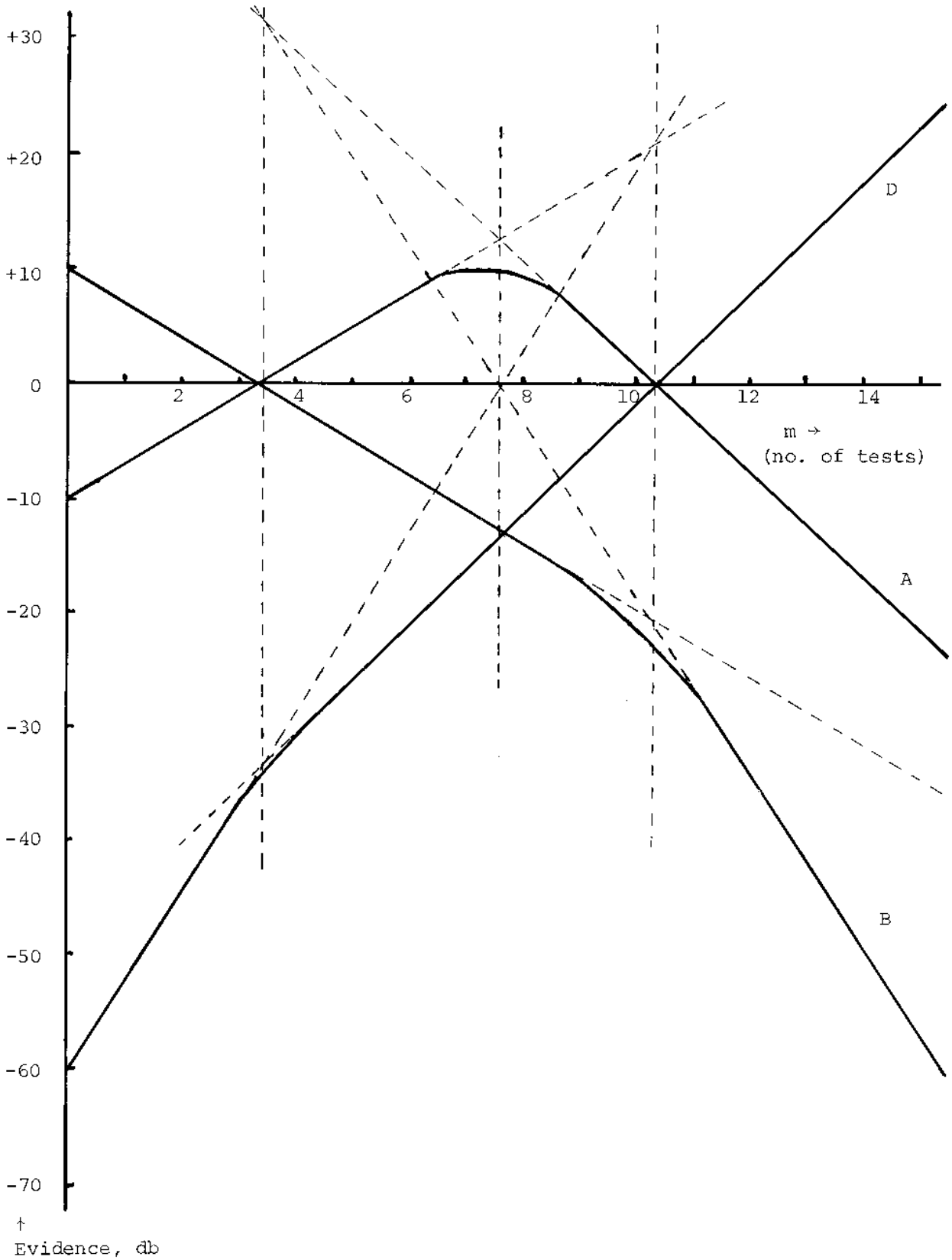

Figure 6.1. Course of a multiple hypothesis test.

the same as in the previous problem; the effect of proposition D does not appear until we have reached the place where D crosses B. At that point, suddenly the character of the A curve changes. The A curve, instead of going on up at this point (at m = 8) has reached its highest value of 10.4 db. Then, it turns around and comes back down. The B curve continues on linearly until it reaches the place where A and D have the same plausibility, and at this point it has a change in slope. From then on, it falls off more rapidly.

Now what is going on here? When D has reached the same plausibility as B, that has a big effect on A. The change in plausibility of A due to one more test arises from the fact that we are testing hypothesis A against two alternative hypotheses: B and D. But initially B is so much more plausible than D, that for all practical purposes, we are simply testing A against B. After enough evidence has accumulated to bring the plausibility of D up to the same level as B, then from that point on, A is essentially being tested against D instead of B, which is a very different situation. All of these changes in slope can be interpreted in this way. Once we see this principle, we see the same thing is going to be true no matter how many hypotheses we have. A change in plausibility of any one hypothesis will always be approximately the result of a test of this hypothesis against a single alternative -- the single alternative being that one of the remaining hypotheses which is most plausible at that time. Whenever the hypotheses are separated by about 10 db or more, then very accurately, multiple hypothesis testing reduces to testing each hypothesis against a single alternative. So, seeing this, you can construct curves of the sort shown in Fig. (6.1) very rapidly without even bothering to look at the equations, because what would happen in the two-hypothesis case is easily seen once and for all.

All the information needed to construct fairly accurate charts resulting from any sequence of good and bad tests is contained in the "plausibility flow diagrams" of Fig. (6.2). They indicate, for example, that finding a good one raises the evidence for B by 1 db if B is being tested against A, and by 19.22 db if it is being tested against D. Similarly, finding a bad one raises the evidence for A by 3 db if A is being tested against B, but lowers it by 4.73 db if it is being tested against D. Likewise, we see that finding a single good one lowers the evidence for D by an amount that cannot be recovered by two bad ones; so D will never attain an appreciable probability unless the observed fraction of bad ones remains persistently greater than 2/3.

Figure (6.1) shows an interesting thing. Suppose we had decided to stop the test and accept hypothesis A if the evidence for it reached plus 10 db. You see, it would reach plus 10 db after about six trials. If we stopped the testing at that point, then of course we would never see the

Good Bad

Figure 6.2. Plausibility flow diagrams.

rest of this curve and see that it really starts going down. If we had continued the testing beyond this point, then we would have changed our mind again. At first glance this seems disconcerting, but notice that it is inherent in all problems of hypothesis testing. If you stop the test at any finite number of trials, then you can never be absolutely sure that you have made the right decision. It is always possible that still more tests would have led you to change your decision.

Evidently, we could extend this example in many different directions. Introducing more "discrete" hypotheses would be perfectly straightforward, as we have seen. More interesting would be the introduction of a continuous range of hypotheses, such as:

$M_f \equiv$ "The machine is putting out a fraction f good." Then instead of a discrete prior probability distribution, our robot would have a continuous distribution in $0 \leq f \leq 1$, and by Bayes' theorem he would calculate the posterior probability distribution of f, on the basis of the observed samples, from which various decisions could be made. In fact, let's just take a glimpse at the equations for that case.

6.2. Testing an Infinite Number of Hypotheses.

We are now testing simultaneously an infinite number of hypotheses about the machine, and as often happens in mathematics, this actually makes things simpler. However, the logarithmic form of Bayes' theorem is now rather awkward, and so we will go back to the original form,

$$(A|BX) = (A|X)\,\frac{(B|AX)}{(B|X)} \tag{6-18}$$

There is a prior probability density

$$(df|X) = p(f)\,df \tag{6-19}$$

which gives the probability that the fraction of good ones is in the range df; and let E stand for the result thus far of our experiment:

E = "N diodes were tested and we found the results GGBGBBG$\cdots$,

containing in all n good ones and (N-n) bad ones."

Then the posterior probability density of f is, by Bayes' theorem,

$$(df|EX) = (df|X)\,\frac{(E|f,X)}{(e|X)} = P(f)\,df \tag{6-20}$$

or,

$$P(f) = p(f)\,\frac{(E|f)}{(E|X)} \tag{6-21}$$

The denominator is just a normalizing constant, which we could calculate directly; but usually it is easier to determine it (if it is needed at all) from requiring that the posterior density satisfy the normalization condition

$$\int_0^1 P(f)\,df = 1 \tag{6-22}$$

The evidence of the experiment thus lies entirely in the f-dependence of the likelihood function $(E|f)$.

Now if we are given that f is the correct fraction of good ones, then the probability of getting a good one at each trial is f, and the probability of getting a bad one is (1-f). The probabilities at different trials are, by hypothesis (i.e., one of the many statements hidden there in X), independent, and so, as in Eq. (5-27),

$$(E|f) = f^n\,(1-f)^{N-n} \tag{6-23}$$

(note that the experimental evidence E told us not only how many good and bad ones were found, but also the order in which they appeared). Therefore, we have the posterior distribution

$$P(f) = \frac{f^n\,(1-f)^{N-n}\,p(f)}{\int_0^1 f^n\,(1-f)^{N-n}\,p(f)\,df} \tag{6-24}$$

You may be startled to realize that all of our previous discussion of quality control is contained in this simple looking equation, as a special case. For example, the multiple hypothesis test starting with (6-1) and including the final results (6-13) - (6-17) is all contained in (6-24)

corresponding to the particular choice of prior density:

$$p(f) = \frac{10}{11}(1 - 10^{-6})\ \delta\left(f - \frac{1}{6}\right) + \frac{1}{11}(1 - 10^{-6})\ \delta\left(f - \frac{1}{3}\right) + 10^{-6}\ \delta(f - 0.99) \qquad (6\text{-}25)$$

where $\delta(f)$ is the Dirac delta-function. The three delta-functions here correspond to the three discrete hypotheses B, A, D respectively, of that example, and they appear in the posterior density with altered coefficients which are just the probabilities given in (6-13), (6-16), (6-17).

Suppose that at the start of this test our robot was fresh from the factory that made him; he had no prior knowledge about the machines at all, except for our assuring him that it is possible for a machine to make a good one, and also possible for it to make a bad one. In this state of knowledge, what prior probability density $p(f)$ should he assign? It seems to me, as it did to Laplace, that in this case the robot has no basis for assigning to any particular interval df a higher probability than to any other interval of the same size; so the only honest way he can describe what he knows is to assign a uniform prior probability density, $p(f) = \text{const}$. To normalize it correctly as in (6-22), we must take

$$p(f) = 1, \qquad 0 \le f \le 1. \qquad (6\text{-}26)$$

It was Bayes himself who first took this step, in his famous work (Bayes, 1762) that started this 200-year-old controversy about probability theory. The problem he considered was, of course, different in statement than ours; but they are mathematically equivalent. Bayes' work was published posthumously, and it appears that he felt a little uneasiness about the validity of (6-26). Laplace took up the subject at this point, and in a series of memoirs from 1772, developed Bayes' work into a general method of statistical inference.

From our viewpoint today, we can say that there is nothing wrong with (6-26); the only valid criticism is that neither Bayes nor Laplace specified clearly the exact state of knowledge in which (6-26) is appropriate. I have tried to give this here, although at this stage the manner in which the result (6-26) follows from my verbal statement cannot be clear. This will be shown later, when we take up transformation groups.

The integral in (6-24) is then the well-known Eulerian integral of the first kind, today more commonly called the complete Beta-function; and (6-24) reduces to

$$P(f) = \frac{(N+1)!}{n!\ (N-n)!}\, f^n\ (1-f)^{N-n} \tag{6-27}$$

This has a single peak in $0 \le f \le 1$, located by differentiation at

$$f = \hat{f} = \frac{n}{N} \tag{6-28}$$

which is the same as the maximum-likelihood estimate of f, and equal to the frequency with which good ones were observed. To find the sharpness of the peak in (6-27), write

$$L(f) \equiv \log P(\hat{f}) = n \log f + (N-n) \log (1-f) + \text{const.} \tag{6-29}$$

and expand L(f) in a Taylor series about $\hat{f}$. The first terms are

$$L(f) = L(\hat{f}) - \frac{N}{\hat{f}\ (1-\hat{f})}\, \frac{(f-\hat{f})^2}{2!} + \cdots \tag{6-30}$$

and so, to this approximation, (6-27) is a <u>gaussian</u>, or <u>normal</u>, distribution

$$P(f) \simeq A \exp\left\{-\frac{(f-\hat{f})^2}{2\sigma^2}\right\} \tag{6-31}$$

where

$$\sigma^2 = \frac{\hat{f}\ (1-\hat{f})}{N} \tag{6-32}$$

and A is a normalizing constant. I leave it for you to convince yourself that (6-31) is actually an excellent approximation to (6-27) in the entire interval $0 < f < 1$, provided that $n \gg 1$ and $(N-n) \gg 1$.

Thus after observing the evidence E = "n good ones in N trials," the robot's state of knowledge about f can be described pretty well by saying that he considers the most likely value of f to be just the observed fraction of good ones, and he considers the accuracy of this estimate to be such that the interval $\hat{f} \pm \sigma$ is reasonably likely to contain the true value. More precisely, from numerical analysis of (6-31), he says that with 50% probability the true value is contained in the interval $\hat{f} \pm 0.68\sigma$; with 90% probability it is contained in $\hat{f} \pm 1.65\sigma$; and with 99% probability it is contained in $\hat{f} \pm 2.57\sigma$. As the number N of tests increases, these intervals shrink, according to (6-32), proportional to $N^{-1/2}$, the usual rule we expect to find in probability theory.

In this way, we see that the robot starts in a state of "complete ignorance" about f; but as he accumulates information from the tests, he acquires more and more definite opinions about f, which correspond very nicely to common sense (except that common sense will hardly give us a definite numerical interval such as $\hat{f} \pm 1.65\sigma$). One caution; all this applies only to the case where, although the numerical value of f is initially unknown, it was known that f is not changing with time.

Still more interesting, and more realistic for actual quality-control situations, would be to introduce the possibility that f might vary with time, and the robot's job is to make the best possible inferences about whether the machine is drifting out of adjustment, with the hope of correcting trouble before it became serious. A simple classification of diodes as bad and good is not too realistic; there is actually a continuous gradation of quality, and by taking that into account we could refine these methods. There might be several important properties in addition to the maximum allowable inverse voltage (for example, forward resistance, noise temperature, rf impedance, low-level rectification efficiency, etc.), and we might also have to control

the quality with respect to all these. There might be a great many different machine characteristics, instead of just M_f, about which we need plausible inference.

You see that we could easily spend years on this problem. But let me just say that although the details can become arbitrarily complicated, there is in principle no difficulty in making whatever generalization you need. It requires no new principles beyond what we have already given.

In the problem of detecting a drift in machine characteristics, you would want to compare our robot's procedure with the ones described by Shewhart (1931). You would find that Shewhart's methods are a pretty good approximation to what our robot would do; in some of the cases involving a normal distribution they are exactly the same. In statisticians' language, the reason for this is that the mean and variance of a sample drawn from a normal distribution are "sufficient statistics" for estimation of the mean and variance of the parent distribution. Translated into our language: in applying Bayes' theorem, the robot always finds that the mean and variance of the sample are the only properties of the sample he needs (i.e., all other details are irrelevant) for making inferences about the machine. These cases are, incidentally, the only ones where Shewhart felt that his procedures were fully satisfactory.

I don't want to go into this further now, because this is really the same problem as that of detecting a signal in noise, which we will study later on. Also, it is equivalent to the problem of deciding from a set of astronomical observations (i.e., positions of the planets) whether there is some unknown systematic effect, or whether discrepancies should be blamed on errors of observation. Laplace was applying this theory from about 1772 in just that way--to calculate the <u>probability</u> that an unknown systematic effect exists, and thus to help him decide which astronomical problems were worth working on.

This use of probability theory led him to some of the most important discoveries in celestial mechanics, and his methodology might well be noted by scientists today.

Of course, I don't mean to set up Laplace as a kind of demigod who could do no wrong. Today, it is easy enough--in fact, it is child's play--to find things to criticize in Laplace's work, if you consider that a worthy occupation. If another 150 years of continuous work in this field had not resulted in any improvement of techniques or clarification of principles, that would certainly make Laplace unique among all scholars who ever lived. But I think that the following judgment of the situation is a fair one: for several generations the dominant school of statisticians has rejected and ridiculed Laplace's whole conception of probability theory, while they slowly and laboriously rediscovered his methods. If past efforts to discredit Laplace had been directed instead toward understanding his contributions and learning how to use them properly, statistical practice would be far more advanced today than it is.

Lecture 7

QUEER USES FOR BAYES' THEOREM

I. J. Good (Good, 1950) has shown how we can use Bayes' theorem backwards to measure our own strengths of belief about propositions. For example, how strongly do you believe in extrasensory perception?

7.1. Extrasensory Perception.

What probability would you assign to the proposition that Mr. Smith has perfect extrasensory perception? He can guess right every time which number you are thinking of. Well now, to say zero--that, of course, is dogmatic. According to our theory, if you start out at $-\infty$ db, this means that you are never going to allow your mind to be changed by any amount of evidence, and you don't really mean that. But where *is* our strength of belief in a proposition like this? Our brains work pretty much the way this robot works, but *we* have an intuitive feeling for plausibility only when it's not too far from 0 db. We feel that something is more than likely to be so or less than likely to be so. We get fairly definite feelings about that. So the trick is to imagine an experiment. How much evidence would it take to bring my state of belief up to the place where I felt very perplexed and unsure about it? Not to the place where I believed it--that would overshoot the mark, and again we'd lose our resolving power. How much evidence would it take to bring you just up to the point where you were beginning to consider the possibility seriously?

We take this man who says he has extrasensory perception, and we will write down some numbers from 1 to 10 on a piece of paper and ask him to guess which numbers we've written down. We'll take the usual precautions to make sure against other ways of finding out. All right, if he guesses the first number correctly, of course we'll say "you're a very lucky person, but I don't believe it." And if he guesses two numbers correctly, we'll say "you're a very lucky person, but I don't believe it." By the time he's guessed four numbers correctly--well, I still wouldn't believe it. So my state of belief is certainly lower than -40 db. How many numbers would he have to guess correctly before you would really seriously consider the hypothesis that he has extrasensory perception? In my own case, I think somewhere around 10. My personal state of belief is, therefore, about -100 db. You could talk me into a $\pm$10 change fairly easily, and perhaps $\pm$20; but not much more than that.

7.2. Bayesian Jurisprudence.

It is interesting also to apply Bayes' theorem to various situations in which we can't really reduce it to numbers very well, but still it shows automatically what kind of information would be relevant to help us do plausible reasoning. Suppose someone in New York City has committed a murder, and you don't know at first who it is. Suppose there are 10 million people in New York City. On the basis of no knowledge but this, $e(\text{Guilty}|X) = -70$ db is the plausibility that any particular person is the guilty one.

How much positive evidence is necessary before we decide some man should be put away? Maybe +40 db, although your first reaction may be that this is not safe enough, and the figure ought to be higher. If we raise this figure, we give increased protection to the innocent, but at the cost of making it more difficult to convict the guilty; and at some point the

interests of society as a whole must take precedence over sentiment.

For example, if a thousand guilty men are set free, we know from only too much experience that two or three hundred of them will immediately proceed to inflict still more crimes upon society, and their escaping justice will encourage a hundred more to take up crime. So, I think it is clear that the damage done to society as a whole by allowing a thousand guilty men to go free, is far greater than that caused by falsely convicting one innocent man. If you have a sentimental reaction against this statement, I ask you to think: if you were a judge, would you rather face one man whom you had convicted falsely; or a hundred victims of crimes resulting from your lenience? Setting the threshold at +40 db will mean, crudely, that on the average not more than one conviction in ten thousand will be in error; a judge following this rule will probably not make one false conviction in a working lifetime on the bench. It seems to me that this is a reasonable figure that we can accept. Obviously, however, this matter ought to be researched much more carefully than we can do here.

So, if we took +40 db starting out from -70, this means that in order to get conviction you would have to produce about 110 db of evidence in favor of the guilt of this particular person.

Suppose now we learn that this person had a motive. What does that do to the plausibility of his guilt? Well, Bayes' theorem says

$$e(\text{Guilty}|\text{Motive}) = e(\text{Guilty}|X) + 10 \log_{10} \frac{(\text{Motive}|\text{Guilty})}{(\text{Motive}|\text{Not Guilty})} \qquad (7\text{-}1)$$

$$\simeq -70 - 10 \log_{10} (\text{Motive}|\text{Not Guilty})$$

since $(\text{Motive}|\text{Guilty}) \simeq 1$; i.e., we consider it quite unlikely that the crime had no motive at all. Thus, the significance of learning that the person had a motive depends almost entirely on the probability $(\text{Motive}|\text{Not Guilty})$ that an innocent person would also have a motive. This evidently

agrees exactly with our common sense; if the deceased were kind and loved by all, hardly anyone would have had a motive to do him in. Learning that, nevertheless, our suspect did have a motive, would then be very significant information. If the victim had been an unsavory character, who took great delight in all sorts of foul deeds, then a great many people would have a motive, and learning that our suspect was one of them, is not so significant. The point of this is that we don't really know what to make of the information that our suspect had a motive, unless we also know something about the character of the deceased. But how many members of juries would realize that, unless it was specifically pointed out to them?

Suppose that a very enlightened judge, with powers not given to judges under present law, had perceived this fact and, when testimony about the motive was introduced, he directed his assistants to obtain for the jury the most reliable data possible on the number of people in New York City who had a motive. This number was N_m. Then

$$(\text{Motive}|\text{Not Guilty}) = \frac{N_m - 1}{(\text{number of people in New York}) - 1} \simeq 10^{-7}(N_m - 1)$$

and equation (7-1) reduces, for all practical purposes, to

$$e(\text{Guilty}|\text{Motive}) \simeq -70 + 10 \log [10^7/(N_m - 1)] = -10 \log (N_m - 1). \quad (7\text{-}2)$$

You see that the population of New York has cancelled out of the equation; as soon as we know the number of people who had a motive, then it doesn't matter any more how large the city was.

Well, you can go on this way for a long time, and I think you will find it both enlightening and entertaining to do so. For example, we now learn that the suspect had bought a gun the day before the crime. Or that he was seen at the scene of the crime shortly before. If you have ever been told not to trust Bayes' theorem, you should follow a few examples like this a good deal further, and see how infallibly it tells you what information

would be relevant, what irrelevant, in plausible reasoning. Even in situations where we would be quite unable to say what numerical values should be used, it still reproduces qualitatively just what your common sense (after perhaps a little meditation) tells you.

7.3. Testing Scientific Theories.

Another class of applications of Bayes' theorem, which has been discussed vigorously by philosophers of science for over a century, concerns the reasoning process of a scientist, by which he accepts or rejects his theories in the light of the observed facts. I mentioned in the second lecture that this consists largely of the use of two forms of syllogism,

$$\text{one strong: } \left\{\begin{array}{l}\text{If A, then B}\\ \underline{\text{\quad B false \quad}}\\ \text{A false}\end{array}\right\}, \text{ and one weak: } \left\{\begin{array}{l}\text{If A, then B}\\ \underline{\text{\quad B true \quad}}\\ \text{A more plausible}\end{array}\right\}$$

We see that these correspond to the use of Bayes' theorem in the forms

$$(A|b) = (A|X)\,\frac{(b|A)}{(b|X)}\,, \qquad (A|B) = (A|X)\,\frac{(B|A)}{(B|X)}$$

respectively. It is at once obvious that Bayes' theorem accounts for the strong syllogism; for if $(B|A) = 1$, Bayes' theorem gives $(A|b) = 0$; our rules for plausible reasoning include those of deductive reasoning as a special case.

Interest here centers on the question whether the second form of Bayes' theorem gives a satisfactory quantitative version of the weak syllogism. Let us consider a specific example given by Professor George Polya [Polya, 1954; Vol. II, pp. 130-132]. The planet Uranus was discovered by Herschel in 1781. Within a few decades (i.e. by the time Uranus had traversed about one third of its orbit), it was clear that it was not following exactly the path prescribed for it by the Newtonian theory (laws of mechanics and gravitation). At this point, a naive application of the strong syllogism might lead one to conclude that the Newtonian theory was demolished. However,

its many other successes had established the Newtonian theory so firmly that to the French astronomer Leverrier, an alternative hypothesis was rendered more plausible: there must be still another planet beyond Uranus, whose gravitational pull is causing the discrepancy.

Working backwards, Leverrier computed the mass and orbit of a planet which could produce the observed deviation and predicted where the new planet would be found. An observatory received Leverrier's prediction on September 23, 1846, and on the evening of the same day, the new planet (Neptune) was discovered within one degree of the predicted position!

Instinctively, we feel that the plausibility of the Newtonian theory was increased by this little drama. The question is, how much? The attempt to apply Bayes' theorem to this problem will give us a good example of the complexity of actual situations faced by scientists, and also of the caution which must be exercised in reading the rather confused literature on these problems.

Following Polya's notation, let T stand for the Newtonian theory, N for the part of Leverrier's prediction that was verified. Then Bayes' theorem gives for the posterior probability of T,

$$(T|N) = (T|X)\,\frac{(N|TX)}{(N|X)}\,. \tag{7-3}$$

Suppose we try to evaluate $(N|X)$. This is the prior probability of N, regardless of whether T is true or not. Since $N = N(T+t) = NT + Nt$, we have, by applying Rule 3, then Rule 1,

$$\begin{aligned}(N|X) &= (NT + Nt|X) = (NT|X) + (Nt|X)\\ &= (N|TX)(T|X) + (N|tX)(t|X)\end{aligned} \tag{7-4}$$

and you see that $(N|tX)$ has intruded itself into the problem. But in the problem as stated this quantity is not defined; the statement t ≡ "Newton's theory is false" has no definite implications until we specify what alterna-

tive we have to put in place of Newton's theory.

For example, if there were only a single possible alternative according to which there could be no planets beyond Uranus, then $(N|tX) = 0$, and Bayes' theorem would again reduce to deductive reasoning, giving $(T|N) = 1$, independently of the prior probability $(T|X)$. On the other hand, if Einstein's theory were the only possible alternative, its predictions do not differ appreciably from those of Newton's theory for this phenomenon, and we would have $(N|tX) = (T|X)$, whereupon Bayes' theorem reduces to $(T|N) = (T|X)$. Verification of Leverrier's prediction might elevate the Newtonian theory to certainty, or it might have no effect at all on its plausibility! It depends entirely on this: Against which specific alternatives are we testing Newton's theory?

Now to a scientist who is judging his theories, this conclusion is the most obvious exercise of common sense. Yet statisticians have developed criteria for accepting or rejecting theories (Chi-squared test, etc.) which make no reference to any alternatives. A practical difficulty of this was pointed out forcefully by Sir Harold Jeffreys (Jeffreys, 1939); there is not the slightest use in rejecting any hypothesis H unless we can do it in favor of some definite alternative H' which better fits the facts.* Bayes' theorem tells us much more than this: unless the observed facts are absolutely impossible on hypothesis H, it is meaningless to ask how much those facts tend "in themselves" to confirm or refute H. Not only the mathematics, but also our common sense (if we think about it for a minute) tells us that we have not asked any definite, well-posed question

*I don't mean to argue against the use of the Chi-squared test itself; later in these lectures, when we take up significance tests, we will see that in some cases it is very nearly the right test to answer a different question, namely: "Within a certain specified class of alternatives H', do any exist which better fit the facts, and how much improvement in fit is possible?"

until we specify the possible alternatives to H.

Of course, as the observed facts approach impossibility on hypothesis H, we are led to worry more and more about H; but mere improbability, however great, cannot in itself be the reason for doubting H. For example, if I toss a coin 1000 times, then no matter what the result is, the specific observed sequence of heads and tails has a probability of only 2^{-1000}, or minus 3000 decibels, on the hypothesis that the coin is honest. If, after having tossed it 1000 times, I still believe that the coin is honest, it can be only because the observed sequence is even more improbable on any alternative hypothesis that I am willing to consider seriously. This situation will be analyzed more deeply later on, where it will lead to a general formulation of significance tests.

We see here that, even when the application is only qualitative, classical probability theory is still useful to us in a normative sense; it is the test by which we can detect inconsistencies in our own reasoning. Some authors have argued strongly against the use of Bayes' theorem for testing hypotheses. But when we take the trouble to learn what it actually says, we find that Bayes' theorem tells immediately what is needed before we have any rational criterion for testing hypotheses.

This brings us to some comparisons with the literature. In Polya's discussion of Bayes' theorem applied to the status of Newton's theory before and after Leverrier's feat, no specific alternative to Newton's theory is stated; but from the numerical values used (loc. cit., p. 131) we can infer that the alternative H' was one according to which it was known that one more planet existed beyond Uranus, but all directions on the celestial sphere were considered equally likely. Unfortunately, in the calculation no distinction was made between $(N|X)$ and $(N|tX)$; and consequently the quantity which Polya interprets as the ratio of posterior to prior probabi-

lities of Newton's theory, is actually the ratio of posterior to prior odds. This is, in our notation, $(N|TX)/(N|tX) = (N|TX)/(N|H'X) \simeq 13{,}000$.

The conclusions are much more satisfactory when we notice this. Whatever prior probability $(T|X)$ we imagine Newton's theory to have, if H' is the only alternative considered, then verification of Leverrier's prediction increased the evidence for Newton's theory by $10 \log_{10}(13{,}000) \simeq 41$ decibels. Actually, if there were a new planet, it would be reasonable to adopt a different alternative hypothesis H", according to which its orbit would lie in the plane of the ecliptic, as Polya points out. If, on hypothesis H", all values of longitude are considered equally likely, we might reduce this figure to about $10 \log_{10}[(N|TX)/(N|H"X)] = 10 \log_{10}(180) \simeq 23$ decibels. In view of the great uncertainty as to just what the alternative is, it seems to me any value between these extremes is more or less reasonable.

There was a difficulty (which Polya interpreted as revealing an inconsistency in Bayes' theorem), that if the probability of Newton's theory were increased by a factor of 13,000, then the prior probability was necessarily lower than (1/13,000); but this contradicts common sense, because Newton's theory was already very well established before Leverrier was born. Recognition that we are, in the above numbers, dealing with odds rather than probabilities, completely removes this objection and makes Bayes' theorem appear quite satisfactory in describing the inductive reasoning of a scientist. This is a good example of the way in which objections to the Bayes-Laplace methods which you find in the literature, disappear when you look at the problem more carefully.

But the example also shows clearly that in practice the situation faced by the scientist is so complicated that there is little hope of applying Bayes' theorem to give quantitative results about the relative status of theories. Also there is no need to do this, because the real difficulty

of the scientist is not in the reasoning process itself; his common sense is quite adequate for that. The real difficulty is in learning how to formulate new alternatives which better fit the facts. Usually, when one succeeds in doing this, the evidence for the new theory soon becomes so overwhelming that no one needs probability theory to tell him what conclusions to draw. So, I would say that in principle the application of Bayes' theorem in the above way is perfectly legitimate; but in practice it is of very little use to a scientist.

7.4. Different Views on Probability Theory.

Professor L. J. Savage (Savage, 1954) has written an excellent survey of the foundations of statistics, in which he clearly recognizes, and gives a rigorous discussion of, many of the points that I am trying to put across here in a more informal way. He gives a broad classification of attitudes toward probability theory into three different camps:

(a) Objectivistic. Probability has nothing whatsoever to do with "degree of reasonable belief" or inductive reasoning. By "probability" we must mean only observable frequencies in independent repetitions of a random experiment.

(b) Personalistic. Probability can be used legitimately to describe the degree of confidence that a particular individual has in the truth of a proposition, but probability assignments are not unique; two individuals having the same prior evidence may assign different probabilities without either being unreasonable.

(c) "Necessary views hold that probability measures the extent to which one set of propositions, out of logical necessity and apart from human opinion, confirms the truth of another. They are generally regarded by their holders as extensions of logic, which

tells when one set of propositions necessitates the truth of another."

Here I have merely summarized Savage's description of objectivistic and personalistic views, but quoted his statement about "necessary" views in full. This is the view which he imputes to Laplace; or more accurately, Laplace's view is described (p. 278) as a "naive necessary one."

I want to say something about each of these adjectives, because I am expounding a viewpoint which I believe is the same as Laplace's (although from this distance in time, there is no way to be sure of that in every detail). Since the term "necessary" was coined by Savage, we have to accept its definition as given above; but we can still ask whether the definition properly describes Laplace's view (or the one I am developing, if there is any difference). Now in order to answer this, it would clearly be absurd to try to consult every statement about probability made by, or in the name of, Laplace. We have to distinguish clearly between probability theory and things that have been said about probability theory; too often, they are entirely different. The only way to find out what Laplace's form of probability theory "really says" about some question is to look at the equations Laplace gave us, in some specific case where the question comes up.

Now, where is an equation which says that probability measures the extent to which one set of propositions, out of logical necessity, confirms the truth of another? Where, indeed, is the relation in logic which tells when one set of propositions necessitates the truth of another? The relations of logic are of the form, "If A implies B, and if B implies C, then ...." There is nothing in logic which tells us whether A does in fact imply B. In other words, the relations of logic are only rules for the consistent manipulation of implications; they do not tell us whether some proposed implications are correct, but only whether they are mutually consistent.

It is exactly the same in probability theory. The basic equations are simply,

$$(AB|C) = (A|BC)(B|C)$$

$$(A|B) + (a|B) = 1$$

These, you see, are again statements of the form, "If C implies B to the extent $(B|C)$, and if BC implies A to the extent $(A|BC)$, then ...." There is nothing which tells whether C does in fact imply B to the extent $(B|C)$. In other words, the relations of probability theory are only rules for the consistent manipulation of partial implications; they do not tell us whether some proposed probability assignments are correct, but only whether they are mutually consistent.

If, on meditation, I decide that my personal probabilities are $(B|C) = 3/4$, $(A|BC) = 4/5$, $(AB|C) = 1/2$, then probability theory tells me that I am reasoning inconsistently. It does not tell me how to resolve that inconsistency.

But we can, in the case of probability theory, make a much stronger statement. What did we just learn? How much did verification of Leverrier's prediction N, out of logical necessity, confirm the truth of Newton's theory T? Bayes' theorem not only did not answer this, but it explicitly stated the opposite of the "necessary" view: Unless N is absolutely impossible on hypothesis T, it is meaningless to ask how much N, in itself, confirms the truth of T.

How about Rule 4? Isn't that an equation that tells us that one proposition does, out of logical necessity, confirm the truth of another to a definite extent? No, it isn't. Mathematically, the rule asserts one thing, and one thing only: if the sum of N equal numbers is unity, then each of the numbers must be N^{-1}. Rule 4 assigns definite numerical values to probabilities only after we have arbitrarily specified the set of propositions $A_1 \ldots A_N$ that

we're going to consider. Nothing in probability theory tells us that this specific set of propositions was the right set to introduce.

Consider two different problems; in problem (1) we have N different propositions, $A_1 \ldots A_N$. In problem (2) we have one more proposition A_{N+1} that must be taken into account. In general, for a given specific piece of evidence E, the probability $(A_1|E)$ will be different in the two problems. We saw this in detail when we studied multiple hypothesis testing in Lecture 6; addition of hypothesis D to the problem completely changed the numerical value of $(A|E)$.

Probability theory not only does not say that evidence E confirms the truth of A to some definite extent; it explicitly denies that any such relation exists. The probability $(A|E)$ does not depend only on A and E; it depends also on which alternatives to A we are considering, and it is mathematically indeterminate until those alternatives have been specified.

So, I think we have to plead "not guilty" to any charge that Laplace's formulation of probability theory is a "necessary" one. Indeed, if anyone is guilty of supposing that one proposition confirms the truth of another to any unique extent, it is the "objectivist" who teaches his students how to accept or reject hypotheses without considering the alternatives. Laplace's theory will not allow us to commit that error of reasoning.

Why have I answered this objection at such great, and repetitious, length? For several decades, authors of works on probability and statistics have been repeating the charge that Laplace's theory is nonsense <u>because</u> it supposes that for any two propositions A, B, there is a definite numerical value of $(A|B)$. The most casual glance at Laplace's <u>equations</u> shows that this is simply not true.

I think the trouble comes ultimately from some unfortunate historical accidents. After Laplace's death, some nineteenth-century philosophers made

ridiculous misapplications of probability theory, asserted that their nonsensical conclusions were "mathematically proved," and invoked the authority of Laplace to back them up. No man's reputation ever suffered more from the antics of enthusiastic but uncritical friends. The rise of the "objectivist" viewpoint in the twentieth century is an understandable, but misdirected, reaction against this lunacy. Instead of analyzing the transgressions and learning how to avoid such mistakes in the future, it was much easier to attack Laplace.

On the other hand, isn't it perfectly obvious that probability theory is an extension of logic, in exactly the sense alluded to by Savage? Probability theory fills in the gap between logical proof and disproof and shows us how to reason consistently in the intermediate region where, of necessity, virtually all of our actual reasoning takes place. It clearly includes deductive logic as a special case. I am continually amazed at the caution with which mathematicians approach this issue, and at their extreme reluctance to take the problem of inductive reasoning seriously. One gets the impression that an extension of logic is some enormously difficult, and probably impossible, problem which ordinary mortals had better leave alone.

Part of our communication gap here lies in the fact that no one has ever given an explicit answer to this question: What is it that we should prove about a proposed extension of logic before mathematicians will take it seriously? What are the tests that it has to pass? If you demand a proof that Laplace's theory is "correct," then I'm afraid I don't know what the question means. If you want to see a proof that it is the only possible extension of logic, then I would reply that it is surely not unique. But I think we have given fairly convincing arguments for the view that it is the only possible extension of logic which is internally consistent and represents degrees of plausibility by real numbers. You can, of course, hope to see

more rigorous and more general arguments than I have given; and I hope that you will. In this connection, let me just mention that the book of Savage (Savage, 1954) contains a great deal of this more refined analysis using measure theory, which is applicable to our problem.

How about other kinds of extensions of logic, in which we don't represent plausibility by real numbers? The possibilities of such "lattice theories" seem endless, and I want to say a little more about them in the last lecture. However, before dashing off to explore them, one should realize this: unless and until some specific failure of Laplace's theory is discovered, we have no rational basis for saying that a different theory is any better than the one we already have, and no clue to tell us in what way we should want another theory to be any different.

So, I would like to propose this as a working procedure. Let's take the good points of Savage's definition of "personalistic" and "necessary" views and combine them into a single definition; and above all, let's acknowledge their proper source:

(d) Laplace's Theory. Probability theory is an extension of logic which describes the consistent inductive reasoning of an idealized being who represents degrees of plausibility by real numbers. The numerical value of any probability $(A|B)$ will in general depend not only on A and B, but also on the entire background of other propositions that this being is taking into account. A probability assignment is "subjective" in the sense that it describes a state of knowledge rather than anything which could be measured in an experiment; but it is completely "objective" in the sense that it is independent of the personality of the user; two beings with the same total background of knowledge must assign the same probabilities.

Now for that other adjective, "naive". This is more difficult to discuss, because it is vague. A dictionary definition of naive is: "of unaffected simplicity." To call any mathematical theory naive in that sense is, I think, very great praise; and praise of which Laplace's theory is fully deserving. But I don't think Savage meant it in that way. I think he meant that Laplace did not hesitate to apply probability theory in all sorts of problems where a modern statistician would fear to tread. Our little excursion into jurisprudence is, no doubt, a good example. But, of course, if probability theory really is an extension of logic, there shouldn't be any restriction on the kind of problem treated; in principle, we ought to be able to apply it to any situation where plausible inference is needed. The only way of judging whether this is so, is simply to apply Laplace's theory to many specific situations, particularly those where the objectivists have warned us not to use it, and see for ourselves just how naive the results are, and whether the objectivist can produce any better results. We have already done some of this in the last three lectures, and many more examples will come up in later ones.

Lecture 8

POINT ESTIMATION WITH BINOMIAL AND POISSON DISTRIBUTIONS

In the next two lectures, I want to take up some applications of Bayes' theorem, and comparisons with maximum likelihood, that are less trivial mathematically and also correspond quite closely to situations faced by many experimentalists. The mathematics to be developed is applicable to a large class of different problems; and let's start by indicating two typical examples.

(A) Each week, a large number N of mosquitos is bred in a stagnant pond near this campus, and we set up a trap on the campus to catch some of them. Each mosquito lives less than a week, during which time it has the probability p of flying onto the campus, and once on the campus, it has the probability "a" of being caught in our trap. We count the numbers c_1, c_2, ... caught each week. What can we then say about the numbers n_1, n_2, ... on the campus each week, and what can we say about N?

(B) We have a radioactive source (say Co^{60} for example), which is emitting particles of some sort (say the γ-rays from Co^{60}). Each radioactive nucleus has the probability p of sending a particle through a counter in one second; and each particle passing through has the probability "a" of producing a count. From measuring the number c_1, c_2, ... of counts in different seconds, what can we say about the numbers n_1, n_2, ... actually passing through the counter

in each second, and what can we say about the strength of the source?

The common feature in these problems is that we have two "random games" played in succession, and we can observe only the outcome of the last one. From this, we are to make the best inferences we can about the original cause and the intermediate conditions, and I want to show how drastically these problems are changed by various changes in the prior information. In our estimates we will want to (1) state the "best" estimate possible on the data; and (2) make a statement about the accuracy of the estimate. These are the classical problems of "point estimation" and "interval estimation." In this lecture we will confine ourselves to point estimation, and take up the second aspect in the next lecture. I will speak in terms of the radioactive source problem, but it will be clear enough that the same arguments apply in many different problems.

8.1. A Simple Bayesian Estimate: Quantitative Prior Information.

First, let's discuss the efficiency of the counter, which I'll denote, as indicated above, by "a." By this I mean that each particle passing through the counter has independently the probability "a" of producing a count. The situation is therefore very much like that of sampling with replacement, discussed in Lecture 5, except that here there is no "urn" to shake, and so we will not question the validity of equations such as (5-34). From the logical standpoint, however, we still have to carry out a sort of bootstrap operation with regard to this quantity; for how is it determined? Intuitively, of course, you have no trouble at all in seeing how you could determine "a" from measurements on the counter. But from the standpoint of strictly logical development, we need to have the calculation about to be given before we can establish the precise connection between the value of "a" and observable quantities. So, for the time being we'll just have to suppose that "a" is a

given number, and later the result of our calculations will show us how it can be measured.

Now if we knew that n particles had passed through the counter, the probability, on this evidence, of getting exactly c counts, is obtained by repeated applications of our Rule 1 and Rule 2, in a way that is given in all the textbooks under the heading, "Bernoulli trials." The result is the binomial distribution that we have already derived in two ways, Equations (5-28) and (5-34). In our present notation, this is

$$(c|n) = \binom{n}{c} a^c (1-a)^{n-c} . \qquad (8-1)$$

In practice, there is a question of resolving time; if the particles come too close together we may not be able to see the counts as separate, either because of limited bandwidth in the detecting circuits or because the counter experiences a "dead time" after a count. These effects are important in many practical situations and there is a voluminous literature on the application of probability theory to them.* But we'll disregard those difficulties for this problem, and imagine that we have infinitely good resolving time (or, what is really the same thing, that the counting rate is so low that there is negligible probability of this happening.)

Now let's also introduce a quantity p which is the probability, in any one second, that any particular nucleus will emit a particle passing though the counter. We're going to assume the number of nuclei N so large and the and the half-life so long, that we don't have to consider N as a variable for this problem. So there are N nuclei, each of which has independently the probability p of sending a particle through our counter in any one second. The quantity p is also, for present purposes, just a given number, because we have not yet seen in terms of probability theory, the line of reasoning

*A bibliography on probability analysis of particle counters is given in appendix B.

by which we could convert experimental measurements on Co^{60} into a numerical value of p (but again, you see intuitively without any hesitation at all, that p is a way of describing the half-life of the source).

Suppose we were given N and p; what is the probability, on this evidence, that in any one second exactly n particles will pass through the counter? Well, that's exactly the same mathematical problem as the above one, so of course it has the same answer, the binomial distribution

$$(n|N,p) = \binom{N}{n} p^n (1-p)^{N-n} \tag{8-2}$$

But in this case there's a good approximation to the binomial distribution. Because the number N is enormously large and p is enormously small. In the limit where $N\to\infty$, $p\to 0$ in such a way that $Np\to s$ = constant, what happens to (8-2)? To find this, write $p = s/N$, and pass to the limit $N\to\infty$. Then

$$\frac{N!}{(N-n)!}\, p^n = N(N-1)\ldots(N-n+1)\left(\frac{s}{N}\right)^n$$

$$= s^n \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\ldots\left(1 - \frac{n-1}{N}\right)$$

which goes into s^n in the limit. Likewise,

$$(1 - p)^{N-n} = \left(1 - \frac{s}{N}\right)^{N-n} \to e^{-s}$$

and so the binomial distribution (8-2) goes over into the simpler Poisson distribution:

$$(n|N,p) = (n|s) = \frac{e^{-s}\, s^n}{n!} \tag{8-3}$$

and it will be handy for us to take this limit. The number s is essentially what the experimenter would call his "source strength."

Now we have enough "formalism" to start solving problems. Suppose we are not given the number of particles n in the counter, but only the source strength s. What is the probability, on this evidence, that we will see exactly c counts in any one second? As we noted in Lecture 6, Eq. (6-9), a

handy trick, which often works in problems of this sort, is to resolve the proposition c into a set of mutually exclusive alternatives; then apply Rule 3 as extended to Eq. (3-21), and then Rule 1. In this case, the propositions cn for all n form such a set, so we can write

$$(c|s) = \sum_{n=0}^{\infty} (cn|s) = \sum_{n=0}^{\infty} (c|ns)(n|s)$$
$$= \sum_{n=0}^{\infty} (c|n)(n|s) \qquad (8\text{-}4)$$

Evidently, if we knew the number of particles in the counter, it wouldn't matter any more what s was, so $(c|ns) = (c|n)$. This is perhaps made clearer by drawing a diagram, Fig. (8.1), which indicates the direction of causal influences; i.e., s partially determines the value of n, which in turn partially determines c; but there is no <u>direct</u> causal influence of s on c. Or, to put it still another way, s can influence c <u>only</u> via its intermediate effect on n.

Since we have worked out both $(c|n)$ and $(n|s)$, we just have to substitute them in, and we get

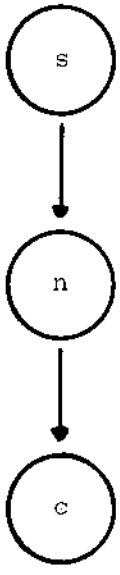

Figure 8.1. Direction of causal influences.

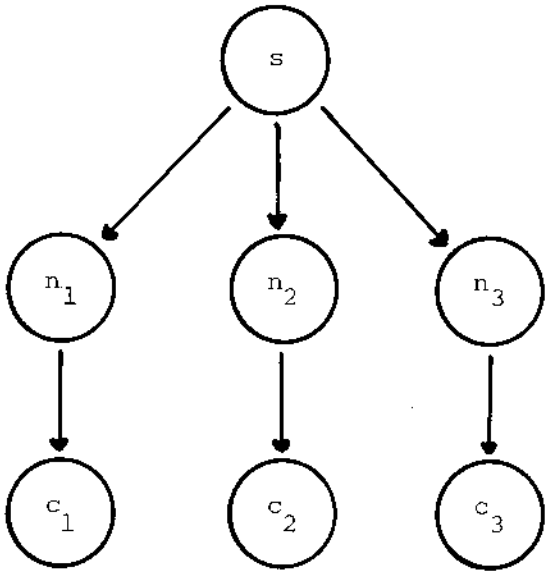

Figure 8.2. Causal influences in successive measurements.

$$(c|s) = \sum_{n=c}^{\infty} \left[\frac{n!}{c!\ (n-c)!}\ a^c\ (1-a)^{n-c} \right] \left[\frac{e^{-s}\ s^n}{n!} \right]$$

$$= \frac{e^{-s}\ a^c\ s^c}{c!} \sum_{n=c}^{\infty} \frac{[s\ (1-a)]^{n-c}}{(n-c)!} = \frac{e^{-s}\ (sa)^c}{c!}\ e^{s(1-a)}$$

or,

$$(c|s) = \frac{e^{-sa}\ (sa)^c}{c!} \tag{8-5}$$

This is a Poisson distribution with mean value

$$\bar{c} = \sum_{c=0}^{\infty} c\ (c|s) = sa. \tag{8-6}$$

Well, our result is not at all surprising. We have the Poisson distribution with a mean value which is the product of the source strength times the efficiency of the counter. Without going through the analysis, that's probably just the guess we would make.

In practice, it is c that is known and n that is unknown. If we knew the source strength s, and also the number of counts c, what would be the probability, on that evidence, that there were exactly n particles passing throggh the counter during that second? This is a problem which arises all the time in physics laboratories, because we may be using the counter as a "monitor," and have it set up so that the particles, after going through the counter, then initiate some other reaction which is the one we're really studying. Not if the particles are γ-rays, I'm afraid, but with almost every other kind of particles, this is an arrangement which has been used many times. It is important to get the best possible estimates of n, because that is one of the numbers we need in calculating the cross-section of this other reaction. Well, this is exactly the sort of problem for which Bayes' theorem was invented, so let's turn it over to our robot and see how he handles it. The probability he needs is

$$(n|cs) = (n|s)\ \frac{(c|ns)}{(c|s)} = \frac{(n|s)\,(c|n)}{(c|s)} \tag{8-7}$$

Again, everything we need for this calculation is on the board, so we just

have to substitute:

$$(n|cs) = \frac{\left[\frac{e^{-s}\, s^n}{n!}\right]\left[\frac{n!}{c!\,(n-c)!}\, a^c\, (1-a)^{n-c}\right]}{\left[\frac{e^{-sa}\,(sa)^c}{c!}\right]}$$

$$= \frac{e^{-s(1-a)}\,[s\,(1-a)]^{n-c}}{(n-c)!} \tag{8-8}$$

So you see the interesting thing is that we still have a Poisson distribution, with parameter s(1-a), but shifted upward by c; because of course, n could not be less than c. The mean value of this distribution is

$$\bar{n} = \sum_n n\ (n|cs) = c + s\ (1-a) \tag{8-9}$$

All right, so what is the best guess the robot can make as to the number of particles responsible for these c counts? In all problems of this sort where you want to make a definite decision, you want the robot to announce one number. There is a probability distribution which describes the robot's state of knowledge as to the number of particles. The number which he will publicly announce as his guess will, of course, depend on what are the consequences of being wrong. We will look at this aspect of the problem more closely later on, when we take up decision theory.

For the time being, we might ask the robot to take as a criterion that he should minimize the expected square of the error. If he announces the estimate n_{est}, but the true value is n, his error will be $(n_{est} - n)$, whose expected square is

$$\overline{(n_{est} - n)^2} = \overline{(n^2_{est} - 2n_{est}n + n^2)}$$

$$= n^2_{est} - 2n_{est}\bar{n} + \overline{n^2}$$

$$= (n_{est} - \bar{n})^2 + (\overline{n^2} - \bar{n}^2) \tag{8-10}$$

The second term $(\overline{n^2} - \bar{n}^2) = \overline{(n - \bar{n})^2}$ is called the <u>variance</u> of the distribution and it is fixed by (8-8) so the robot can do nothing to minimize it. But he can remove the first term entirely by taking as his estimate just the mean

value $n_{est} = \bar{n}$ that we just calculated in (8-9).

Evidently, this result holds generally whatever the form of the distribution; the mean square error criterion always leads to taking the mean value $\bar{n}$ (i.e., the "center of gravity" of the distribution) as his "best" guess. Or, if we ask him to state the one in which he believes most strongly, then he will take the most probable value, i.e. the one which maximizes (8-8). But the difference is negligible in this case, because in a Poisson distribution the most probable value (which we will denote by $\hat{n}$) always lies between $\bar{n}$ and $(\bar{n}-1)$. So, let's suppose that the mean value is the one he is to announce.

At this point, a statistician of the "orthodox" or "objectivistic" school of thought pays a visit to our laboratory. We describe the properties of the counter to him, and invite him to give us <u>his</u> best estimate as to the number of particles. He will, of course, use maximum likelihood because his textbooks have told him that (Cramer, 1946; p. 498): "From a theoretical point of view, the most important general method of estimation so far known is the method of maximum likelihood." His likelihood function is, in our notation, $(c|n)$. The value of n which maximizes it is found, within one unit, from setting

$$\frac{(c|n)}{(c|n-1)} = \frac{n\ (1-a)}{n-c} = 1$$

or

$$(n)_{\text{max. likelihood}} = \frac{c}{a} \tag{8-11}$$

You may find the difference between these two estimates rather startling, if we put in some numbers. Suppose our counter has an efficiency of 10 per cent; in other words, a = 0.1, and the source strength is s = 100 particles per second, so that the expected counting rate according to Equation (8-6) is $\bar{c}$ = 10 counts per second. But in this particular second, we got 15 counts.

What should we conclude about the number of particles? Well, probably the first answer one would give without thinking is that, if the counter has an efficiency of 10 per cent, then in some sense each count must have been due to about 10 particles; so if there were 15 counts, then there must have been about 150 particles. That is, as a matter of fact, exactly what the maximum likelihood estimate (8-11) would be in this case. But what does the robot tell us? Well, he says the best estimate is only

$$\bar{n} = 15 + 100\ (1 - 0.1) = 15 + 90 = 105\ . \qquad (8\text{-}12)$$

More generally, we could write Equation (8-9) this way:

$$\bar{n} = s + (c - \bar{c})\ ; \qquad (8\text{-}13)$$

if you see k more counts than you should have in one second, according to the robot that is evidence for only k more particles, not 10k.

This example turned out to be quite surprising to some experimental physicists engaged in work along these lines. Let's see if we can reconcile it with our common sense. If we have an average number of counts of 10 per second with this counter, then we would guess, by rules well known, that a fluctuation in counting rate of something like the square root of this, ±3, would not be at all surprising even if the number of incoming particles per second stayed strictly constant. On the other hand, if the average rate of flow of particles is s = 100 per second, the fluctuation in this rate which would not be surprising is about $\pm\sqrt{100} = \pm 10$. But this corresponds to only ±1 in the number of counts.

This shows that you cannot use a counter to measure fluctuations in the rate of arrival of particles, unless the counter has a very high efficiency. If the efficiency is high, then you know that practically every count corresponds to one particle, and you are reliably measuring the fluctuations in beam current. If the efficiency is low and you know that there is a definite, fixed source strength, then fluctuations in counting rate are much more likely

to be due to things happening in the counter than to actual changes in the rate of arrival of particles.

What caused the difference between the Bayes and maximum likelihood solutions? It's due to the fact that we had prior information contained in this source strength s. The maximum likelihood estimate simply maximizes the probability of getting c counts, given n particles, and maximizing that gives you 150. In Bayes' solution, we will multiply this by the prior probability, which represents our knowledge of the laws of radioactivity, before maximizing, and we'll get an entirely different value for the estimate. Prior information can make a big change in the conclusions we draw from a random experiment.

Now, we really have to apologize to the statistician at this point; what we did was not entirely fair to him. Because, of course, this number "s" does represent a substantial piece of quantitative information which we didn't let him use. I think that as soon as this comparison was out, his common sense would lead him to agree readily enough that in this problem the Bayes estimate was far superior to the maximum likelihood estimate, and he would not object to the use of Bayes' theorem. He would say that in this case we did have a good prior probability distribution, with an evident frequency interpretation (which we have not so far mentioned, because it has no bearing on the robot's problem), so that Bayes' theorem is perfectly valid.

But now I want to extend this problem a little bit, to a case where there is no quantitative prior information, but only one qualitative fact. We are now going to use Bayes' theorem in four problems where the "objectivist" statistician says categorically that use of Bayes' theorem is nonsense because it has no frequency interpretation; and again compare its results with the ones obtained by the statistician's methods.

8.2. Effect of Qualitative Prior Information.

Two robots, Mr. A and Mr. B, who have different amounts of prior information about the source of the particles, are watching this counter. The source is hidden in another room which they are not allowed to enter. Mr. A has no knowledge at all about the source of the particles; for all he knows, it might be an accelerating machine which is being turned on and off in an arbitrary way, or the other room might be full of little men who run back and forth, holding first one radioactive source, then another, up to the exit window. Mr. B has one additional qualitative fact; he knows that the source is a radioactive sample of long lifetime, in a fixed position. But he does not know anything about its source strength (except, of course, that it is not infinite because, after all, the laboratory is not being vaporized by its presence. Mr. A is also given assurance that he will not be vaporized during the experiment). They both know that the counter efficiency is 10 per cent. Again, we want them to estimate the number of particles passing through the counter, from knowledge of the number of counts. We denote their prior information by X_A, X_B respectively.

All right, we commence the experiment. During the first second, $c_1 = 10$ counts are registered. What can Mr. A and Mr. B say about the number n_1 of particles? Bayes' theorem for Mr. A reads,

$$(n_1|c_1X_A) = (n_1|X_A)\frac{(c_1|n_1X_A)}{(c_1|X_A)} = \frac{(n_1|X_A)(c_1|n_1)}{(c_1|X_A)} \tag{8-14}$$

The denominator is just a normalizing constant, and could also be written,

$$(c_1|X_A) = \sum_{n_1} (c_1|n_1)(n_1|X_A) \quad . \tag{8-15}$$

But now we seem to be stuck, for what is $(n_1|X_A)$? The only information about n_1 contained in X_A is that n_1 is not large enough to vaporize the laboratory. How can we assign prior probabilities on this kind of evidence? This has

been the point of controversy for a good long time, for in any frequency theory of probability, we certainly have no basis at all for assigning the probabilities $(n_1|X_A)$.

Now, of course, Mr. A is going to assign a uniform prior probability here, and our statistician friend will object on the grounds that this is a completely unwarranted assumption. He will say, "How do you know that all values of n_1 are equally likely? They might not be equally likely at all. You just don't know, and you have no basis for applying Bayes' theorem until you have found the correct prior probability distribution." Note that this is not because our friend has any particular dislike for a uniform distribution; for he would object just as strongly (and in fact, I suspect, even more strongly) to any other prior probability assignment we might propose to use. It would always seem, to him, like an unwarranted assumption which would invalidate all our conclusions.

I am belaboring this point because it lies at the heart of the most persistently held misconception about the Laplace-Bayes theory. Unless we understand clearly what we're doing when we assign a uniform prior probability, we're going to be faced with tremendous conceptual difficulties from here on. This is what Mr. A replies to the statistician:

"Your objection shows that the word 'probability' has entirely different meanings to you and me. When you say that I cannot apply Bayes' theorem until I have determined the 'correct' prior probability distribution, you are implying that the event n_1 possesses some intrinsic 'absolute' probability. I deny this. n_1 is what it is; simply an unknown number. The only meaning of the word 'probability' which makes any sense at all to me, is simply the best indication of the truth of a proposition, based on whatever evidence we do <u>in fact</u> have. To me, a probability assignment is not an assertion about experience, real or potential. When I say, 'the probability of event E is p,' I

am not describing any property of the event. I am describing my state of knowledge concerning the event.

"Now, evidently, each of us believes that the other is suffering from a very fundamental and dangerous confusion about the proper use of probability theory. But we can never settle this by philosophical arguments about the meaning of words. The only real way of settling the question, which of these conceptions of probability is best, is to put them to the test in specific problems. You say that my uniform prior probability assignment is foolish. If so, then it ought to lead to at least one foolish result. So I'm just going to ignore your warning and go ahead with my calculation. If I get a foolish result, then from studying how it happened, I can learn something. But if I get a sensible result, then maybe you are the one who can learn something.

"According to Bayes' theorem, I need to find the probability assignment $(n_1|X_A)$ which represents my state of knowledge before I observed that $c_1 = 10$ counts. At that time, n_1 might have been 0, 1, 137, 2069, or 10^5 for all I knew. There was nothing in my prior knowledge which would justify saying that any one of those was more likely than any other, and assigning the same probability to all of them is simply my way of stating that fact. n_1 might easily have been as large as 10^7, for all I knew. But there is some upper limit N, for which I knew that $n_1 < N$. For example, if n_1 had been $10^{10^{10}}$, then not only the laboratory, but our entire galaxy, would have been vaporized by the energy in the beam. I could justify a considerably lower value of N than that, and if it turns out to make a difference in my conclusions, I'll have to think harder about just how low I could take it. But before going to all that work, I'd better find out whether it does make any difference. So, I'll just take

$$(n_1|X_A) = \begin{cases} \frac{1}{N}, & 0 \le n_1 < N \\ 0, & N \le n_1 \end{cases} \tag{8-16}$$

and see what Bayes' theorem gives me."

Well, Mr. A turns out to be lucky, for nicely enough, the 1/N cancels out of Equations (8-14), (8-15), and we are left with

$$(n_1|c_1\ X_A) = \begin{cases} \dfrac{(c_1|n_1)}{\sum_{n_1=0}^{N-1} (c_1|n_1)} & ,\quad 0 \le n_1 < N \\ 0 & ,\quad N \le n_1 \end{cases} \tag{8-17}$$

We have noted, in Equation (8-11), that as a function of n, $(c|n)$ attains its maximum at $n = c/a$ (=100, in this problem). For n large compared to this, $(c|n)$ falls off like $n^c(1-a)^n \simeq n^c\, e^{-an}$. Therefore, the sum in (8-17) converges so rapidly that if N is as large as a few hundred, there is no appreciable difference between

$$\sum_{n=0}^{N-1} (c|n) \qquad \text{and} \qquad \sum_{n=0}^{\infty} (c|n)$$

So, unless the prior information could justify an upper limit N lower than about 200, the value of N turns out not to make any difference. The sum to infinity is easily evaluated, and we get the result

$$(n_1|c_1X_A) = a\ (c_1|n_1) = \binom{n_1}{c_1} a^{c_1+1}\ (1-a)^{n_1-c_1}\ . \tag{8-18}$$

So, to Mr. A, the most probable value of n_1 is the same as the maximum-likelihood estimate:

$$(\hat{n}_1)_A = \frac{c}{a} = 100 \tag{8-19}$$

while the mean value estimate is calculated as follows:

$$\overline{n}_1 - c_1 = \sum_{n_1=c_1}^{\infty} \frac{n_1!}{c_1!\ (n_1-c_1-1)!}\, a^{c_1+1}\ (1-a)^{n_1-c_1}$$

$$= a^{c_1+1}\ (1-a)(c_1+1) \sum_{n_1=c_1+1}^{\infty} \binom{n_1}{n_1-c_1-1} (1-a)^{n_1-c_1-1}\ .$$

The sum is equal to

$$\sum_{m=0}^{\infty} \binom{m+c_1+1}{m} (1-a)^m = \sum_{m=0}^{\infty} (-)^m \binom{-c_1-2}{m} (1-a)^m$$

$$= [1 - (1-a)]^{-c_1-2} = \frac{1}{a^{c_1+2}} \tag{8-20}$$

and, finally, we get

$$(\bar{n}_1)_A = c_1 + (c_1+1)\,\frac{1-a}{a} = \frac{c_1+1-a}{a} = 109 \quad . \tag{8-21}$$

Now, how about the other robot, Mr. B? Does his extra knowledge help him here? He knows that there is some definite source strength s. And, because the laboratory is not being vaporized, he knows that there is some upper limit S_o. Suppose that he assigns a uniform prior probability density for $0 \leq s < S_o$. Then he will obtain

$$(n_1|X_B) = \int_0^{\infty} (n_1|s)(s|X_B)\, ds = \frac{1}{S_o} \int_0^{S_o} (n_1|s)\, ds$$

$$= \frac{1}{S_o} \int_0^{S_o} \frac{s^{n_1} e^{-s}}{n_1!}\, ds \quad . \tag{8-22}$$

Now, if n_1 is appreciably less than S_o, the upper limit of integration can for all practical purposes, be taken as infinity, and the integral is just unity. So, we have

$$(n_1|X_B) = (s|X_B) = \frac{1}{S_o} = \text{const., if } n_1 < S_o \quad . \tag{8-23}$$

In putting this into Bayes' theorem with $c_1 = 10$, the significant range of values of n_1 will be of the order of 100, and unless S_o is lower than about 200, we will have exactly the same situation as before; Mr. B's extra knowledge didn't help him at all, and he comes out with exactly the same distribution and the same estimates:

$$(n_1|c_1X_B) = (n_1|c_1X_A) = a\,(c_1|n_1) \quad . \tag{8-24}$$

Jeffreys (1939; Chap. 3) has proposed a different way of handling this problem. He suggests that the proper way to express "complete ignorance" of

a continuous variable known to be positive, is to assign uniform prior probability to its logarithm; i.e. the prior probability density is

$$(s|X_J) = \frac{1}{s} \tag{8-25}$$

Of course, you can't normalize this, but that doesn't stop you from using it, because when we expand the denominator of Bayes' theorem as in (8-15), we see that the prior probability appears in both numerator and denominator [the same reason that N cancelled out of (8-17)]. So, in applying Bayes' theorem, it doesn't really matter whether the prior probabilities are normalized or not.

Jeffreys justified (8-25) on the grounds of invariance under certain changes of parameters; i.e. instead of using the parameter s, what prevents us from using $t \equiv s^2$, or $u \equiv s^3$? Evidently, to assign a uniform prior probability density to s, is not at all the same thing as assigning a uniform prior probability to t; but if we use the Jeffreys prior, we are saying the same thing whether we use s or any power s^m as the parameter. There is the germ of an important principle here; but it was only recently that the situation has been fairly well understood. When we take up the theory of transformation groups later on, we will see that the real justification of Jeffreys' rule cannot lie merely in the fact that the parameter is positive; but that our desideratum of consistency in the sense (b) of Lecture 2 (p. 26) uniquely determines the Jeffreys rule in the case when s is a "scale parameter." The question then reduces to whether s can properly be regarded as a scale parameter in this problem. However, this takes us far beyond the present topic, so I don't want to spend a lot of time now arguing either for or against (8-25); but, in the spirit of this problem, we can put it to the test and see what it gives. The calculations are all very easy, and we find these results:

$$(n_1|X_J) = \frac{1}{n_1}, \qquad (c_1|X_J) = \frac{1}{c_1}$$

$$(n_1|c_1X_J) = \frac{c_1}{n_1}(c_1|n_1) \quad . \tag{8-26}$$

This leads to the most probable and mean value estimates:

$$(\hat{n}_1)_J = \frac{c_1 - 1 + a}{a} = 91 \qquad (8\text{-}27)$$

$$(\bar{n}_1)_J = \frac{c}{a} = 100 \quad . \qquad (8\text{-}28)$$

The amusing thing emerges that Jeffreys' prior probability rule just lowers the most probable and mean value by 9 each, bringing the mean value right back to the maximum likelihood estimate!

This comparison is valuable in showing us how little difference there is numerically between the consequences of different prior probability assignments which are not sharply peaked, and helps to put arguments about them into proper perspective. We made a rather drastic change in the prior probabilities, in a problem where there was really very little information contained in the result of the random experiment, and it still made less than 10 per cent difference in the result. This is, as we will see in the next lecture, small compared to the probable error in the estimate which was inevitable in any event. In a more realistic problem where a random experiment is repeated many times to give us a good deal more information, the difference would be very much smaller still. So, from a pragmatic standpoint, the arguments about which prior probabilities correctly express a state of "complete ignorance" usually amount to quibbling over pretty small peanuts.* From the standpoint of principle, however, they are very important and have to be thought about a great deal.

Now we are ready for the interesting part of this problem. For during the next second, we see $c_2 = 16$ counts. What can Mr. A and Mr. B now say about the numbers n_1, n_2, of particles responsible for c_1, c_2? Well, Mr. A has no reason to expect any relation between what happened in the two time

*This is most definitely not true if the prior probabilities are to describe a definite piece of prior knowledge, as the next example shows.

intervals, and so to him the increase in counting rate is evidence only of an increase in the beam intensity. His calculation for the second time interval is exactly the same as before, and he will give as the most probable value

$$(\hat{n}_2)_A = \frac{c_2}{a} = 160 \tag{8-29}$$

and his mean value estimate is

$$(\bar{n}_2)_A = \frac{c_2 + 1 - a}{a} = 169 \quad . \tag{8-30}$$

Knowledge of c_2 doesn't help him to get any improved estimate of n_1, which stays the same as before.

But now, Mr. B is in an entirely different position than Mr. A; his extra qualitative information suddenly becomes very important. For knowledge of c_2 enables him to improve his previous estimate of n_1. Bayes' theorem now gives

$$(n_1|c_2c_1X_B) = (n_1|c_1X_B)\,\frac{(c_2|n_1c_1X_B)}{(c_2|c_1X_B)}$$

$$= (n_1|c_1X_B)\,\frac{(c_2|n_1X_B)}{(c_2|c_1X_B)} \tag{8-31}$$

Again, the denominator is just a normalizing constant, which we can find by summing the numerator. We see that the significant thing is $(c_2|n_1X_B)$. Using our trick of resolving c_2 into mutually exclusive alternatives, this is

$$(c_2|n_1X_B) = \int_0^\infty (c_2s|n_1X_B)\,ds = \int_0^\infty (c_2|sn_1)(s|n_1)\,ds$$

$$= \int_0^\infty (c_2|s)(s|n_1)\,ds \quad . \tag{8-32}$$

We have already found $(c_2|s)$ in Equation (3-7), and we need only

$$(s|n_1) = (s|X_B)\,\frac{(n_1|s)}{(n_1|X_B)} = (n_1|s) \quad , \quad \text{if } n_1 \ll S_o \tag{8-33}$$

where we have used Equation (8-23). We have found $(n_1|s)$ in Equation (8-3), so we have

$$(c_2|n_1X_B) = \int_0^\infty \left[\frac{e^{-sa}(sa)^{c_2}}{c_2!}\right]\left[\frac{e^{-s}s^{n_1}}{n_1!}\right] ds = \binom{n_1+c_2}{c_2} \frac{a^{c_2}}{(1+a)^{n_1+c_2+1}} . \quad (8\text{-}34)$$

Now we just substitute (8-24) and (8-34) into (8-31), carry out an easy summation to get the denominator, and the result is

$$(n_1|c_2c_1X_B) = \frac{(2a)^{c_1+c_2+1}}{(c_1+c_2)!\ (1-a)^{c_1}\ (1+a)^{c_2+1}} \frac{(n_1+c_2)!}{(n_1-c_1)!} \left(\frac{1-a}{1+a}\right)^{n_1} . \quad (8\text{-}35)$$

Note that we could also have derived this by direct application of our trick:

$$(n_1|c_2c_1X_B) = \int_0^\infty (n_1s|c_2c_1X_B)\, ds = \int_0^\infty (n_1|sc_1)(s|c_2c_1)\, ds . \quad (8\text{-}36)$$

We have already found $(n_1|sc_1)$ in (8-8), and it is easily shown that $(s|c_2c_1)$ = (const.) x $(c_2|s)(c_1|s)$, which is therefore given by (8-5). This, of course, leads to the same result (8-35); this provides another test of the consistency of our rules, which we sought to ensure by the functional equation arguments in Lecture 3.

To find Mr. B's new most probable value of n_1, we set

$$\frac{(n_1|c_2c_1X_B)}{(n_1-1|c_2c_1X_B)} = \frac{n_1 + c_2}{n_1 - c_1}\frac{1 - a}{1 + a} = 1 ,$$

or,

$$(\hat{n}_1)_{B_2} = \frac{c_1}{a} + (c_2 - c_1)\frac{1-a}{2a}$$

$$= \frac{c_1 + c_2}{2a} + \frac{c_1 - c_2}{2}$$

$$= 127 \quad (8\text{-}37)$$

His new mean-value estimate is also readily calculated, and is equal to

$$(\bar{n}_1)_{B_2} = \frac{c_1 + 1 - a}{a} + (c_2 - c_1 - 1)\frac{1 - a}{2a}$$

$$= \frac{c_1 + c_2 + 1 - a}{2a} + \frac{c_1 - c_2}{2}$$

$$= 131.5 \quad . \tag{8-38}$$

You see that both estimates are considerably raised, and the difference between most probable and mean value is only half what it was before. If we want Mr. B's estimates for n_2, then from symmetry we just interchange the subscripts 1 and 2 in the above equations. This gives for his most probable and mean value estimates, respectively,

$$(\hat{n}_2)_B = 135 \tag{8-39}$$

$$(\bar{n}_2)_B = 137.5 \tag{8-40}$$

Now, can we understand what is happening here? Intuitively, the reason why Mr. B's extra qualitative prior information makes a difference is that knowledge of both c_1 and c_2 enables him to make a better estimate of the source strength s, which in turn is relevant for estimating n_1. The situation is indicated more clearly by the diagrams, Fig. (8.2). To Mr. A, each sequence of events $n_i \rightarrow c_i$ is entirely independent of the others, so knowledge of one doesn't help him in reasoning about any other. In each case, he must reason from c_i directly to n_i, and no other route is available. But to Mr. B, there are two routes; he can reason directly from c_1 to n_1 as Mr. A does, as described by $(n_1|c_1X_A) = (n_1|c_1X_B)$; but because of his knowledge that there is a fixed source strength s "presiding over" both n_1 and n_2, he can also reason along the route $c_2 \rightarrow n_2 \rightarrow s \rightarrow n_1$. If this were the only route available to him (i.e., if he didn't know c_1), he would obtain the distribution

$$(n_1|c_2X_B) = \int_0^\infty (n_1|s)(s|c_2X_B)\, ds$$

$$= \frac{a^{c_2+1}}{c_2!\ (1+a)^{c_2+1}} \ \frac{(n_1 + c_2)!}{n_1!\ (1+a)^{n_1}} \tag{8-41}$$

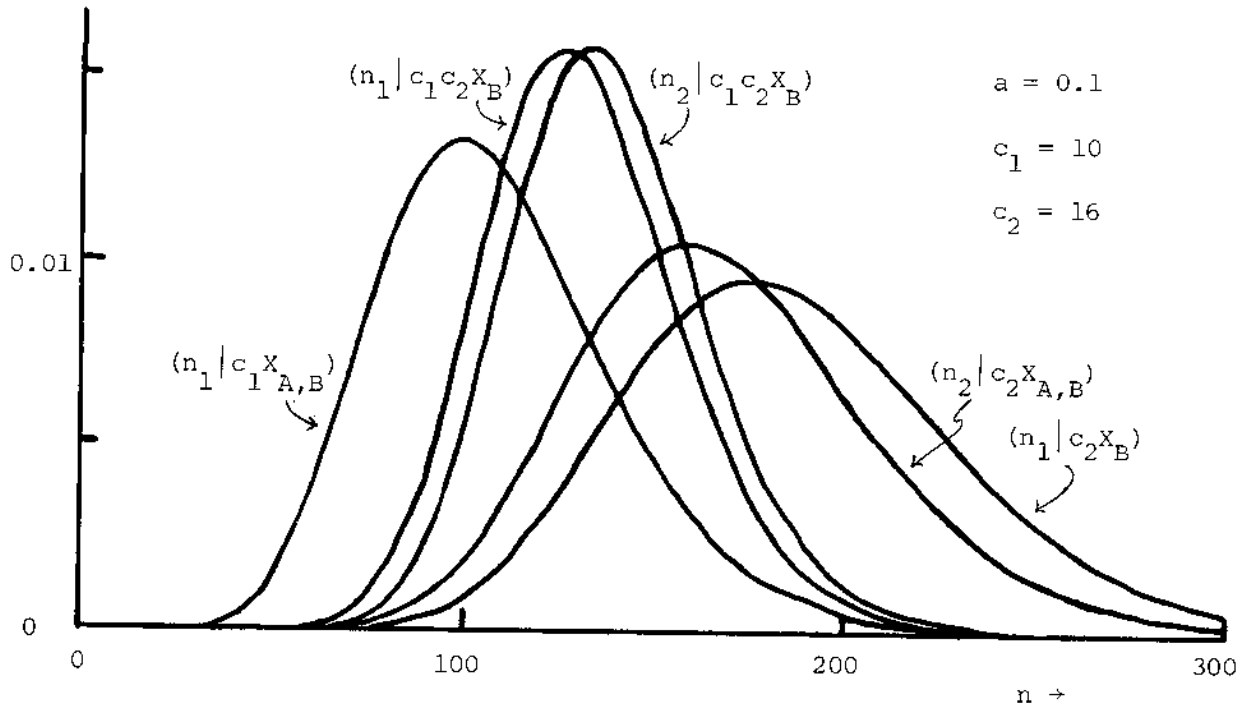

Figure 8.3. The various distributions (8-18), (8-35), (8-41), showing the effect of varying prior information.

and, comparing the above relations, we see that Mr. B's final distribution (8-35) is, except for normalization, just the product of the ones found by reasoning along his two routes:

$$(n_1|c_1c_2X_B) = (\text{const.}) \cdot (n_1|c_1X_B)(n_1|c_2X_B) \tag{8-42}$$

The information (8-41) about n_1 obtained by reasoning along the new route $c_2 \to n_2 \to s \to n_1$ thus introduces a "correction factor" in the distribution obtained from the direct route $c_1 \to n_1$, enabling Mr. B to improve his estimates.

This suggests that, if Mr. B could obtain the number of counts in a great many different seconds, c_3, c_4, ..., c_m, he would be able to do better and better; and perhaps in the limit $m \to \infty$ his estimate of n_1 might become as good as the one we found from Eq. (8-8), in which the source strength was considered known exactly. In the next Lecture we will check this surmise by working out the degree of reliability of these estimates, and by generalizing these distributions to arbitrary m, from which we can obtain the asymptotic forms.

Lecture 9

INTERVAL ESTIMATION AND ASYMPTOTIC PROPERTIES

There is still an essential feature missing in the comparison of Mr. A and Mr. B in our particle-counter problem. We would like to have some measure of the degree of reliability which they attach to their estimates, especially in view of the fact that their estimates are so different. Clearly, the best way of doing this would be to draw the entire probability distributions

$$(n_1 | c_2 c_1 X_A) \qquad \text{and} \qquad (n_1 | c_2 c_1 X_B)$$

and from this make statements of the form, "90 per cent of the posterior probability is concentrated in the interval $\alpha < n_1 < \beta$." But, for present purposes, we will be content to give the standard deviations [i.e., the square root of the variance as defined in Eq. (8-10)] of the various distributions we have found. An inequality due to Tchebycheff then asserts that, if σ is the standard deviation, then the amount p of probability concentrated between the limits $(\overline{n}_1 \pm t\sigma)$ satisfies

$$p \geq 1 - \frac{1}{t^2} \tag{9-1}$$

This tells us nothing when $t \leq 1$, but it tells us more and more as t increases beyond unity. For example, at least 3/4 of the probability must be assigned to the nange $\overline{n} \pm 2\sigma$, and at least 8/9 to the range $\overline{n} \pm 3\sigma$.

9.1. Calculation of Variance.

The variances σ^2 of all the distributions we have found in the last

lecture are readily calculated. In fact, the calculation of any moment of these distributions is easily performed by making use of the general formula

$$\sum_{m=0}^{\infty} \binom{m+a}{m} {}^{n}x^{m} = \left(x \frac{d}{dx}\right)^{n} \frac{1}{(1-x)^{a+1}}, \quad |x| < 1, \tag{9-2}$$

which we have already used in calculation of the mean value in (8-21). For Mr. A and Mr. B, and the Jeffreys prior probability distribution, we find the variances

$$\text{Var } (n_1 | c_1 X_A) = \frac{(c_1+1)(1-a)}{a^2} \tag{9-3}$$

$$\text{Var } (n_1 | c_2 c_1 X_B) = \frac{(c_1+c_2+1)(1-a^2)}{4a^2} \tag{9-4}$$

$$\text{Var } (n_1 | c_1 X_J) = \frac{c_1(1-a)}{a^2} \tag{9-5}$$

and the variances for n_2 are found from symmetry.

This has been a rather long discussion, so let's summarize all our results so far in a table. I'll give, for problem 1 and problem 2, the most probable values of number of particles as found by Mr. A and Mr. B, and also the (mean value) ± (standard deviation), which provides a reasonable interval estimate.

From this table we see that Mr. B's extra information not only has led him to change his estimates considerably from those of Mr. A, but it has enabled him to make an appreciable decrease in his probable error. Prior information which has nothing to do with frequencies can greatly alter the conclusions we draw from a random experiment, and their degree of reliability.

It is also of interest to ask how good Mr. B's estimate of n_1 would be if he knew only c_2; and therefore had to use the distribution (8-41) representing reasoning along the route $c_2 \to n_2 \to s \to n_1$ of Fig. (8.2). From (8-41) we find the most probable, and the (mean) ± (standard deviation) estimates

		Problem 1 $c_1 = 10$	Problem 2 $c_1 = 10$ $c_2 = 16$	
		n_1	n_1	n_2
A	most prob.	100	100	160
	mean ± s.d.	109 ± 31	109 ± 31	169 ± 39
B	most prob.	100	127	133
	mean ± s.d.	109 ± 31	131.5 ± 26	137.5 ± 26
J	most prob.	91		
	mean ± s.d.	100 ± 30		

$$\hat{n}_1 = \frac{c_2}{a} = 160 \qquad (9\text{-}6)$$

$$\text{mean} \pm \text{s.d.} = \frac{c_2+1}{a} \pm \frac{\sqrt{(c_2+1)(a+1)}}{a} = 170 \pm 43.3 \qquad (9\text{-}7)$$

In this case he would obtain slightly poorer estimate (i.e. a larger probable error) than Mr. A even if the counts $c_1 = c_2$ were the same, because the variance (9-3) for the direct route contains a factor (1-a), which gets replaced by (1+a) if we have to reason over the indirect route. Thus, if the counter has low efficiency, the two routes give nearly equal reliability for equal counting rates; but if it has high efficiency, a ≃ 1, then the direct route $c_1 \rightarrow n_1$ is far more reliable. I think your common sense will tell you that this is just as it should be.

9.2. Generalization and Asymptotic Forms.

Now in the last lecture we conjectured that Mr. B might be helped a good deal more in his estimate of n_1 by acquiring still more data $\{c_3, c_4, \ldots, c_m\}$. Let's investigate that further. The standard deviation of the distribution (8-8) in which the source strength was known exactly, is only $\sqrt{s(1-a)} = 10.8$ for $s = 130$; and from the table, Mr. B's standard deviation for his estimate of n_1 is now about 2.5 times this value. What would happen if we gave him more and more data from other time intervals, such that his estimate of s approached 130? To answer this, note that, if $1 \le k \le m$, we have (now dropping the X_B because we will be concerned only with Mr. B from now on):

$$(n_k|c_1\ldots c_m) = \int_0^\infty (n_k s|c_1\ldots c_m)\, ds$$

$$= \int_0^\infty (n_k|sc_k)(s|c_1\ldots c_m)\, ds \qquad (9\text{-}8)$$

in which we have put $(n_k|sc_1\ldots c_m) = (n_k|sc_k)$ because, from Fig. (8.2), if s is known, then all the c_i with $i \neq k$ are irrelevant for inference about n_k. The second factor in the integrand of (9-8) can be evaluated by Bayes' theorem:

$$(s|c_1\ldots c_m) = (s|X_B)\,\frac{(c_1\ldots c_m|s)}{(c_1\ldots c_m|X_B)}$$

$$= (\text{const.})\cdot(s|X_B)(c_1|s)(c_2|s)\ldots(c_m|s)$$

Using (8-5) and normalizing, this reduces to

$$(s|c_1\ldots c_m) = \frac{(ma)^{c+1}}{c!}\, s^c\, e^{-msa} \qquad (9\text{-}9)$$

where $c \equiv c_1 + \ldots + c_m$ is the total number of counts in the m seconds.

Let's note in passing the properties of this distribution. The most probable, mean, and variance of the distribution (9-9) are respectively

$$\hat{s} = \frac{c}{ma} \qquad (9\text{-}10)$$

$$\bar{s} = \frac{c+1}{ma} \qquad (9\text{-}11)$$

$$\mathrm{var}(s) = \overline{s^2} - \bar{s}^2 = \frac{c+1}{m^2 a^2} = \frac{\bar{s}}{ma} \tag{9-12}$$

So it turns out, as we might have expected, that as $m \to \infty$, the distribution $(s|c_1 \ldots c_m)$ becomes sharper and sharper, the most probable and mean value estimates of s get closer and closer together, and in the limit we would have just a δ-function:

$$(s|c_1 \ldots c_m) \to \delta(s-s')$$

where

$$s' \equiv \lim_{m\to\infty} \frac{c_1 + c_2 + \ldots + c_m}{ma} \tag{9-13}$$

So, in the limit, Mr. B does acquire exact knowledge of the source strength.

Returning to (9-8), both factors in the integrand are now known from (8-8) and (9-9), and so

$$(n_k|c_1 \ldots c_m) = \int_0^\infty \frac{e^{-s(1-a)}[s(1-a)]^{n_k-c_k}}{(n_k-c_k)!} \frac{(ma)^{c+1}}{c!} s^c e^{-msa}\, ds$$

or

$$(n_k|c_1 \ldots c_m) = \frac{(n_k-c_k+c)!}{(n_k-c_k)!\,c!} \frac{(ma)^{c+1}}{(1+ma-a)^{n_k-c_k+c+1}} (1-a)^{n_k-c_k} \tag{9-14}$$

which is the promised generalization of (8-35). In the limit $m \to \infty$, $c \to \infty$, $(c/ma) \to s' = \text{const.}$, this goes into the Poisson distribution

$$(n_k|c_1 \ldots c_m) \to \frac{e^{-s'(1-a)}}{(n_k-c_k)!}[s'(1-a)]^{n_k-c_k} \tag{9-15}$$

which is identical with (8-8). We therefore confirm that, given enough additional data, Mr. B's standard deviation can be reduced from 26 to 10.8, compared to Mr. A's 31.

For finite m, the mean value estimate of n_k from (9-14) is

$$\bar{n}_k = c_k + \bar{s}(1-a) \tag{9-16}$$

where $\bar{s} = (c+1)/ma$ is the mean value estimate of s from (9-11). Equation (9-16), which is to be compared to (8-9), includes (8-21) and (8-38) as special cases. Likewise, the most probable value of n_k according to (9-14), is

$$\hat{n}_k = c_k + \hat{s}(1-a) \tag{9-17}$$

where $\hat{s}$ is given by (9-10).

Note that Mr. B's revised estimates in problem 2 still lie within the range of reasonable error assigned by Mr. A. It would be rather disconcerting if this were not the case, as it would then appear that probability theory is giving Mr. A an unduly optimistic picture of the reliability of his estimates. There is, however, no theorem which guarantees this; for example, if the counting rate had jumped to $c_2 = 80$, then Mr. B's revised estimate of n_1 would be far outside Mr. A's limits of reasonable error. But in this case, Mr. B's common sense would lead him to doubt the reliability of his prior information X_B; we would have another example like that in Lecture 6, of a problem where one of those alternative hypotheses down at -100 db, which we don't even bother to formulate until they are needed, is resurrected by very unexpected new evidence.

9.3. Comparison of Bayesian and Orthodox Results.

Well, in the last lecture I said I was going to compare the results of Bayes' theorem with those obtained by the orthodox statistician's methods in this problem. I have already done that in the case of Mr. A; for his most probable values of n_1 and n_2 were in all cases just the same as the direct maximum likelihood estimates. The statistician accepts Bayes' theorem in the initial example where the source strength was known. He rejects it in the problem where the source strength was unknown, and says that (Wald, 1941): "These problems cannot be solved by any theorems of the calculus of probabi-

lities alone. Their solution requires some additional principles besides the axioms on which the calculus of probabilities is based." The new principle which he introduces is maximum likelihood; but mathematically, he ends up doing exactly what he would have done if he had stayed with Bayes' theorem. In order to form some idea of the degree of reliability of the estimate, he introduces still another ad hoc principle, the confidence interval. Our robot obtains all of these results automatically, by application of a single principle which is contained in the calculus of probabilities, as formulated by Laplace.

But how does this comparison look in the case of Mr. B? We have seen how Bayes' theorem automatically "digests" his qualitative prior information: X_B = "there is a constant but unknown source strength s," and how it enables him to improve his estimates and lower his probable error. How would the orthodox statistician make use of this information? In the first place, his ideology forbids him to use any of the equations (8-22), (8-23), (8-32), (8-36), (8-41), (9-8), (9-9) which formed the backbone of our various derivations, for he contends that "Probability statements can be made only about random variables. It is meaningless to speak of the probability that s lies in a certain interval, because s is not a random variable, but only an unknown constant." According to his doctrines, the distinction between a "random" and a "nonrandom" quantity is very essential; the methods he will use for inference, (and the conclusions he will arrive at,) depend on his decision as to which quantities are random, which are not.

I want to point out some difficulties with this position in a minute; however, right now our job is not to criticize the orthodox statistician's methods, but to describe them. If he refuses to use Bayes' theorem in the way our robot did, how would he handle it? I can't really be sure; and in fact I'll wager that different statisticians would handle it in different ways, because orthodox teaching has just not produced any unique method for such

problems. But I think I can suggest one ad hoc procedure that he might invent, and which most of his colleagues would accept. Consider the problem where we know that $c_1 = 10$, $c_2 = 16$. If anyone were to refuse to use the prior information X_B, on the grounds that it does not consist of frequency data, then he would have little choice but to estimate n_1 and n_2 by direct maximum likelihood, i.e., by maximization of $(c_1|n_1)$ and $(c_2|n_2)$; and it would collapse back to the problem of Mr. A. But, as I said in Lecture 4, if we do have prior information which is clearly relevant to the problem, common sense will tell all but the most pedantic not to use direct maximum-likelihood estimation. Without departing from orthodox principles, one can use the prior information X_B to formulate the problem in a different way. Here is one possible line of reasoning that he might use.

"The unknown constant s determines the objective statistical properties of n and c; i.e., the relative frequencies with which the random variables n and c would assume various values in the long run. Therefore, if I knew the value of s, it would be perfectly legitimate to use Bayes' theorem in the form

$$(n_1|c_1c_2s) = (n_1|s)\,\frac{(c_1c_2|n_1s)}{(c_1c_2|s)} \tag{9-18}$$

since every probability here has a clear frequency interpretation. Furthermore, since

$$(c_1c_2|n_1s) = (c_1|c_2n_1s)(c_2|n_1s) = (c_1|n_1)(c_2|s)$$

and

$$(c_1c_2|s) = (c_1|c_2s)(c_2|s) = (c_1|s)(c_2|s) \quad , \tag{9-19}$$

the calculation would reduce to

$$(n_1|c_1c_2s) = \frac{(n_1|s)(c_1|n_1)}{(c_1|s)} = (n_1|c_1s) \tag{9-20}$$

i.e., if s is known, then knowledge of c_2 is not relevant for estimation of n_1.

This leads, according to equation (8-9), to the mean-value estimate

$$\bar{n}_1 = c_1 + s(1-a) \ . \qquad (9\text{-}21)$$

Now if I had a reasonable estimate of s, then substituting it into (9-21) should give me an estimate of n_1 which is in some sense equally reasonable. So, instead of estimating n_1 by direct maximum-likelihood, I'll use an indirect method: first estimate s by maximum-likelihood, and use the result in (9-21)."

From (9-19) and (8-5) we have

$$\log (c_1 c_2 | s) = (c_1 + c_2) \log s - 2sa + (\text{const.})$$

where the (const.) is independent of s. So, the maximum-likelihood estimate of s, given c_1 and c_2, is found from $\frac{d}{ds} \log (c_1 c_2 | s) = 0$, or

$$(s)_{\text{max. likelihood}} = \frac{c_1 + c_2}{2a} = \frac{10 + 16}{2 \times 0.1} = 130$$

and his estimate of n_1 is then

$$\bar{n}_1 = 10 + 130(1 - 0.1) = 127 \ , \qquad (9\text{-}22)$$

which is the same as Mr. B's most probable value (8-37)! The fact that these estimates turn out exactly the same is, of course, fortuitous; but we see from equations (9-10) and (9-17) that in this problem the agreement would still hold no matter how many counts $\{c_1, c_2, \ldots, c_m\}$ had been observed.

This comparison shows how, in practice, the orthodox statistician who uses a little common sense in formulating the problem, can often manage to get very acceptable results and make use of his prior information without ever using a probability for a "nonrandom" quantity. But if now we asked him to make some definite statements about the reliability of the estimate (9-22), he would be faced with a quite sticky problem. He would probably set up a confidence interval to describe the uncertainty in s; but then he would have to find some way of "folding" this uncertainty with the uncertainty in n_1, inherent in (9-21) even when s is known exactly. I will not presume to guess how he would do this; again, since orthodox teaching has produced no unique

way of handling such problems, we can be pretty sure that different workers would do it in different ways, and come out with different conclusions. With reasonable common sense, however, the orthodox conclusions would not differ greatly from the ones our robot obtains from the posterior probability distribution $(n_1 | c_1 c_2 X_B)$. From a purely pragmatic standpoint, which sees no value in the fact that the robot's method comes from a more general and unified set of basic principles, the robot's procedure still has the advantage that he obtains all of these results from a single elementary calculation.

There is a further point which should be made on these estimation problems. We have seen that the most probable value and the mean value estimates are not the same in general. Which is best? The answer, evidently, depends on the use to be made of the theory, and on the form of the posterior probability distribution. For example, in Figure (9.1a) we have a distribution for which the most probable value is not only intuitively a poorer estimate than the mean value, but is also very unstable; very small changes in the data could tilt the curve the other way, making a large change in the estimate, which seems like a clear violation of common sense. But in Figure (9.1b) we have a case where the most probable value is quite likely to be the correct one, while the mean value is known to be an impossible one. In all cases, however, the mean value is the estimate which minimizes the expected square of the error. Generally, if the distribution has a single peak, the mean value would seem preferable. At any rate, any principle which denies us the choice between them cannot possibly be the best in all cases. We are concerned here with value judgments rather than inference; this will be studied in more detail when we consider decision theory.

In summary, what can we now say about the principle of maximum likelihood? If you ask a statistician about these things, one answer you are likely to get is that the real justification of maximum likelihood is not found in problems

of the sort just examined, but in its asymptotic properties, as we accumulate more and more random data. But, of course, in that limit the various "laws of large numbers" guarantee that all these methods approach the same thing. Indeed, in the "large sample" limit the evidence stares you in the face, and anybody can see what general conclusions are indicated, with hardly any need for a formal statistical theory. Scientists and engineers have been getting along fairly well for a long time without statistical training, for just that reason. It is in the small and medium sample case we considered here, that our unaided common sense lacks sufficient discrimination, and we need the guidance of a mathematical theory in order to make definite and defensible judgments.

In any event, whatever desirable properties maximum likelihood might have, asymptotic or otherwise, are also enjoyed by Bayes' theorem with uniform prior probabilities, because mathematically they amount to the same thing. But it

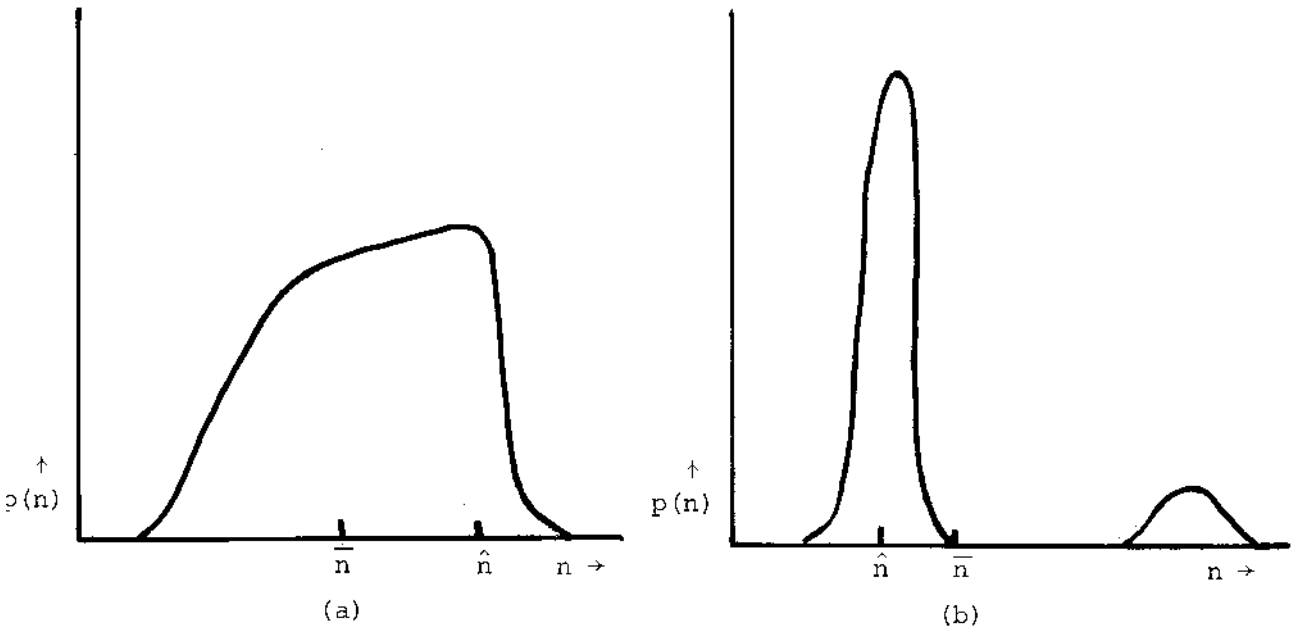

Figure 9.1. (a) A likelihood function for which the maximum-likelihood estimate is not a reasonable one. (b) A case where the maximum-likelihood estimate is more reasonable than the mean value estimate.

is still best to interpret the rules as an application of Bayes' theorem, for the following reason. Statisticians are well aware that the maximum-likelihood estimate may be very poor in the small-sample case. But these are just the cases in which situations like that depicted in Figure (9.1a) arise. In the small sample case, the Bayesian mean-value estimate (i.e., the "center of gravity" of the likelihood function) is often far more reasonable than the maximum-likelihood estimate.

It seems to me that we have to conclude from this that there is no sound reason for ever introducing the notion of maximum likelihood as a separate principle. It is already contained in Bayes' theorem as a special case, and whenever it is the appropriate method to use, Bayes' theorem will reduce to maximum likelihood. From this point of view, we will see later (when we take up decision theory) that it is possible to define precisely the conditions in which maximum likelihood is the optimal procedure [see Sec. 13.5.].

9.4. The Trouble with "Random" Variables.

Now let's take a glimpse at some of the difficulties that face the orthodox statistician because of his belief that use of probability theory requires us to distinguish between random and nonrandom quantities. In the example just studied, he didn't face any serious impediment because in this problem there was really no difficulty in deciding intuitively that s is a "constant", while n and c are "random variables". There is little danger that anyone would make a different decision. But there are other problems of inference in which it is not at all clear how this distinction is to be drawn. We will study some cases of this in detail when we take up linear regression (which means simply: fitting the best straight line to a plot of experimental points). This is probably the most common of all statistical problems faced by the experimental scientist; yet it is in just this problem that the distinction

between random and nonrandom quantities is so obscure that you sometimes have to resort to black magic to draw any distinction at all.

This situation has led to some really hilarious proposals for data reduction, solemnly advocated in the orthodox literature. Here is one way it can happen: the abscissa of our graph represents some physical quantity that has a true value X; but this is unknown because the value x actually read from a meter suffers from some experimental error $\varepsilon = x - X$. Nobody ever doubts that ε is random; but then which of the quantities x, X is random?

To change from one value of X to another, the experimenter typically turns a knob on his apparatus. According to some orthodox writers (Berkson, 1950; Mandel, 1964; Chap. 12), if he turns it without particularly noticing just where the "x-meter" ends up, then X is an unknown constant, and x a random variable. Orthodox theory then tells us how to analyze the data.

But another experimenter, even though he turns the knob in exactly the same way and stops at exactly the same place, does so with the conscious intention of stopping when the meter reads the value x. In this case, we are told, x is the "constant," and X the "random variable". Although there is absolutely no difference in the physical conditions of the experiment, orthodox teaching then tells us that we should analyze the data in an entirely different way, which can lead to different estimates of the slope and intercept of that line, and to widely different conclusions about the reliability of those estimates. If this isn't black magic, I would like to know what it is.

If, now, the second experimenter flips a coin to help him decide at what value of x to set the knob, then both x and X become random variables; and orthodox theory says we should use still a third method of data analysis, leading to a third set of conclusions!

I think most of us are persuaded that the import of the experiment ought to depend on this: how were the knobs actually turned, and what data resulted?

It does not depend on what thoughts flitted through anybody's imagination while turning the knobs; a given experimental procedure and resulting data have exactly the same import whether the knobs were turned by a man or a chimpanzee.

Orthodox theory fails to meet this rather elementary desideratum; if you give an orthodox statistician only the actual procedure and the actual data, plus one of the usual hypotheses about the errors, he has no definite way of getting started on the problem, because for him it is taboo to write down any probability distribution $p(x)$ unless it has been established that x is random; and this information gives him no basis for deciding which quantities are random. Although common sense tells you it cannot be relevant, he wants to know also something about the "state of mind" of the experimenter; and his final conclusions will depend on this. The fact that orthodox practice has to invoke psychokinesis in order to set up some problems hardly supports the claim (Bross, 1963) that orthodox methods, unlike Bayesian, are "objective" and "fact-oriented."

The Bayesian analysis does conform to our desideratum, because it is liberated from that taboo, and therefore has no need to draw artificial distinctions which have nothing to do with the physical conditions of the experiment. Given the above information, our robot can proceed immediately with definite calculations; he is not afraid to introduce probability distributions for any quantity about which he needs inference, and the question whether it is or is not "random" just never comes up at all. Because of his liberation from a taboo that has no justification and serves no purpose, probability theory is, for our robot, an enormously more powerful mathematical reasoning device than it is for one whose ideology forbids the use of that mathematics in its full generality. We will see some spectacular examples of this later when we compare Bayesian and orthodox significance tests and inter-

val estimation methods.

But orthodox taboos can lead to even worse consequences. They force one to attach such supreme importance to this random-nonrandom distinction that, in addition to introducing irrelevancies, many writers will not hesitate to throw away practically all the relevant data of a problem, in order to achieve the situation of "independent random errors" which their theory presupposes. For example, in the problem of fitting a straight line to experimental points, if there is cumulative error (i.e., the error in one value x_i is propagated into all subsequent x_j, $j > i$) Mandel (1964; Chap. 12) advocates that we estimate the slope of the line using only the first and last points; and simply throw away all the intermediate ones! To our robot--and also to the poor experimenter who labored to get the data--this is a far graver offense against reason than merely dabbling in a little black magic. As we will see later, throwing away the highly relevant evidence of the intermediate points can increase the probable error of your estimate by more than an order of magnitude in real problems.

The Bayesian analysis never requires us to do such absurd things, because it contains no artificial presuppositions about "randomness". If there is cumulative error, that is just an additional mathematical detail that Bayes' theorem takes into account without any difficulty, while retaining all of the relevant evidence.

Yet in spite of all this emphasis on the necessity of specifying the "random" quantities, no worker in probability theory, orthodox or otherwise, has produced any definition of "random variable" which could actually be applied in real life situations. Here is, for example, a quotation from the book of Savage (1954; p. 45): "The concept of a random variable enters into almost any discussion of probability. Experts are fairly well agreed on the following definition. A random variable is a function x attaching a value

x(s) in some set X to every s in a set S on which a probability measure P is defined." Definitions essentially equivalent to this can be found in most of the modern books on statistics. While this may be fine for setting up an abstract mathematical theory, the most obvious thing about it is that the definition is absolutely useless in helping us decide whether some specific quantity, such as the number of beans in a can, is or is not "random".

If you read the literature carefully, I think you will see that whenever the orthodox statistician gets down to a very specific problem, he uses the word "random" merely as shorthand for "likely to be different in different situations." In Laplace's theory there is no need to emphasize, or even to define, any sharp distinction between random and nonrandom quantities, for the commonsense reason that in the specific problem at hand, the quantity I am reasoning about (in the problem just discussed, n_1) is <u>always</u> simply a definite, but unknown number. Whether this number would or would not be the same in some other situation that I am <u>not</u> reasoning about, is just not relevant to my problem; to adopt different methods on such grounds is to commit the most obvious inconsistency of reasoning.

All right, I hope this little excursion into polemics has given you a clearer understanding of why, in the theory we are developing, the word "random" just doesn't appear; and of the kind of troubles we would get into if we did try to use it. In the next lecture, I want to return to the constructive development of the theory.

Lecture 10

DISCRETE PRIOR PROBABILITIES--THE ENTROPY PRINCIPLE

I would like to return to the job of designing this robot. We've got part of his brain designed, and we have seen how he would reason in a few simple problems of hypothesis testing and estimation. But he is still not a very versatile reasoning machine, because he has only one means by which he can translate raw information into numerical values of probability; the "principle of indifference," Rule 4. Consistency requires him to recognize the relevance of prior information, and so in almost every problem he is faced at the outset with the problem of assigning prior probabilities. He can use Rule 4 for this if he can break the situation up into mutually exclusive, exhaustive possibilities in such a way that no one of them is preferred to any other by the evidence he has. But often he will have prior information that does give him some reason for preferring one possibility to another. What do we do in this case?

10.1 A New Kind of Prior Information.

Let's imagine a certain class of problems in which the robot's prior information consists of average values of certain things. Suppose, for example, we tell him that statistics were collected in a recent earthquake and that out of 100 windows broken, there were 1,000 pieces found. We will state this in the form: "the average window is broken into 10 pieces." That is the way it would be reported. Given only that information, what is the probability

that a window would be broken into exactly m pieces? There is nothing in the theory so far that will answer that question. Let's imagine some other problems where the same situation would arise. Here's a fairly elaborate one.

Suppose I have a table which I cover with a black cloth, and I have some dice, which I am going to toss onto this table, but for reasons that will be clear in a minute, let's make these dice black with white spots. I toss a die onto the black table. Above I have a camera. Every time I toss it, I take a snapshot. The camera will record only the white spots. Now I don't change the film in between, so we end up with a multiple exposure; uniform blackening of the film after we have done this a few thousand times. From the density of the film, we can infer the average number of spots which were on top, but not the frequencies with which various faces came up. Suppose that the average number of spots on top turned out to be 4 1/2 instead of the 3 1/2 that we might expect from an honest die. What probability should our robot assign to the n'th face coming up?

To give still another example of a problem where the information available consists of average values, suppose that we have a string of 1,000 cars, bumper to bumper, and they occupy the full length of say three miles. We know the total length of this string of cars, and as they drive onto a rather large ferry boat, the distance that it sinks into the water tells us their total weight. So we know the average length and the average weight of the 1,000 cars. We can look up statistics from the manufacturers, and find out how long the Volkswagen is, how heavy it is; how long a Cadillac is, and how heavy it is, and so on, for all the other brands. From knowledge only of the average length and the average weight of these cars, what can we then infer about the number of cars of each make that were in the cluster? That is a problem where we have two average values given to us.

Now, it is not at all obvious how our robot should handle problems of this sort. So let's think about how we would want him to behave in this situation. We would not want him to jump to conclusions which are not warranted by the evidence he has. He should always frankly admit the full extent of his ignorance. We have seen that a uniform probability assignment represents a state of mind completely noncommittal with regard to all possibilities; it favors no one over any other, and thus leaves the entire decision to subsequent information which the robot may receive. The knowledge of average values does give the robot a reason for preferring some possibilities to others, but we would like him to assign a probability distribution which is, in some sense, as uniform as it can get while agreeing with the available information. The most conservative, noncommittal distribution is the one which is as "spread-out" as possible. In particular, the robot must not ignore any possibility--he must not assign zero probability to any situation unless his information really rules out that situation.

So, the aim of avoiding unwarranted conclusions leads us to ask whether there is some reasonable numerical measure of how uniform a probability distribution is, which the robot could maximize subject to constraints which represent his available information. Let's approach this in the way all problems are solved; the time-honored method of trial and error. We just have to invent some measures of uncertainty, and put them to the test to see what they give us.

One measure of how broad this distribution is would be its variance. Would it make sense if we build into the robot the property that whenever he is given information about average values, he will assign probabilities in such a way that the variance is maximized subject to that information? Well, consider the distribution of maximum variance for a given $\overline{m}$ if the values of m are unlimited, as in the broken window problem. Then the maximum variance

solution would be just the one where we assign a very large probability for no breakage at all, and an enormously small probability for a window to be broken into billions and billions of pieces. You can get an arbitrarily high variance this way, while keeping the average at 10. In the dice problem, the solution with maximum variance would be to assign all the probability to the one and the six, in such a way that you come out with the right average. So that, evidently, is not the way we would want our robot to behave; if he used the principle of maximum variance, he would be assigning zero probability to many cases which were not at all impossible on the information we gave him.

10.2. Minimum $\sum p_i^2$.

Another kind of measure of how spread out a probability distribution is, which has been used a great deal in statistics, is the sum of the squares of the probabilities assigned to each of the possibilities. The distribution which minimizes this expression, subject to constraints represented by average values, might be a reasonable way for our robot to behave. Let's see what sort of a solution this would lead to. I want to make

$$\sum_m p_m^2$$

a minimum, subject to the constraints that the sum of all p_m shall be unity, and the average over the distribution is $\overline{m}$. A formal solution is obtained by writing

$$\delta\left[\sum_m p_m^2 - \lambda \sum_m m p_m - \mu \sum_m p_m\right]$$

$$= \sum_m (2p_m - \lambda m - \mu)\, \delta p_m = 0 \qquad (10\text{-}1)$$

where λ and μ are Lagrange multipliers. So p_m will always be a linear function of m:

$$2p_m - \lambda m - \mu = 0.$$

Now, μ and λ are found from

$$\sum_m p_m = 1 \quad , \qquad \sum_m m p_m = \bar{m} \quad , \tag{10-2}$$

where $\bar{m}$ is the average value of m.

Let's investigate this and draw the graph for a simple version. Let's say that m can take on only the values 1, 2, and 3. Then we easily find that the formal solution for minimum $\sum_m p_m^2$ is

$$\begin{aligned} p_1 &= \frac{4}{3} - \frac{\bar{m}}{2} \\ p_2 &= \frac{1}{3} \\ p_3 &= \frac{\bar{m}}{2} - \frac{2}{3} \end{aligned} \tag{10-3}$$

In Figure (10.1) these results are plotted. This shows that p_1 and p_3 become negative. In these regions let's say we will replace the negative values by zero and then adjust the other probabilities to agree with the given value of $\bar{m}$. If we do this the results are shown in Figure (10.2).

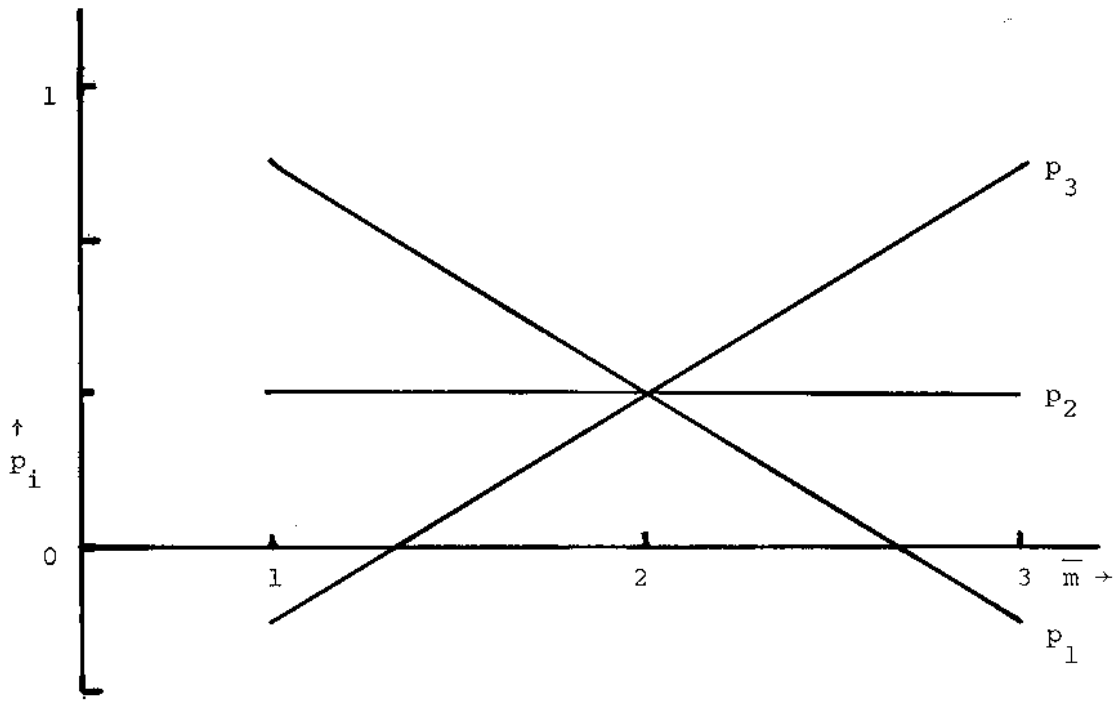

Figure 10.1. Formal solution for minimum $\sum p^2$.

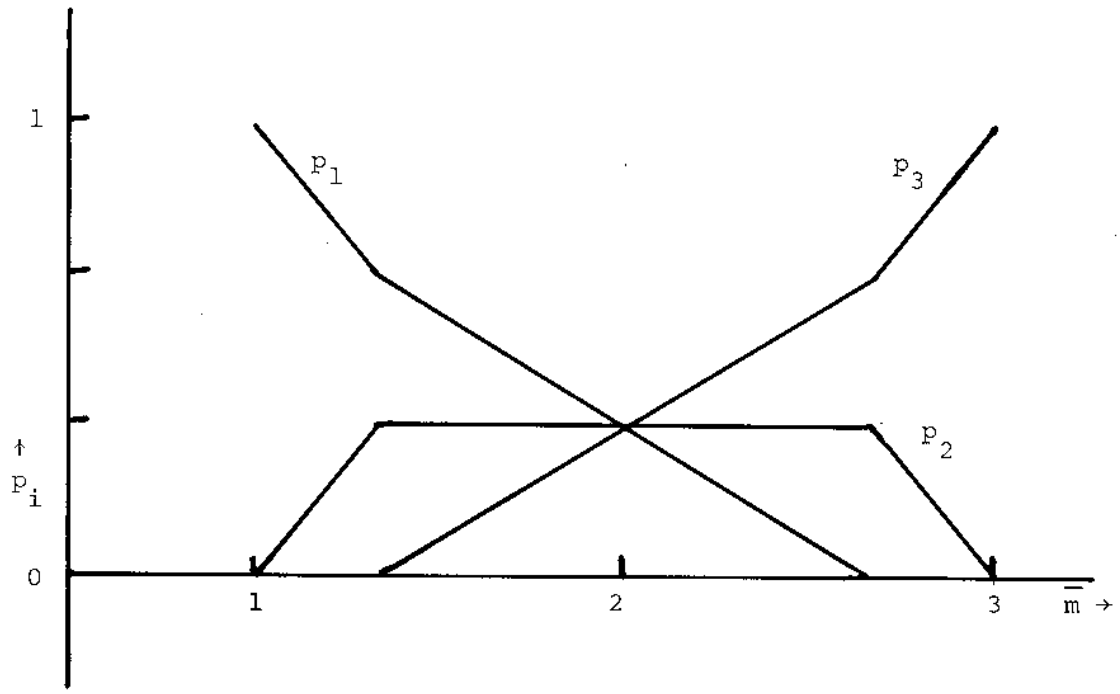

Figure 10.2. Corrected formal solution.

All right, so that's what this criterion will give to us. Now, is the robot behaving in a reasonable way if we build this behavior pattern into him? This is certainly a big improvement over maximum variance, but he is still, in certain ranges of $\overline{m}$, assigning zero probability to one of the possibilities, and there is nothing in the data we gave him which said one was impossible. So he is still jumping to unjustified conclusions. But the idea behind it still looks like a good one. There should be some consistent measure of the uniformity, or "amount of uncertainty" of a probability distribution which we can maximize, subject to constraints, and which will have the property that it forces the robot to be completely honest about what he knows, and in particular it does not permit the robot to draw any conclusions unless those conclusions are really justified by the evidence he has.

10.3. Entropy: Shannon's Theorem.

Well, at this stage we turn to the most quoted theorem in Shannon's work on information theory (Shannon, 1948; Shannon and Weaver, 1949). This is the theorem. If there exists a consistent measure of the "amount of uncertainty" represented by a probability distribution, there are certain conditions it will have to satisfy. I am going to state them in a way which will remind you of the arguments we gave in Lecture 3; in fact, this is really a continuation of the basic development of probability theory. Here is the line of reasoning:

(1) We assume that some numerical measure $H_n(p_1, p_2, \cdots, p_n)$ exists; i.e., that it is possible to set up some kind of association between "amount of uncertainty" and real numbers.

(2) We assume a continuity property: H_n is a continuous function of the p_i. For otherwise an arbitrarily small change in the probability distribution would still lead to the same big change in the amount of uncertainty.

(3) We require that this measure should correspond qualitatively to common sense in that when there are many possibilities, we are more uncertain than when there are few. This condition takes the form that in case the p_i are all equal, the quantity

$$h(n) = H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

shall be a monotonic increasing function of n. This merely establishes the "sense of direction."

(4) We require that the measure H_n be consistent in the same sense as before; i.e., if there is more than one way of working out its value, we've got to get the same answer for every possible way.

Previously, our conditions of consistency took the form of the functional equations (3-2), (3-7), (3-11). Now we have instead a hierarchy of functional equations relating the different H_n to each other. Suppose the robot perceives two alternatives, to which he assigns probabilities p_1 and $q = 1 - p_1$. Then the "amount of uncertainty" represented by this distribution is $H_2(p_1, q)$. But now the robot learns that the second alternative really consists of two possibilities, and he assigns probabilities p_2, p_3 to them, satisfying $p_2 + p_3 = q$. What is now his full uncertainty $H_3(p_1, p_2, p_3)$ as to all three possibilities? Well, the process of choosing one of the three can be broken down into two steps. First, he decides whether the first possibility is or is not true; his uncertainty for this decision is the original $H_2(p_1, q)$. Then, with probability q, he encounters an additional uncertainty as to events 2, 3, leading to

$$H_3(p_1,p_2,p_3) = H_2(p_1,q) + qH_2\left(\frac{p_2}{q},\frac{p_3}{q}\right) \tag{10-4}$$

In general, a function H_n can be broken down in many different ways, relating it to the lower order functions by a large number of equations like this.

Note that equation (10-4) says rather more than our previous functional equations did. It says not only that the H_n are consistent in the aforementioned sense, but also that they are to be additive. So this is really an additional assumption which we should have included in our list. The most general equation of consistency would be a functional equation which is satisfied by any monotonic increasing function of H_n, but I don't know how to write it.

At any rate, the next step is perfectly straightforward mathematics; let's see the full proof of Shannon's theorem, now dropping the unnecessary subscript on H_n.

First, let's find the most general form of the composition law (10-4)

for the case that there are n mutually exclusive propositions $(A_1,\ldots,A_n)$ to consider, to which we assign probabilities $(p_1,\ldots,p_n)$ respectively. Instead of giving the probabilities of the $(A_1,\ldots,A_n)$ directly, we might first group the first k of them together as the proposition denoted by $(A_1+A_2+\ldots+A_k)$ in Boolean algebra, and give its probability which by Eq. (3-21) is equal to $w_1 = (p_1+\ldots+p_k)$; then the next m propositions are combined into $(A_{k+1}+\ldots+A_{k+m})$, for which we give the probability $w_2 = (p_{k+1}+\ldots+p_{k+m})$, etc. When this much has been specified, the amount of uncertainty as to the composite propositions is $H(w_1\ldots w_r)$. Next we give the conditional probabilities $(p_1/w_1,\ldots,p_k/w_1)$ of the propositions $(A_1,\ldots,A_k)$, given that the composite proposition $(A_1+\ldots+A_k)$ is true. The additional uncertainty $H(p_1/w_1,\ldots p_k/w_1)$ is then encountered with probability w_1. Carrying this out for the other composite propositions $(A_{k+1}+\ldots+A_{k+m})$, etc., we arrive ultimately at the same state of knowledge as if the $(p_1,\ldots,p_n)$ had been given directly; so if our measure of "amount of uncertainty" is to be consistent, we must obtain the same ultimate uncertainty no matter how the choices were broken down in this way. Thus we must have

$$H(p_1\ldots p_n) = H(w_1\ldots w_r) + w_1 H(p_1/w_1,\ldots,p_k/w_1) + w_2 H(p_{k+1}/w_2,\ldots,p_{k+m}/w_2) + \ldots \qquad (10\text{-}5)$$

which is the general form of the functional equation (10-4). For example, $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + (1/2)\, H(2/3, 1/3)$.

Since $H(p_1\ldots p_n)$ is to be continuous, it will be sufficient to determine it for all rational values

$$p_i = \frac{n_i}{\Sigma n_i} \qquad (10\text{-}6)$$

with n_i integers. But then (10-5) determines the function H already in terms of the quantities $h(n) \equiv H(1/n,\ldots,1/n)$ which measure "amount of uncertainty" in the case of n equally likely alternatives. For we can regard a choice of

one of the alternatives $(A_1,\ldots,A_n)$ as the first step in the choice of one of

$$\sum_{i=1}^{n} n_i$$

equally likely alternatives in the manner just described, the second step of which is also a choice between n_i equally likely alternatives. As an example, with n=3, we might choose $n_1 = 3$, $n_2 = 4$, $n_3 = 2$. For this case the composition law (10-5) becomes

$$h(9) = H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) + \frac{3}{9}\,h(3) + \frac{4}{9}\,h(4) + \frac{2}{9}\,h(2)$$

For a general choice of the n_i, (10-5) reduces to

$$h(\Sigma n_i) = H(p_1\ldots p_n) + \sum_i p_i h(n_i) \qquad (10\text{-}7)$$

Now we can choose all $n_i = m$; whereupon (10-7) collapses to

$$h(mn) = h(m) + h(n) \qquad (10\text{-}8)$$

Evidently, this is solved by setting

$$h(n) = k \log n \qquad (10\text{-}9)$$

where k is a constant. But is this solution unique? If m, n were continuous variables, this would be easy to answer; differentiate with respect to m, set m = 1, and integrate the resulting differential equation with the initial condition h(1) = 0 evident from (10-8), and you have proved that (10-9) is the only solution. But in our case, (10-8) need hold only for integer values of m, n; and this elevates the problem from a trivial one of analysis to an interesting little exercise in number theory.

First, note that (10-9) is no longer unique; in fact, (10-8) has an infinite number of solutions for integer m, n. For, each positive integer N has a unique decomposition into prime factors; and so by repeated application of (10-8) we can express h(N) in the form $\sum_i m_i h(q_i)$ where q_i are the prime numbers and m_i non-negative integers. Thus we can specify $h(q_i)$ <u>arbitrarily</u> for the prime numbers q_i, whereupon (10-8) is just sufficient to determine

h(N) for all positive integers.

To get any unique solution for h(n), we have to add our qualitative requirement that h(n) be monotonic increasing in n. To show this, note first that (10-8) may be extended by induction:

$$h(nmr\cdots) = h(n) + h(m) + h(r) + \cdots$$

and setting the factors equal in the k'th order extension gives

$$h(n^k) = k\, h(n) \tag{10-10}$$

Now let t, s be any two integers not less than 2. Then for arbitrarily large n, we can find an integer m such that

$$\frac{m}{n} \leq \frac{\log t}{\log s} < \frac{m+1}{n} \tag{10-11}$$

or,

$$s^m \leq t^n < s^{m+1}$$

Since h is monotonic increasing,

$$h(s^m) \leq h(t^n) \leq h(s^{m+1})$$

or from (10-10),

$$m\, h(s) \leq n\, h(t) \leq (m+1)\, h(s)$$

which can be written as

$$\frac{m}{n} \leq \frac{h(t)}{h(s)} \leq \frac{m+1}{n} \tag{10-12}$$

Comparing (10-11), (10-12), we see that

$$\left|\frac{h(t)}{h(s)} - \frac{\log t}{\log s}\right| \leq \frac{1}{n}$$

or

$$\left|\frac{h(t)}{\log t} - \frac{h(s)}{\log s}\right| \leq \varepsilon \tag{10-13}$$

where

$$\varepsilon \equiv \frac{h(s)}{n \log t}$$

is arbitrarily small. Thus h(t)/log t must be a constant, and the uniqueness

of (10-9) is proved.

Now different choices of k amount to the same thing as taking logarithms to different bases; so if we leave the base arbitrary for the moment, we can just as well write $h(n) = \log n$. Substituting this into (10-7), we have Shannon's theorem: the only function $H(p_1,\ldots,p_n)$ satisfying the conditions we have imposed on a reasonable measure of "amount of uncertainty" is

$$H(p_1,\ldots,p_n) = -\sum_{i=1}^{n} p_i \log p_i \qquad (10\text{-}14)$$

Accepting this interpretation, it follows that the distribution $(p_1\ldots p_n)$ which maximizes (10-14) subject to constraints imposed by the available information, will represent the "most honest" description of what we know about the propositions $(A_1,\ldots,A_n)$. The only arbitrariness is that we have the option of taking the logarithm to any base we please, corresponding to a multiplicative constant in H. This, of course, has no effect on the values of $(p_1,\ldots,p_n)$ which maximize H.

The function H will be called the <u>entropy</u> of the distribution $(p_1,\ldots,p_n)$ from now on. It is a new measure of how uniform a probability distribution is--any change in the direction of equalizing the different probabilities will increase the entropy.

I don't think that either this demonstration or the ones we gave in the third lecture are anywhere near in satisfactory form yet. In particular, the functional equation (10-4) does not seem quite so intuitively compelling as our previous ones were. You might ask why the factor q must appear in the last term, and the only answer I can give is that if you leave it out, the solution of the functional equation will collapse to $H_n(p_1\ldots p_n) = (n-1)$, independently of the p_i, and you will lose everything we had hoped to get from this argument. In this case, I think the trouble is just that neither I nor any other writer known to me has yet learned how to verbalize the argument leading to (10-4) in a fully convincing manner. Perhaps this will inspire

you to try your hand at improving the verbiage that I used just before writing (10-4).

For this reason, it is comforting to know that there are several other possible arguments which will also lead to the same conclusion (10-14). Khinchin (1957) has given a slightly different set of conditions. They are:

(1) For given n, $H_n(p_1 \ldots p_n)$ attains its maximum value when $p_k = (1/n)$, $k = 1, 2, \ldots, n$.

(2) If we include in our enumeration a new situation which is, however, known to be impossible, our state of uncertainty is not really changed. Therefore, we should have $H_{n+1}(p_1 \ldots p_n, 0) = H_n(p_1 \ldots p_n)$.

(3) A composition law essentially equivalent to (10-4) although stated in slightly different terms.

Khinchin shows that the only continuous function satisfying these requirements is the entropy expression (10-14).

10.4. The Wallis Derivation.

Another, and quite amusing, way of deriving the maximum-entropy principle resulted from a suggestion made to me by Dr. Graham Wallis (although the argument to follow differs slightly from his). We are given information I, which is to be used in assigning probabilities $\{p_1 \ldots p_m\}$ to m different possibilities. We have a total amount of probability

$$\sum_{i=1}^{m} p_i = 1$$

to allocate among them. Now in judging the reasonableness of any particular allocation we are limited to a consideration of I and the laws of probability theory; for to call upon any other evidence would be to admit that we had not used all the available information in the first place.

The problem can also be stated as follows. Choose some integer $n >> m$, and imagine that we have n little "quanta" of probability, each of magnitude

$\delta = n^{-1}$, to distribute in any way we see fit. In order to ensure that we have a "fair" allocation, in the sense that none of the m possibilities shall knowingly be given either more or fewer of these quanta than it "deserves," in the light of the information I, we might proceed as follows.

Suppose we were to scatter these quanta at random among the m choices--you can make this a blindfolded penny-pitching game into m equal boxes if you like. If we simply toss these "quanta" of probability at random, so that each box has an equal probability of getting them, nobody can claim that any box is being unfairly favored over any other. If we do this, and the first box receives exactly n_1 quanta, the second n_2, etc., we will say that the random experiment has generated the probability assignment

$$p_i = n_i \delta = n_i/n, \quad i = 1, 2, \ldots, m$$

The probability that this will happen is

$$m^{-n} \frac{n!}{n_1! \ldots n_m!}$$

Now imagine that a blindfolded friend repeatedly scatters the n quanta at random among the m possibilities. Each time he does this we examine the resulting probability assignment. If it happens to conform to the information I, we accept it; otherwise we reject it and tell him to try again. We continue until some probability assignment $\{p_1 \ldots p_m\}$ is accepted.

What is the most likely probability distribution to result from this game? It is the one which maximizes

$$W \equiv \frac{n!}{n_1! \ldots n_m!} \tag{10-15}$$

subject to whatever constraints are imposed by the information I. We can refine this procedure by choosing smaller quanta; i.e. large n. In the limit we have, by the Stirling approximation

$$\log n! = n \log n - n + \sqrt{2\pi n} + \frac{1}{12n} + O\left(\frac{1}{n^2}\right) \tag{10-16}$$

where $O(1/n^2)$ denotes terms that tend to zero as $n\to\infty$, as $(1/n^2)$ or faster. Using this result, and writing $n_i = np_i$, we easily find that as $n\to\infty$, $n_i\to\infty$, in such a way that $n_i/n \to p_i = \text{const.}$,

$$\frac{1}{n}\log W = \frac{1}{n}\left[\log n! - \sum_{i=1}^{n}\log (np_i)!\right]$$

$$\to \log n - 1 - \frac{1}{n}\sum_{i=1}^{n}\left[np_i \log (np_i) - np_i\right]$$

Since $\Sigma p_i = 1$, several terms cancel, and we are left with

$$\frac{1}{n}\log W \to -\sum_{i=1}^{n} p_i \log p_i = H(p_1 \ldots p_n) \tag{10-17}$$

and so, the most likely probability assignment to result from this game, is just the one that has maximum entropy subject to the given information I.

You might object that this game is still not entirely "fair," because we have stopped at the first acceptable result without seeing what other acceptable ones might also have turned up. In order to remove this objection, we can consider all possible acceptable distributions and choose the average $\overline{p_i}$ of them. But here the "laws of large numbers" come to our rescue. I leave it for you to prove that in the limit of large n, the overwhelming majority of all acceptable probability allocations that can be produced in this game are arbitrarily close to the maximum-entropy distribution.

This derivation is, in several respects, the best one yet produced. It is entirely independent of Shannon's functional equation (10-5); it does not require any postulates about connections between probability and frequency; nor does it suppose that the different possibilities {1 ... m} are themselves the result of any repeatable random experiment. Furthermore, it leads automatically to the prescription that H is to be maximized--and not treated in some other way--without the need for any quasi-philosophical interpretation of H in terms of such a vague notion as "amount of uncertainty." Let me stress this point. It is a big mistake to try to read too much philosophical signifi-

cance into theorems which lead to equation (10-14). In particular, the association of the word "information" with entropy expressions seems in retrospect quite unfortunate, because it persists in carrying the wrong connotations to so many people. Shannon himself, with really prophetic insight into the reception his work would get, tried to play it down by pointing out immediately after stating his theorem, that it was in no way necessary for the theory to follow. By this he meant that the inequalities which H satisfies are already quite sufficient to justify its use; it does not really need the further support of the theorem which deduces it from functional equations expressing intuitively the properties of "amount of uncertainty." However, while granting that this is perfectly true, I would like now to try to show that if we do accept the expression for entropy, very literally, as the correct expression for the "amount of uncertainty" represented by a probability distribution, this will lead us to a much more unified picture of probability theory in general. It will enable us to see that the principle of indifference, Rule 4, and many frequency connections of probability are special cases of a single principle, and that statistical mechanics and communication theory are both instances of a single method of reasoning.

10.5. An Example.

First, let's test this principle and see how it would work out if we ask the robot to assign probabilities in such a way that the entropy (10-14) is maximized subject to the available information, in the simple example discussed in Sec. 10.2, in which m can take on only the values 1, 2, 3 and $\overline{m}$ is given.

We can use our Lagrange multiplier argument again to solve this problem; i.e., as in (10-1),

$$\delta\left[H(p_1 \ldots p_3) - \lambda \sum_{m=1}^{3} m p_m - \mu \sum_{m=1}^{3} p_m \right] =$$

$$= \sum_{m=1}^{3} \left[\frac{\partial H}{\partial p_m} - \lambda m - \mu \right] \delta p_m = 0.$$

Now,

$$\frac{\partial H}{\partial p_m} = - \log p_m - 1 \tag{10-18}$$

so our solution is

$$p_m = e^{-\lambda_o - \lambda m} \tag{10-19}$$

where $\lambda_o \equiv \mu + 1$.

So the distribution which has maximum entropy, subject to a given average value, will always be in exponential form, and we have to fit the constants λ_o and λ by forcing this to agree with the constraints that the sum of the p's must be one and that the average value must be equal to the average $\overline{m}$ that we assigned. Well, the mathematics that you have to go through in order to do this is very straightforward and comes out very beautifully if you define a function

$$Z(\lambda) \equiv \sum_{m=1}^{3} e^{-\lambda m} \tag{10-20}$$

which we call the <u>partition function</u>. The equations (10-2) which fix our Lagrange multipliers then take the form

$$\lambda_o = \log Z(\lambda) \tag{10-21}$$

and

$$\overline{m} = - \frac{\partial}{\partial \lambda} \log Z(\lambda) \tag{10-22}$$

We find easily that $p_1(\overline{m})$, $p_2(\overline{m})$, $p_3(\overline{m})$ are given in parametric form by

$$p_k = \frac{\exp[(2-k)\lambda]}{1 + 2 \cosh \lambda}, \quad k = 1, 2, 3. \tag{10-23}$$

$$\overline{m} = \frac{e^{2\lambda} + 2e^{\lambda} + 3}{e^{2\lambda} + e^{\lambda} + 1}. \tag{10-24}$$

In a more complicated problem we would just have to leave it in parametric form, but in this particular case we can eliminate the parameter λ algebra-

ically, leading to the explicit solution

$$p_1 = \frac{3 - \overline{m} - p_2}{2}$$

$$p_2 = \frac{1}{3}\left[\sqrt{4 - 3(\overline{m}-2)^2} - 1\right] \qquad (10\text{-}25)$$

$$p_3 = \frac{\overline{m} - 1 - p_2}{2}$$

These results are plotted in Figure (10.3). p_2 is the arc of an ellipse which comes in with unit slope at the ends. p_1 and p_3 are also arcs of ellipses, but slanted one way and the other.

Let's just notice that we have finally arrived here at a solution which meets the objections we had to the first two criteria. The maximum entropy distribution automatically has the property $p_m \geq 0$ because the logarithm has a singularity at zero which we could never get past. It has, furthermore, the property that it never allows the robot to assign zero probability to any possibility unless the evidence forces that probability to be zero. The only

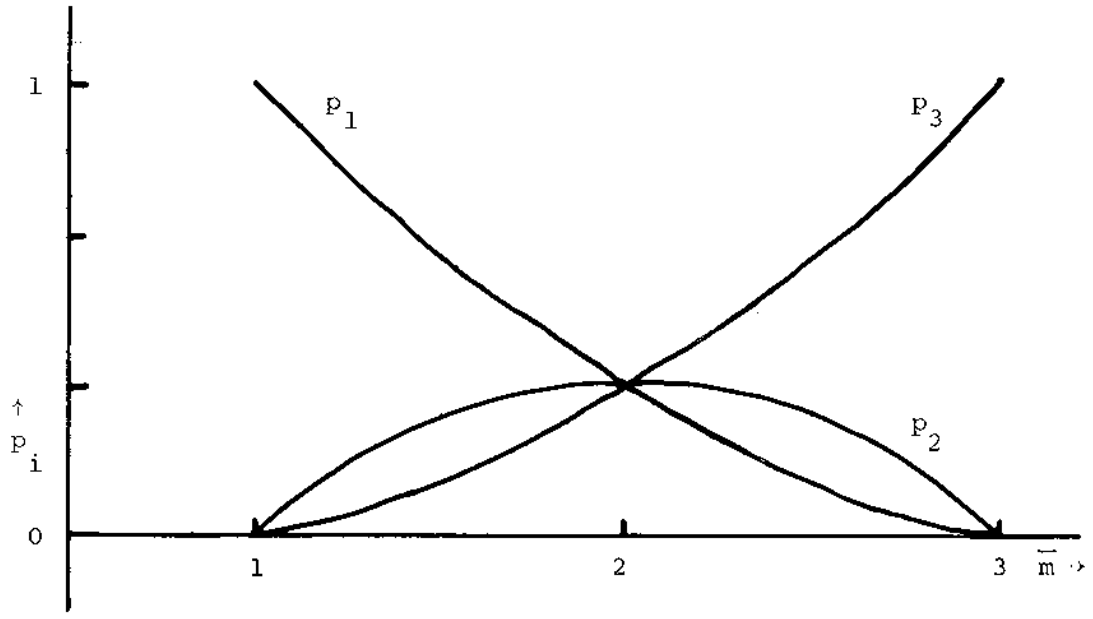

Figure 10.3. Maximum-Entropy solution.

place where a probability goes to zero is in the limit where the $\overline{m}$ is exactly one or exactly three. But of course, in that case, some probabilities did have to be zero.

We see the comparison between these two criteria is very interesting. The criterion that

$$\sum_m p_m^2 = \text{minimum}$$

gives [Fig. (10.2)] the same value and the same slope as the maximum entropy solution, at the end points and at the middle. It represents, in a sense, the best straight-line approximation you could have made to the maximum entropy solution.

10.6. Generalization: A More Rigorous Proof.

The maximum-entropy solution can be generalized in many ways. Suppose a variable x can take on n different discrete values $(x_1 \ldots x_n)$, which correspond to the n different propositions $(A_1 \ldots A_n)$ above; and that there are m different functions of x

$$f_k(x) \quad , \qquad 1 \leq k \leq m, \quad m < n, \tag{10-26}$$

for which we know the mean values. What probabilities $(p_1 \ldots p_n)$ will the robot assign to the possibilities $(x_1 \ldots x_n)$? The average of $f_k(x)$ is supposed known for each of the possible values of k, i.e.,

$$F_k \equiv \langle f_k(x) \rangle = \sum_{i=1}^{n} p_i \, f_k(x_i) \quad , \tag{10-27}$$

and the robot will find the set of p_i's which has maximum entropy subject to all these constraints simultaneously. Let's see what he'll come out with. We just have to introduce as many Lagrange multipliers as there are constraints imposed on the problem.

$$\delta [H(p_1 \ldots p_n) - (\lambda_o - 1) \sum_i p_i - \lambda_1 \sum_i p_i \, f_1(x_i) - \ldots - \lambda_m \sum_i p_i \, f_m(x_i)]$$

$$= \sum_i \left[\frac{\partial H}{\partial p_i} - (\lambda_o - 1) - \lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i) \right] \delta p_i = 0$$

and so from (10-18) our solution is the following:

$$p_i = e^{-\lambda_o - \lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)} . \qquad (10\text{-}28)$$

That's the form of the distribution, and we still have to find how he is going to evaluate these constants. In the first place, the sum of all probabilities will have to be unity, i.e.,

$$1 = \sum_i p_i = e^{-\lambda_o} \sum_i e^{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)} . \qquad (10\text{-}29)$$

If we now define a partition function as

$$Z(\lambda_1 \ldots \lambda_m) \equiv \sum_{i=1}^{n} e^{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)} \qquad (10\text{-}30)$$

then (10-29) reduces to

$$\lambda_o = \log Z(\lambda_1 \ldots \lambda_m) \qquad (10\text{-}31)$$

The average value (10-27) of $f_k(x)$ is then

$$F_k = e^{-\lambda_o} \sum_i e^{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)} f_k(x_i) ,$$

or,

$$F_k = - \frac{\partial}{\partial \lambda_k} \log Z \qquad (10\text{-}32)$$

What is the maximum value of the entropy that we get from this probability distribution? After an entropy has been maximized, I will call it S, the way physicists do, instead of H, the way information theory people do:

$$S \equiv (H)_{max} = \left(- \sum_{i=1}^{n} p_i \log p_i \right)_{max} \qquad (10\text{-}33)$$

From (10-28) we find that

$$S = \lambda_o + \lambda_1 F_1 + \ldots + \lambda_m F_m \qquad (10\text{-}34)$$

Now these results open up so many new applications that it is important to have as rigorous a proof as possible. But to solve a maximization problem

by variational means, as we just did, isn't 100 per cent rigorous. Our Lagrange multiplier argument has the nice feature that it gives you the answer instantaneously. It has the bad feature that after you've done it, you're not quite sure it <u>is</u> the answer. Suppose we had a function like the one in Fig (10.4), and our job was to locate the maximum of it. Well, if we state it as a variational problem and set the derivative equal to 0, we'll get solutions at A, B, C, etc. And, of course, we could investigate these separately and see which one is really a minimum, which one is a maximum. But after we prove that A is a local maximum, still we have doubt as to whether it's an absolute maximum. Maybe there is some other point that is still higher. Even after we've proved that we have the highest value that can be reached by variational methods, it is still possible that the function reaches a still higher value at some cusp E that we can't locate by variational methods. There would always be a little grain of doubt remaining if we do only the variational problem.

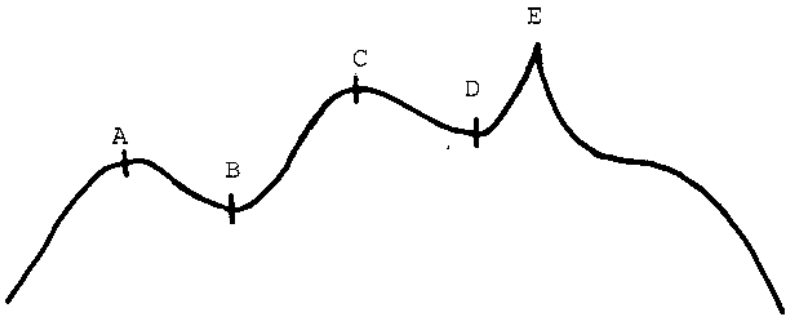

Figure 10.4.

So, I would like now to give you an entirely different derivation which is strong just where the variational argument is weak. For this I want a lemma. Let p_i be any set of numbers which could be a possible probability distribution; in other words, they add up to one and they are not negative,

$$\sum_{i=1}^{n} p_i = 1 \quad , \qquad p_i \geq 0 \tag{10-35}$$

and let u_i be another possible probability distribution,

$$\sum_{i=1}^{n} u_i = 1 \quad , \qquad u_i \geq 0. \tag{10-36}$$

Now let's think for a moment about the function log x. The graph of log x looks like this, Fig. 10.5.

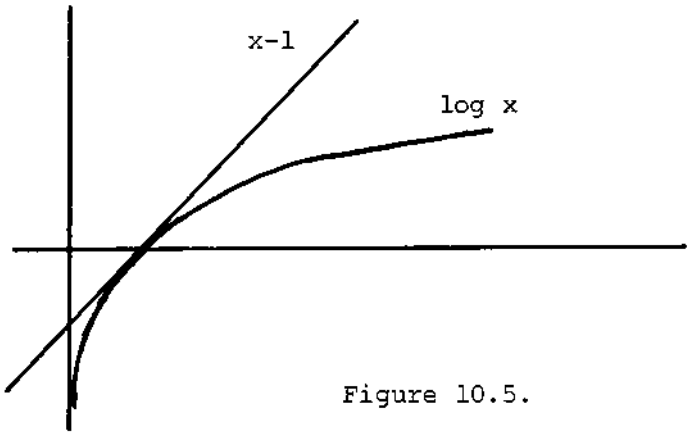

Figure 10.5.

It passes through the point (1,0) with unit slope. So if I draw a tangent to it at this point, the straight line has the equation y = x - 1. You see that log x always has curvature downward and so it stays below the tangent; therefore,

$$\log x \leq (x - 1) \quad , \qquad 0 < x < \infty \tag{10-37}$$

with equality if and only if x = 1. Therefore,

$$\sum_{i=1}^{n} p_i \log \left(\frac{u_i}{p_i}\right) \leq \sum_{i=1}^{n} p_i \left(\frac{u_i}{p_i} - 1\right) = 0$$

or,

$$H(p_1 \cdots p_n) \leq \sum_{i=1}^{n} p_i \log \left(\frac{1}{u_i}\right) \tag{10-38}$$

with equality if and only if $p_i = u_i$, i = 1, 2, ..., n. This is the lemma we need.

I'm going to simply pull a distribution u_i out of the hat;

$$u_i \equiv \frac{1}{Z(\lambda_1 \cdots \lambda_m)} \exp\{-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)\} \; . \tag{10-39}$$

where $Z(\lambda_1 \cdots \lambda_m)$ is defined by (10-30). Never mind why I chose u_i this

particular way; we'll see why in a minute. But now let's play with the inequality (10-38). We can now write it as

$$H \leq \sum_{i=1}^{n} p_i [\log Z + \lambda_1 f_1(x_i) + \ldots + \lambda_m f_m(x_i)]$$

or

$$H \leq \log Z + \lambda_1 \langle f_1(x) \rangle + \ldots + \lambda_m \langle f_m(x) \rangle \ . \qquad (10\text{-}40)$$

Now, let the p_i vary over the class of all possible probability distributions that satisfy the constraints (10-27) of the problem. The right-hand side of (10-40) stays constant. Our lemma now says that H attains its absolute maximum, making (10-40) an equality, if and only if the p_i are chosen as the canonical distribution (10-39).

This is the rigorous proof, which is independent of the things that might happen if you try to do it as a variational problem. This argument is, as we see, strong just where the variational argument is weak. On the other hand, this argument is weak where the variational argument is strong, because I just had to pull the answer out of a hat in writing (10-39). I had to know the answer before I could prove it. If you have both arguments side by side, then you have the whole story.

10.7. Formal Properties of Maximum-Entropy Distributions.

Now I want to put down a list of the general formal properties of this canonical distribution (10-39). This is a bad way of doing it in one sense; it sounds very abstract and you don't see the connection to any physical problem yet. On the other hand, we get all the things we want a lot faster if we first become aware of all the formal properties that are going to be in this theory in any application; and then later I'll go into specific physical problems and we'll see that every one of these formal relations turns out to have many different useful physical meanings, depending on the particular problem.

Now the maximum attainable H that we can get by holding these averages fixed depends, of course, on the average values we specified,

$$(H)_{max} = S(F_1 \ldots F_m) = \log Z + \sum_{k=1}^{m} \lambda_k F_k \qquad (10\text{-}41)$$

H itself we can regard as a measure of the "amount of the uncertainty" in any probability distribution. After I have maximized it, it becomes a function of the definite physical data of the problem, and I'll call it S. It's still a measure of "uncertainty", but it's uncertainty when all the information we have consists of just these numbers. It is "subjective" in the sense that it still measures uncertainty; but it is completely "objective" in the sense that it depends only on the data of the problem, and not on anybody's personality.

If S is to be only a function of $(F_1 \ldots F_m)$, then in (10-41) the $(\lambda_1 \ldots \lambda_m)$ must also be thought of as functions of $(F_1 \ldots F_m)$. At first, the λ's were just unspecified constants flapping around loose, but eventually we have to find what they are. If I choose different λ_i, I am writing down different probability distributions (10-39); and we saw in (10-32) that the averages over this distribution agree with the given averages F_k if

$$F_k = \langle f_k \rangle = -\frac{\partial}{\partial \lambda_k} (\log Z) \quad , \quad k = 1, 2, \ldots, m \qquad (10\text{-}42)$$

So we are now to regard (10-42) as a set of m simultaneous equations which are to be solved for the λ_i in terms of the given data F_k; at least one would like to dream about this. Generally, when you get to non-trivial problems, this is so involved that you have to leave the λ_i where they are, and express things in parametric form. If you've got more than about two λ_i in the problem, it is generally impractical to solve for them explicitly. Actually, this isn't such a tragedy, because the λ_i usually turn out to have such important physical meanings that we are quite happy to use them as the independent variables. However, I would like to show you that if we can evaluate the function $S(F_1 \ldots F_m)$, then we can give the λ_i as explicit functions of the

given data.

Suppose I take S and differentiate it. I make a small change in one of the values F_k that we fed into the problem; how does this change the maximum attainable H? We have from (10-41),

$$\frac{\partial S}{\partial F_k} = \sum_{j=1}^{m} \frac{\partial \log Z}{\partial \lambda_j} \frac{\partial \lambda_j}{\partial F_k} + \sum_{j=1}^{m} \frac{\partial \lambda_j}{\partial F_k} F_k + \lambda_k$$

which, thanks to (10-42), collapses to

$$\lambda_k = \frac{\partial S}{\partial F_k} \tag{10-43}$$

in which λ_k is given explicitly.

Compare this equation with (10-42); one gives F_k explicitly in terms of the λ_k, the othergives the λ_k explicitly in terms of the F_k. If I specify log Z as a function of the λ_k; or if I specify S as a function of the given data F_k, these are equivalent in the sense that each gives full information about the probability distribution. The complete story is contained in either function, and in fact you see that (10-41) is just the Legendre transformation that takes us from one representative function to another.

We can derive some more interesting laws simply by differentiating the two we already have. Let me differentiate (10-42) with respect to λ_j:

$$\frac{\partial F_k}{\partial \lambda_j} = \frac{\partial^2}{\partial \lambda_j \partial \lambda_k} (\log Z) = \frac{\partial F_j}{\partial \lambda_k} \tag{10-44}$$

since the second cross derivatives of log Z are symmetric in j and k. So here's a general reciprocity law which will hold in any problem that we do by maximizing the entropy. Likewise, if I differentiate (10-43) a second time, I'll have

$$\frac{\partial \lambda_k}{\partial F_j} = \frac{\partial^2 S}{\partial F_j \partial F_k} = \frac{\partial \lambda_j}{\partial F_k} \tag{10-45}$$

another reciprocity law, which is however not independent of (10-44), because if we define the matrices $A_{jk} \equiv \partial \lambda_j / \partial F_k$, $B_{jk} \equiv \partial F_j / \partial \lambda_k$, you easily see that

they are inverse matrices: $A = B^{-1}$, $B = A^{-1}$. These reciprocity laws might appear trivial from the ease with which we derived them here; but when we get around to applications we'll see that they have highly nontrivial and non-obvious physical meanings.

Now let's consider the possibility that one of these functions $f_k(x)$ has an extra parameter α in it which can be varied. If you want to think of applications, you can say $f_k(x_i;\alpha)$ stands for the i'th energy level of some system and α represents the volume of the system. The energy levels depend on the volume. Or, if it's a magnetic resonance system, you can say this represents the energy of the i'th state of the spin system and α represents the magnetic field that's applied. Very often we want to make a prediction of how certain quantities change as I change α. I want to calculate the pressure; or the susceptibility. By the criterion of minimum mean square error, the best estimate I can make of that derivative would be the mean value over the probability distribution. If I write it out, it will be

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = \frac{1}{Z} \sum_i \exp\{-\lambda_1 f_1(x_i) - \ldots - \lambda_k f_k(x_i,\alpha) - \ldots - \lambda_m f_m(x_i)\} \frac{\partial f_k(x_i,\alpha)}{\partial \alpha}$$

which reduces to

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{1}{\lambda_k Z} \frac{\partial}{\partial \alpha} \sum_i \exp\{\quad\}$$

$$= -\frac{1}{\lambda_k} \frac{\partial}{\partial \alpha} \log Z \ . \qquad (10\text{-}46)$$

In this derivation, I supposed that this parameter α only shows up in one function f_k. If the same parameter shows up in several different f_k, then I'll leave it for you to verify that this generalizes to

$$\sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial}{\partial \alpha} \log Z \ . \qquad (10\text{-}47)$$

This general rule contains, among other things, the equation of state of any system.

When we add α to the problem, the maximum entropy S is a function not only of the specified average values $\langle f_k \rangle$, but it depends now on α too. Likewise, Z depends on α. If we differentiate log Z or S, we get the same thing:

$$-\frac{\partial S}{\partial \alpha} = \sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial}{\partial \alpha} \log Z \qquad (10\text{-}48)$$

with one tricky point that isn't brought out too clearly in this notation. In S the independent variables are $\{F_k, \alpha\}$. In other words, $S = S(F_1 \ldots F_m; \alpha)$. But in log Z they are $\{\lambda_k, \alpha\}$: $\log Z = \log Z(\lambda_1 \ldots \lambda_m; \alpha)$. So in (10-48) we have to understand that in $(\partial S/\partial \alpha)$ we are holding the F_k fixed, while in $(\partial \log Z/\partial \alpha)$ we are holding the λ_k fixed. The equality of these derivatives then follows from the Legendre transformation (10-41). Evidently, if there are several different parameters $\{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ in the problem, a relation of the form (10-48) will hold for each of them.

Now let's note some general "fluctuation laws," or moment theorems. First, a comment about notation: we're using the symbols F_k, $\langle f_k \rangle$ to stand for the same <u>number</u>. They are equal because I specified that the expectation values $\{\langle f_1 \rangle \ldots \langle f_m \rangle\}$ are to be set equal to the given data $\{F_1 \ldots F_m\}$ of the problem. When I want to emphasize that these quantities are expectation values over the canonical distribution (10-39), I'll use the notation $\langle f_k \rangle$. When I want to emphasize that they are the given data, I'll call them F_k. At the moment, I want to do the former, and so the reciprocity law (10-44) can be written equally well as

$$\frac{\partial \langle f_k \rangle}{\partial \lambda_j} = \frac{\partial \langle f_j \rangle}{\partial \lambda_k} = \frac{\partial^2}{\partial \lambda_j \partial \lambda_k} \log Z \qquad (10\text{-}48)$$

In varying the λ's here, we're changing from one canonical distribution (10-39) to a slightly different one in which the $\langle f_k \rangle$ are slightly different. Since

the new distribution corresponding to $(\lambda_k + d\lambda_k)$ is still of canonical form, it is still a maximum-entropy distribution corresponding to slightly different data $(F_k + dF_k)$. Thus we are comparing two slightly different maximum entropy problems. For later physical applications it will be important to recognize this in interpreting the reciprocity law (10-48).

But now I want to show that the quantities in (10-48) also have an important meaning with reference to a single maximum entropy problem. In the canonical distribution (10-39), how are the different quantities $f_k(x)$ correlated with each other? More specifically, how are departures from their mean values $\langle f_k \rangle$ correlated? The measure of this is the covariance or second central moments of the distribution:

$$\langle (f_j - \langle f_j \rangle)(f_k - \langle f_k \rangle) \rangle$$

$$= \langle [f_j f_k - f_j \langle f_k \rangle - \langle f_j \rangle f_k + \langle f_j \rangle \langle f_k \rangle] \rangle$$

$$= \langle f_j f_k \rangle - \langle f_j \rangle \langle f_k \rangle \qquad (10\text{-}49)$$

If a value of f_k greater than the average $\langle f_k \rangle$ is likely to be accompanied by a value of f_j greater than its average $\langle f_j \rangle$, the covariance is positive; if they tend to fluctuate in opposite directions, it is negative; and if their variations are uncorrelated, the covariance is zero. If $j = k$, this reduces to the variance:

$$\langle (f_k - \langle f_k \rangle)^2 \rangle = \langle f_k^2 \rangle - \langle f_k \rangle^2 \geq 0 \ . \qquad (10\text{-}50)$$

To calculate these quantities directly from the canonical distribution (10-39), we can first find

$$\langle f_j f_k \rangle = \frac{1}{Z(\lambda_1 \ldots \lambda_m)} \sum_{i=1}^{n} f_j(x_i) f_k(x_i) \exp\{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\}$$

$$= \frac{1}{Z} \sum_{i=1}^{n} \frac{\partial^2}{\partial \lambda_j \partial \lambda_k} \exp\{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\}$$

$$= \frac{1}{Z}\frac{\partial^2 Z}{\partial\lambda_j\partial\lambda_k} \tag{10-51}$$

Then, using (10-42), the covariance becomes

$$\langle f_j f_k\rangle - \langle f_j\rangle\langle f_k\rangle = \frac{1}{Z}\frac{\partial^2 Z}{\partial\lambda_j\partial\lambda_k} - \frac{1}{Z^2}\frac{\partial Z}{\partial\lambda_j}\frac{\partial Z}{\partial\lambda_k}$$

$$= \frac{\partial^2}{\partial\lambda_j\partial\lambda_k}\log Z \tag{10-52}$$

But this is just the quantity (10-48); therefore the reciprocity law takes on a bigger meaning,

$$\langle f_j f_k\rangle - \langle f_j\rangle\langle f_k\rangle = -\frac{\partial\langle f_j\rangle}{\partial\lambda_k} = -\frac{\partial\langle f_k\rangle}{\partial\lambda_j} \tag{10-53}$$

That second derivative of log Z which gave us the reciprocity law also gives us the covariance of f_j and f_k in our distribution.

Note that (10-53) is in turn only a special case of a more general rule: Let q(x) be any function; then the covariance with $f_k(x)$ is, as you easily verify,

$$\langle q f_k\rangle - \langle q\rangle\langle f_k\rangle = -\frac{\partial\langle q\rangle}{\partial\lambda_k} \quad , \tag{10-54}$$

a relation that I hadn't noticed in several years of using this formalism, until it was pointed out to me by my former student, Dr. Baldwin Robertson.

From comparing (10-42), (10-48), (10-53) we might expect that still higher derivatives of log Z would correspond to higher moments of the distribution (10-39). This is easily checked; for the third central moments of the f_k we have

$$\langle (f_j - \langle f_j\rangle)(f_k - \langle f_k\rangle)(f_r - \langle f_r\rangle)\rangle$$

$$= \langle f_j f_k f_r\rangle - \langle f_j\rangle\langle f_k f_r\rangle - \langle f_k\rangle\langle f_j f_r\rangle - \langle f_r\rangle\langle f_j f_k\rangle + 2\langle f_j\rangle\langle f_k\rangle\langle f_r\rangle$$

$$= -\frac{\partial^3}{\partial\lambda_j\partial\lambda_k\partial\lambda_r}\log Z \tag{10-55}$$

and in general, all the central moments are given by

$$\langle (f_i - \langle f_i \rangle)^{m_i} (f_j - \langle f_j \rangle)^{m_j} \ldots \rangle$$

$$= (-)^{m_i+m_j+\cdots} \left(\frac{\partial^{m_i}}{\partial \lambda_i^{m_i}} \frac{\partial^{m_j}}{\partial \lambda_j^{m_j}} \cdots \right) \log Z \qquad (10\text{-}56)$$

For noncentral moments, it is customary to define a <u>moment generating function</u>

$$\phi(\beta_1 \ldots \beta_m) \equiv \langle \exp[\beta_1 f_1 + \ldots + \beta_m f_m] \rangle \qquad (10\text{-}57)$$

which evidently has the property

$$\langle f_i^{m_i} f_j^{m_j} \ldots \rangle = \left(\frac{\partial^{m_i}}{\partial \beta_i^{m_i}} \frac{\partial^{m_j}}{\partial \beta_j^{m_j}} \cdots \right) \phi(\beta_1 \ldots \beta_m) \Big|_{\beta_k \neq 0} \qquad (10\text{-}58)$$

However, we find from (10-57)

$$\phi(\beta_1 \ldots \beta_m) = \frac{Z[(\lambda_1 - \beta_1), \ldots, (\lambda_m - \beta_m)]}{Z(\lambda_1 \ldots \lambda_m)} \qquad (10\text{-}59)$$

so that the partition function Z serves this purpose; instead of (10-58) we may write equally well,

$$\langle f_i^{m_i} f_j^{m_j} \ldots \rangle = \frac{1}{Z} \left(\frac{\partial^{m_i}}{\partial \lambda_i^{m_i}} \frac{\partial^{m_j}}{\partial \lambda_j^{m_j}} \cdots \right) Z \qquad (10\text{-}60)$$

which is the generalization of (10-51).

Now, we might ask, what are the covariances of the derivatives of f_k with respect to a parameter α? Let's define

$$g_k \equiv \frac{\partial f_k}{\partial \alpha} \quad , \qquad (10\text{-}61)$$

if f_k is the energy and α is the volume then $-g_k$ is the pressure. The law

for the fluctuation of these is, by a similar derivation that I'll leave for you to work out,

$$\sum_{j=1}^{m} \lambda_j [\langle g_j g_k \rangle - \langle g_j \rangle \langle g_k \rangle] = \left\langle \frac{\partial g_k}{\partial \alpha} \right\rangle - \frac{\partial \langle g_k \rangle}{\partial \alpha} \tag{10-62}$$

a very interesting thing. I had found and used special cases of this for some time, before I finally realized it's actually completely general.

Other derivatives of log Z are related to various moments of the f_k and their derivatives with respect to α. For example, closely related to (10-62) is

$$\frac{\partial^2 \log Z}{\partial \alpha^2} = \sum_{jk} \lambda_j \lambda_k [\langle g_j g_k \rangle - \langle g_j \rangle \langle g_k \rangle] - \sum_k \lambda_k \left\langle \frac{\partial g_k}{\partial \alpha} \right\rangle \tag{10-63}$$

The cross-derivatives give us a simple and useful relation

$$\frac{\partial^2 \log Z}{\partial \alpha \partial \lambda_k} = -\frac{\partial \langle f_k \rangle}{\partial \alpha}$$

$$= \sum_j \lambda_j [\langle f_k g_j \rangle - \langle f_k \rangle \langle g_j \rangle] - \langle g_k \rangle \tag{10-64}$$

which also follows from (10-48) and (10-54); and by taking further derivatives an infinite hierarchy of similar moment relations is obtained. As we will see later, the above theorems have many applications in calculating the fluctuations in pressure of a gas or liquid, the voltage fluctuations, or "noise" generated by a reversible electric cell, etc.

Again, it is evident that if several different parameters $\{\alpha_1 \ldots \alpha_r\}$ are present, relations of the above form will hold for each of them; and new ones like

$$\frac{\partial^2 \log Z}{\partial \alpha_1 \partial \alpha_2} = \sum_k \lambda_k \left\langle \frac{\partial^2 f_k}{\partial \alpha_1 \partial \alpha_2} \right\rangle - \sum_{kj} \lambda_j \lambda_k \left[\left\langle \frac{\partial f_k}{\partial \alpha_1} \frac{\partial f_j}{\partial \alpha_2} \right\rangle - \left\langle \frac{\partial f_k}{\partial \alpha_1} \right\rangle \left\langle \frac{\partial f_j}{\partial \alpha_2} \right\rangle \right] \tag{10-65}$$

will appear.

Well, these moment theorems are quite numerous, but easy to derive. Because of the relation (10-41) between $\log Z(\lambda_1 \ldots \lambda_m; \alpha_1 \ldots \alpha_m)$ and

$S(\langle f_1\rangle \ldots \langle f_m\rangle ; \alpha_1 \ldots \alpha_r)$, you can see that they can all be stated also in terms of derivatives (i.e. variational properties) of S. In the case of S, however, there is a still more general and important variational property that I want to develop.

In (10-43) we supposed that the definitions of the functions $f_k(x)$ were fixed once and for all, the variation in $\langle f_k\rangle$ being due only to variations in the p_i. We now derive a more general variational statement in which both of these quantities are varied. Let $\delta f_k(x_i)$ be specified arbitrarily and independently for each value of k and i, let $\delta\langle f_k\rangle$ be specified independently of the $\delta f_k(x_i)$, and consider the resulting change from one maximum-entropy distribution p_i to a slightly different one $p_i' = p_i + \delta p_i$, the variations δp_i and $\delta\lambda_k$ being determined in terms of $\delta f_k(x_i)$ and $\delta\langle f_k\rangle$ through the above equations. In other words, we are now considering two slightly different maximum-entropy problems in which all conditions of the problem--including the definitions of the functions $f_k(x)$ on which it is based--are varied arbitrarily. The variation in log Z is

$$\delta \log Z = \frac{1}{Z} \sum_{i=1}^{n} \Big\{ \sum_{k=1}^{m} [-\lambda_k \delta f_k(x_i) - \delta\lambda_k f_k(x_i)] \cdot \exp\Big[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\Big] \Big\}$$

$$= -\sum_{k=1}^{m} [\lambda_k \langle \delta f_k\rangle + \delta\lambda_k \langle f_k\rangle] \qquad (10\text{-}66)$$

and thus from the Legendre transformation (10-41)

$$\delta S = -\sum_k \lambda_k [\delta\langle f_k\rangle - \langle \delta f_k\rangle]$$

or,

$$\delta S = \sum_k \lambda_k Q_k \qquad (10\text{-}67)$$

where

$$\delta Q_k \equiv \delta\langle f_k\rangle - \langle \delta f_k\rangle$$

$$= \sum_{i=1}^{n} f_k(x_i)\ \delta p_i \qquad (10\text{-}68)$$

This result, which generalizes (10-43), shows that the entropy S is stationary not only in the sense of the maximization property which led to the canonical distribution (10-39); it is also stationary with respect to small variations in the functions $f_k(x_i)$ if the p_i are held fixed.

As a special case of (10-67), suppose that the functions f_k contain parameters $\{\alpha_1 \ldots \alpha_r\}$ as in (10-65), which generate the $\delta f_k(x_i)$ by

$$\delta f_k(x_i,\alpha_j) = \sum_{j=1}^{r} \frac{\partial f_k(x_i,\alpha)}{\partial \alpha_j}\ \delta\alpha_j \quad . \qquad (10\text{-}69)$$

While δQ_k is not in general the exact differential of any function $Q_k(\langle f_1\rangle \ldots \langle f_m\rangle;\alpha_1 \ldots \alpha_r)$, Eq. (10-67) shows that λ_k is an integrating factor such that $\sum \lambda_k \delta Q_k$ is the exact differential of a "state function" $S(\langle f_1\rangle \ldots \langle f_m\rangle;\alpha_1 \ldots \alpha_r)$. At this point, perhaps all this is beginning to sound vaguely familiar.

Finally, I leave it for you to prove from (10-67) that

$$\sum_{k=1}^{m} \langle f_k\rangle \frac{\partial \lambda_k}{\partial \alpha} = 0 \qquad (10\text{-}70)$$

where $\langle f_1\rangle \ldots \langle f_m\rangle$ are held constant in the differentiation.

Evidently, there's now a large new class of problems which we can ask the robot to do, which he can solve in rather a wholesale way. He first evaluates this partition function Z, or better still, log Z. Then just by differentiating that with respect to everything in sight, he obtains all sorts of predictions in the form of mean values. This is quite a neat mathematical procedure, and, of course, you recognize what we have been doing here. These equations are all just the standard equations of statistical mechanics, in a disembodied form with all the physics removed. In the next lecture, we'll examine that application; but from the way we derived it, it's already clear that this same mathematics also has a lot of other applications outside of physics.

10.8. Conceptual Problems--Frequency Correspondence.

The principle of maximum entropy is basically a simple and straightforward idea, and in the case that the given information consists of average values it leads, as we have just seen, to a surprisingly concise mathematical formalism, since essentially everything is known if we can evaluate a single function $\log Z(\lambda_1 \ldots \lambda_m; \alpha_1 \ldots \alpha_r)$. Nevertheless, it seems to generate some serious conceptual difficulties, particularly to people who have been trained to think of probability only in the frequency sense. Therefore, before turning to applications, I want to examine, and hopefully resolve, some of these difficulties.

Here are some of the objections that have been raised against the principle of maximum entropy: (A) If the only justification for the canonical distribution (10-39) is "maximum uncertainty," that is a negative thing which can't possibly lead to any useful predictions; you can't get reliable results out of mere ignorance. (B) The probabilities obtained by maximum entropy cannot be relevant to physical predictions because they have nothing to do with frequencies--there is absolutely no reason to suppose that distributions observed experimentally would agree with ones found by maximizing entropy. (C) The principle cannot lead to any definite physical results because different people have different information, which would lead to different distributions--so the results are basically arbitrary. (D) The principle is restricted to the case where the constraints are average values--but almost always the given data $\{F_1 \ldots F_n\}$ are not averages over anything. They are definite measured numbers. When you set them equal to averages, $F_k = \langle f_k \rangle$, you are committing a logical contradiction, for the given data said that f_k had the value F_k; yet you immediately write down a probability distribution that assigns non-zero probabilities to values of $f_k \neq F_k$.

Objection (A) is, of course, nothing but a play on words. The "uncertainty" was always there. Our maximizing the entropy did not create any "ignorance" or "uncertainty;" it is rather the means for honestly recognizing the full extent of the uncertainty already present. It is failure to do this--and as a result using a distribution that implies more knowledge than we really have--that would lead to dangerously unreliable conclusions.

Of course, the information put into the theory as constraints on our maximum-entropy distribution, may be so meager that no reliable predictions can be made from it. But in that case, as we will see later, the theory automatically tells us this. If we emerge with a very broad probability distribution for some quantity θ of interest (such as pressure, magnetization, electric current density, rate of diffusion, etc.), that is the robot's way of telling us: "You haven't given me enough information to determine any definite prediction." But if we get a very sharp distribution for θ [for example--and typical of what does happen in many real problems--if the theory says the odds on θ being in the interval $\theta_o(1 \pm 10^{-6})$ are greater than 10^{10}:1], then the given information was sufficient to make a very definite prediction. But in both cases, and in the intermediate ones between these extremes, the distribution for θ tells us just what conclusions we are entitled to draw about θ, on the basis of the information which was put into the equations.

Now to answer objection (B), I want to show that the situation is vastly more subtle than that. The principle of maximum entropy has, fundamentally, nothing to do with any "random experiment," and some of the most important applications are to cases where the probabilities p_i in (10-39) have no frequency connection for just that reason--the x_i are simply an enumeration of the possibilities, and there are no "random variables" in the problem. However, nothing prevents us from applying the principle of maximum entropy also to those cases where the x_i may be regarded as produced by some random

experiment; and in this case, the question of the relation between maximum-entropy probabilities and observable frequencies is capable of mathematical analysis.

I want to give you this analysis now, and demonstrate that (1) in this case the maximum-entropy probabilities do have a precise connection with frequencies; (2) in most real problems, however, this relation is unnecessary for the usefulness of the method; (3) in fact, the principle of maximum entropy is most useful to us in just those cases where the empirical frequency distribution does not agree with the maximum-entropy probability distribution.

Suppose now that the value of x is determined by some random experiment; at each repetition of the experiment the final result is one of the values x_i, $i = 1, 2, \ldots, n$. But now, instead of asking for the probability p_i, let's ask an entirely different question: on the basis of the available information, what can we say about the relative frequencies f_i with which the various x_i will occur in the long run?

Let the experiment consist of N trials (we are particularly interested in the limit $N \to \infty$, because that is the situation contemplated in the usual frequency theory of probability), and let every conceivable sequence of results be analyzed. Each trial could give, independently, any one of the results $\{x_1 \ldots x_n\}$; and so there a priori n^N conceivable outcomes of the whole experiment. But many of these will be incompatible with the given information (let's suppose again that this consists of average values of serveral functions $f_k(x)$, $k = 1, 2, \ldots, m$; in the end it will be clear that the final conclusions are independent of whether it takes this form or some other). We will, of course, assume that the result of the experiment agrees with this information --if it didn't, then the given information was false and we are doing the wrong problem. In the whole experiment, the result x_1 will be obtained n_1 times, x_2 will be obtained n_2 times, etc. Of course,

$$\sum_{i=1}^{n} n_i = N \tag{10-71}$$

and if the specified mean values F_k are in fact obtained, we have the additional relations

$$\sum_{i=1}^{n} n_i f_k(x_i) = NF_k \quad , \qquad k = 1, 2, \ldots, m \tag{10-72}$$

If $m < n-1$, the relations (10-71), (10-72) are insufficient to determine the relative frequencies $f_i = n_i/N$. Nevertheless, we do have good and strong grounds for preferring some choices of the f_i to others. For, out of the original n^N conceivable outcomes, how many would lead to a given set of sample numbers $\{n_1, n_2, \ldots, n_n\}$? The answer is, of course, the multinomial coefficient

$$W = \frac{N!}{n_1!n_2!\ldots n_n!} = \frac{N!}{(Nf_1)!(Nf_2)!\ldots(Nf_n)!} \tag{10-73}$$

The set of frequencies $\{f_1 \ldots f_n\}$ which can be realized in the greatest number of ways is therefore the one which maximizes W subject to the constraints (10-71), (10-72). Now you see it coming--we can equally well maximize any monotonic increasing function of W, in particular $N^{-1} \log W$; but as $N \to \infty$ we have, as we already saw in (10-17),

$$\frac{1}{N} \log W \to -\sum_{i=1}^{n} f_i \log f_i = H_f \tag{10-74}$$

So you see that, in (10-71), (10-72), (10-74) we have formulated exactly the same mathematical problem as in the maximum-entropy derivation of Sec. (10.6), so the two problems will have the same solution. This derivation is mathematically very reminiscent of the Wallis derivation that I gave you a few minutes ago, but of course the equations now have an entirely different meaning.

You also see that this identity of the mathematical problems will persist whether or not the constraints take the form of mean values. If the given information does consist of mean value--and I want to say more about that in

a moment--then the mathematics is particularly neat, leading to the partition function, etc. But, for given information which places any definite kind of constraint on the problem, we have the same conclusion: the probability distribution which maximizes the entropy is numerically identical with the frequency distribution which can be realized in the greatest number of ways.

The maximum in W is, furthermore, enormously sharp. To show this, let $\{f_1 \ldots f_n\}$ be the set of frequencies which maximizes W and has entropy H_f; and let $\{f_1' \ldots f_n'\}$ be any other set of possible frequencies [i.e. a set which satisfies the constraints (10-71), (10-72)] and has entropy $H_{f'} < H_f$. The ratio (number of ways in which f_i could be realized)/(number of ways in which f_i' could be realized) grows asymptotically, according to (10-74), as

$$\frac{W}{W'} \to \exp\{N(H_f - H_{f'})\} \tag{10-75}$$

and passes all bounds as $N \to \infty$. Therefore, the distribution predicted by maximum entropy can be realized experimentally in overwhelmingly more ways than can any other.

We have here another precise and quite general connection between probability and frequency; once again, it had nothing to do with the definition of probability, but emerged as a mathematical consequence of probability theory, interpreted as the "calculus of inductive reasoning." Two more kinds of connection between probability and frequency, whose precise mathematical statements are different in form, but which have the same practical consequences, will appear later, in lectures 12 and 17.

Now let's turn to objection (C) and analyze the situation there. Does this connection between probability and frequency justify our predicting that the maximum-entropy distribution will in fact be observed in a real random experiment? Clearly not, in the sense of deductive proof; for just as objection (C) points out, we have to concede that different people may

have different amounts of information, which will lead them to writing down different distributions, and they can't all be right. But let's look at this more closely. Consider a specific case: Mr. A knows the mean values $\langle f_1(x)\rangle$, $\langle f_2(x)\rangle$. Mr. B knows in addition $\langle f_3(x)\rangle$. Each sets up a maximum-entropy distribution on the basis of his information. Since Mr. B's entropy is maximized subject to one further constraint, we will have

$$H_B \leq H_A \tag{10-76}$$

Suppose that Mr. B's extra information was redundant, in the sense that it was only what Mr. A would have predicted from his distribution. Now Mr. A has maximized his entropy with respect to all variations of the probability distribution which hold $\langle f_1\rangle$, $\langle f_2\rangle$ fixed at the specified values F_1, F_2. Therefore, he has a fortiori maximized it with respect to the smaller class of variations which also hold $\langle f_3\rangle$ fixed at the value finally attained. Therefore Mr. A's distribution also solves Mr. B's problem in this case; $\lambda_3 = 0$, and Mr. A and Mr. B have identical probability distributions. In this case, and only in this case, we have equality in (10-76).

From this example we learn two things: (1) two people with different given information do not necessarily arrive at different maximum-entropy distributions; this is the case only when Mr. B's extra information was "surprising" to Mr. A. (2) In setting up a maximum-entropy problem, it is not necessary to determine whether the different pieces of information used are independent: any redundant information will not be "counted twice," but will drop out of the equations automatically.

Now suppose the opposite extreme: Mr. B's extra information was logically contradictory to what Mr. A knows. For example, it might turn out that $f_3(x) = f_1(x) + 2f_2(x)$, but Mr. B's data failed to satisfy $F_3 = F_1 + 2F_2$. Evidently, there is no probability distribution with this property. How

does our robot tell us this? Mathematically, you will then find that the equations

$$F_k = -\frac{\partial}{\partial \lambda_k} \log Z(\lambda_1,\lambda_2,\lambda_3) \tag{10-77}$$

have no simultaneous solution with real λ_k. In the example just mentioned,

$$\begin{aligned} Z(\lambda_1,\lambda_2,\lambda_3) &= \sum_{i=1}^{n} \exp[-\lambda_1 f_1(x_i) - \lambda_2 f_2(x_i) - \lambda_3 f_3(x_i)] \\ &= \sum_{i=1}^{n} \exp[-(\lambda_1+\lambda_3) f_1(x_i) - (\lambda_2+2\lambda_3) f_2(x_i)] \end{aligned} \tag{10-78}$$

and so

$$\frac{\partial Z}{\partial \lambda_3} = \frac{\partial Z}{\partial \lambda_1} + 2\,\frac{\partial Z}{\partial \lambda_2} \tag{10-79}$$

and so (10-77) cannot have solutions for λ_1, λ_2, λ_3 unless $F_3 = F_1 + 2F_2$. So, when a new piece of information logically contradicts previous information, the principle of maximum entropy breaks down, as it should, giving us no distribution at all.

The most interesting case is the intermediate one where Mr. B's extra information was neither redundant nor contradictory. He then finds a maximum-entropy distribution different from that of Mr. A, and the inequality holds in (10-76), indicating that Mr. B's extra information was "useful" in further narrowing down the range of possibilities allowed by Mr. A's information. The measure of this range is just W; and from (10-75) we have

$$\frac{W_A}{W_B} \sim \exp\{N(H_A - H_B)\} \tag{10-80}$$

For large N, even a slight decrease in the entropy leads to an enormous decrease in the number of possibilities.

Suppose now that we start performing the random experiment with Mr. A and Mr. B watching. Since Mr. A predicts a mean value $\langle f_3 \rangle$ different from the correct one known to Mr. B, it is clear that the experimental distribution

cannot agree in all respects with Mr. A's prediction. We cannot be sure in advance that it will agree with Mr. B's prediction either, for there may be still further constraints $f_4(x)$, $f_5(x)$, ..., etc. operating in the experiment but unknown to Mr. B.

However, the property demonstrated above does justify the following weaker statement of frequency correspondence: If the information incorporated into the maximum-entropy analysis includes all the constraints actually operative in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally, because it can be realized in overwhelmingly the greatest number of ways.

Conversely, if the experiment fails to confirm the maximum-entropy prediction, and this disagreement persists on indefinite repetition of the experiment, then we will conclude that the physical mechanism of the experiment must contain additional constraints which were not taken into account in the maximum-entropy calculation. The observed deviations then provide a clue as to the nature of these new constraints. In this way, Mr. A can discover empirically that his information was incomplete.

Now the little scenario just described is an accurate model of just what did happen in one of the most important applications of statistical analysis, carried out by J. Willard Gibbs. By the year 1901 it was known that in classical statistical mechanics, use of the canonical ensemble (which Gibbs derived as the maximum-entropy distribution over classical phase volume, based on a specified mean value of the energy) failed to predict thermodynamic properties (heat capacities, equations of state, equilibrium constants, etc.) correctly. Analysis of the data showed that the entropy of a real physical system was always less than the value predicted. At that time, therefore, Gibbs was in just the position of Mr. A in the scenario, and he drew the conclusion that the microscopic laws of physics must involve additional

constraints not contained in the laws of classical mechanics. Unfortunately, Gibbs died in 1903 and it was left to others to find the nature of this constraint; first by Planck in the case of radiation, then by Einstein and Debye for solids, and finally by Bohr for isolated atoms. The constraint consisted in the discreteness of the possible energy values, thenceforth called energy levels. By 1927, the mathematical theory by which these could be calculated had been developed by Heisenberg and Schrödinger.

Thus it is an historical fact that the first clues indicating the need for the quantum theory, and indicating some necessary features of the new theory, were uncovered by a seemingly "unsuccessful" application of the principle of maximum entropy. We may expect that such things will happen again in the future, and this is the basis of the remark that the principle of maximum entropy is most useful to us in just those cases where it fails to predict the correct experimental facts.

Gibbs (1902) wrote his probability density in phase space in the form

$$w(q_1 \cdots q_n; p_1 \cdots p_n) = \exp[\eta(q_1 \cdots p_n)] \qquad (10\text{-}81)$$

and called the function η the "index of probability of phase." He derived his canonical and grand canonical ensembles from constraints on average energy, and average energy and particle numbers, respectively, as (loc. cit., p. 143) "the distribution in phase which without violating this condition gives the least value of the average index of probability of phase $\bar{\eta}$ ...." This is, of course, just what we would describe today as maximizing the entropy subject to constraints.

Unfortunately, Gibbs did not give any clear explanation, and we can only conjecture whether he possessed one, as to why this particular function is to be minimized on the average, in preference to all others. Consequently, his procedure appeared arbitrary to many, and for sixty years there was

controversy over the validity and justification of Gibbs' method. In spite of its enormous practical success when adapted to quantum statistics, few attempts were made to extend it beyond problems of thermal equilibrium.

It was not until the work of Shannon in our own time that the full significance of Gibbs' method could be appreciated. Once we had Shannon's theorem establishing the uniqueness of entropy as an "information measure," it was clear that Gibbs' procedure was an example of a general method for inductive inference, whose applicability is in no way restricted to equilibrium thermodynamics or to physics.

Lecture 13

INTRODUCTION TO DECISION THEORY

"Your act was unwise," I exclaimed "as you see
by the outcome." He solemnly eyed me.
"When choosing the course of my action," said he,
"I had not the outcome to guide me."

----Ambrose Bierce

At this stage we have accumulated quite a few loose ends, which I would now like to clear up. In almost every lecture so far, I had to insert one or more parenthetic remarks to the effect that "there is still an essential point missing here, which will be supplied when we take up decision theory." Actually, we began seeing what it is, as soon as we started applying the theory to our first problem. When we illustrated the use of Bayes' theorem by sequential testing in Lecture 5, we noted that there is nothing in probability theory *per se* which could tell us where to put the threshold levels at which we make our decision: whether to accept the batch, reject it, or make another test. At that time, I said only that the location of this threshold level obviously depends in some way on our judgment as to what are the consequences of making wrong decisions, and what are the costs of making further tests. Qualitatively, this is clear enough; but before we can claim to have a really complete design for our robot, we must re-state this in quantitative terms.

The same situation occurred in Lecture 8 when we studied particle counters, and the robot was faced with the job of estimating the number of particles which had passed through the counter under various conditions. Probability theory told us only the robot's state of knowledge as to the number of particles; it did not tell us what estimate he should in fact make. We noted at that time that taking the mean value over the posterior distribution was the same as making that decision which minimizes the expected square of the error; and in Lecture 11 we followed the same procedure for statistical mechanics. In both of those cases, this seems to be a fairly sensible criterion, and leads to results in good correspondence with common sense. However, why was it the square of the error that we minimized? Why not some other function of the error? The criterion of minimum mean square error has obvious mathematical advantages, because the mean value of a distribution is generally easy to calculate; but in principle it appears to be entirely arbitrary.

You see the common feature of all these problems. In every case, probability theory can give us only a probability distribution which represents the robot's final state of knowledge with all the available data taken into account; but in practice his job is to make a definite decision. He must act as though one hypothesis were true, he must make a definite numerical estimate of some parameter, and so on. The essential thing which is still missing in our design of this robot is the rule by which he converts his final probability assignment into a definite course of action.

13.1. Daniel Bernoulli's Suggestion.

As you might expect from the way this situation appeared in the most elementary applications of probability theory, this problem is by no means new. It was clearly recognized, and a definite solution offered for a certain class of problems, by Daniel Bernoulli in the year 1738. In a cruder form, the

same principle had been seen even earlier, at the time when probability theory was concerned almost exclusively with problems of gambling. The notion which seemed very intuitive to the first workers in probability theory was "expectation of profit." By this we mean, of course, that I consider each possibility, $i = 1, 2, \ldots, n$, assign probabilities p_i to them, and also assign numbers M_i which represent the profit I would obtain if the i'th possibility should in fact turn out to be true. Then the quantity

$$\langle M \rangle = \sum_{i=1}^{n} p_i M_i \qquad (13\text{-}1)$$

is what we call the "expectation of profit." It seemed obvious to the first workers in probability theory that a gambler acting in pure self-interest should always behave in such a way as to maximize his expected profit. This, however, led to some paradoxes (particularly in the famous St. Petersburg problem) which led Bernoulli to recognize that simple expectation of profit is not always a sensible criterion of action.

To give a very simple example, suppose that I assign probability 0.51 to heads in a certain slightly biased coin. Now I am given the choice of two actions: (1) to bet every cent I have at even money, on heads for the next toss of this coin; (2) not to bet at all. According to the criterion of expectation of profit, I should always choose to gamble when faced with this choice. My expectation of profit, if I do not gamble, is zero; but if I do gamble, it is

$$\langle M \rangle = 0.51\, M_o + 0.49\, (-M_o) = 0.02\, M_o > 0 \qquad (13\text{-}2)$$

where M_o is the amount I have now. Nevertheless it seemed obvious to Bernoulli, and I think it does also to you, that very few people would really choose the first alternative in the problem as stated. This means that our common sense, in some cases, rejects the criterion of maximizing expected profit.

Suppose that you are offered the following opportunity. You can bet any

amount you want on the basis that, with probability $(1 - 10^{-6})$, you will lose your money; but with probability 10^{-6}, you will win 1,000,001 times the amount you had wagered. Again, the criterion of maximizing expected profit says that you should bet all the money you have. Our common sense rejects this solution even more forcefully; no sane person would risk all his fortune, which he is practically certain to lose, for an infinitesimal chance of winning a very much larger sum.

Daniel Bernoulli proposed to resolve these paradoxes by recognition that the true value to a person, of receiving a certain amount of money, is not measured simply by the amount received; it depends also upon how much he has already. In other words, Bernoulli said that we should recognize that the mathematical expectation of profit is not the same thing as its "moral expectation." A modern economist is expressing exactly the same idea when he speaks of the "diminishing marginal utility of money."

The original St. Petersburg game consists of the following--we toss an honest coin until it comes up heads for the first time. The game is then terminated. If heads occurs for the first time at the n'th throw, the player receives 2^n dollars. The question is: what is a "fair" entrance fee for him to pay, for the privilege of playing this game? If we use the criterion that a fair game is one where the entrance fee is equal to the expectation of profit, you see what happens. This expectation is

$$\sum_{K=1}^{\infty} (2^{-K})(2^{K}) = \sum_{K=1}^{\infty} 1 \qquad (13\text{-}3)$$

and this is infinite. Nevertheless it is clear again that no sane person would be willing to risk more than a very small part of his fortune for the privilege of playing this game. Let me quote Laplace (1819) at this point:

"Indeed, it is apparent that one franc has much greater value for him who possesses only 100 than for a millionaire. We ought then to distinguish

the absolute value of the hoped-for benefit from its relative value. The latter is regulated by the motives which make it desirable, whereas the first is independent of them. The general principle for appreciating this relative value cannot be given, but here is one proposed by Daniel Bernoulli which will serve in many cases: The relative value of an infinitely small sum is equal to its absolute value divided by the total fortune of the person interested."

In other words, Bernoulli proposed that the "moral value," or what the modern economist would call the "utility" of money should be taken proportional to its logarithm. Laplace, in discussing the St. Petersburg problem and this criterion, reports the following result without giving the calculation: a person whose total fortune is 200 francs ought not reasonably to stake more than 9 francs on the play of this game. I took the trouble of checking this. The fair fee f(200) is found by equating his present utility with his expected utility if he pays the fee and plays the game; a computer gives the root of

$$\log 200 = \sum_{n=1}^{\infty} \frac{1}{2^n} \log(200 - f + 2^n)$$

as f(200) = 8.7204. Likewise, $f(10^3) = 10.98$, $f(10^4) = 14.24$ $f(10^6) = 20.87$.

It seems to me that this kind of numerical result is entirely reasonable. However the logarithmic assignment of utility is not to be taken literally either in the case of extremely small fortunes (as Laplace points out), or in the case of extremely large ones, as the following example of Savage (1954) shows. Suppose your total fortune is 10,000,000 dollars; then if your utility for money is proportional to the logarithm of the amount, the theory says that you should be as willing as not to accept a wager in which, with probability one-half, you'll be left with only 10,000 dollars; and with probability one-half, you will be left with 10,000,000,000 dollars. I think that most of us would consider such a bet to be distinctly disadvantageous to a person with that initial fortune. This shows that our intuitive "utility" for money actually must increase even less rapidly than the logarithm for extremely

large values. There are some who even claim that it is bounded.

The gist of Daniel Bernoulli's suggestion was therefore that, in the gambler's problem of decision making under uncertainty, one should act so as to maximize the expected value, not necessarily of the profit itself, but of some function of the profit which he called the "moral value". In more modern terminology the optimist will call this "maximizing expected utility," while the pessimist will speak instead of "minimizing expected loss", the loss function being taken as the negative of the utility function.

The logarithmic assignment of utility is reasonable for many purposes, as long as it is not pushed to extremes. It is also, incidentally, very closely connected with the notion of entropy, as shown by an argument of Kelly (1956), extended by Bellman and Kalaba (1956). Here, a gambler who receives advance tips on a game which are only partly reliable, acts (i.e., decides on which side and how much to bet) so as to maximize the expected logarithm of his fortune. They show that (1) one can never go broke following this strategy, in contrast to the strategy of maximizing expected profit, where it is easily seen that with probability one this will eventually happen, and (2) the amount one can reasonably expect to win on any one game is clearly proportional to the amount M_o he has to begin with, so after n games, one could hope to have an amount $M = M_o e^{\alpha n}$. With the logarithmic utility function, one acts so as to maximize the expected value of α. The maximum attainable $\langle\alpha\rangle$ turns out to be just $(S_o - S)$, where S is the entropy which describes the gambler's uncertainty as to the truth of his tips, and S_o is the maximum possible entropy, if the tips were completely unreliable. This suggests that, with a little more development of the theory, entropy might have an important place in guiding the strategy of a stock market investor.

Daniel Bernoulli's solution to the problem of decision making has suffered the same fate as did Laplace's solution to the problem of inductive reasoning.

The "objectivist" or "orthodox" school of thought either ignored it or condemned it as metaphysical nonsense until just a few years ago. In one of the best known books on probability theory (Feller, 1950; p. 199), Daniel Bernoulli's solution of the St. Petersburg paradox is rejected without even being described, except to assure the reader that he "tried in vain to solve it by the concept of moral expectation." Well, we will see next just how vain Daniel Bernoulli's efforts were.

13.2. The New Formulation of the Decision Problem.

In the late 1940's a general theory of decision making in the face of uncertainty was developed, largely by Wald (1950) which in its initial stages had no apparent connection with probability theory. I mentioned it briefly in Lecture 5, and now I would like to give you a more specific account of some of the ideas it involved.

We begin by imagining (i.e. enumerating) a set of possible unknown "states of nature", $\{\theta_1, \theta_2, \ldots, \theta_N\}$ whose number might be finite or infinite. The θ_j might also form a continuum. In the quality-control example of Lecture 5, the "state of nature" is the unknown number of defectives in the batch, and the θ_j are discrete. In the particle-counter problem of Lecture 8, the state of nature could be taken as the unknown source strength s, and the θ_j are continuous.

There are certain illusions that tend to grow and propagate here. Let me dispel one right now by noting that, in enumerating the different states of nature, we are not describing any objective (measurable) property of nature --for, one and only one of them is in fact true. The enumeration is only a means of describing our state of ignorance. It is, therefore, meaningless to ask whether one particular enumeration in "correct" without first asking, "what is the information that is being described by the set of θ_j?" Two

observers with different amounts of information may enumerate θ_j differently without either being inconsistent.

The next step in our theory is to make a similar enumeration of the possible decisions $\{D_1, D_2, \ldots, D_k\}$ that might be made. In the quality-control example, there were three possible decisions at each stage:

$$D_1 = \text{accept the batch}$$
$$D_2 = \text{reject it}$$
$$D_3 = \text{make another test} \qquad (13\text{-}4)$$

In the particle counter problem of Mr. B, where we are to estimate the number n_1 of particles passing through the counter in the first second, there are an infinite number of possible decisions:

$$D_i = \text{"}n_1 \text{ is estimated as equal to } i\text{,"} \quad i = 0, 1, 2, \ldots \qquad (13\text{-}5)$$

If we are to estimate the source strength, then there is a continuum of possible decisions.

This theory is clearly of no use unless by "making a decision" we mean "deciding to act as if the decision were correct". It is idle to "decide" that $n_1 = 150$ is the best estimate unless we are then prepared to act on the assumption that $n_1 = 150$. Thus the enumeration of the D_i is a means of describing our knowledge as to what kinds of actions are <u>feasible</u>; it is idle to consider any decision which we know in advance corresponds to an impossible course of action.

There is another reason why a particular decision might be eliminated; even though D_1 is easy to carry out, we might know in advance that it would lead to intolerable consequences. An automobile driver can make a sharp left turn at any time; but his common sense usually tells him not to. Here we see two more points: (1) there is a continuous gradation--the consequences of an action might be serious without being absolutely intolerable, and (2) the consequences of an action (=decision) will in general depend on what is the

true state of nature--a sharp left turn does not always lead to disaster.

This suggests a third concept we need--the loss function $L(D_i,\theta_j)$, which is a set of numbers representing our judgment as to the "loss" incurred by making decision D_i if θ_j should turn out to be the true state of nature. If the D_i and θ_j are both discrete, this becomes a loss matrix L_{ij}.

Quite a bit can be done with just the θ_j, D_i, L_{ij} and there is a rather extensive literature dealing with criteria for making decisions with no more than this. The material we need for our purposes has been summarized in a very readable and entertaining form by Luce and Raiffa (1957), and in the elementary textbook of Chernoff and Moses (1959). The minimax criterion is this: for each D_i find the maximum possible loss $M_i = \max_j(L_{ij})$; then choose that D_i for which M_i is a minimum. The minimax criterion would be a reasonable one if we regard nature as an intelligent adversary who foresees our decision and deliberately chooses the state of nature so as to cause us the maximum frustration. In the theory of some games, this is not a completely unrealistic way of describing the situation, and consequently minimax strategies are of fundamental importance in game theory. But in the decision problems of the scientist or engineer the minimax criterion is that of the long-faced pessimist who concentrates all his attention on the worst possible thing that could happen, and thereby misses out on the favorable opportunities.

Equally unreasonable for us is the opposite extreme of the starry-eyed optimist who uses this "minimin" criterion: for each D_i find the minimum possible loss $m_i \equiv \min_j(L_{ij})$ and choose the D_i that makes m_i a minimum.

Evidently, a reasonable decision criterion for the scientist and engineer is, in some sense, intermediate between minimax and minimin. Many other criteria have been suggested, which go by the names of maximum utility (Wald), α-optimism-pessimism (Hurwicz), minimax regret (Savage), etc. The usual procedure, as described in detail by Luce and Raiffa, has been to analyze any

proposed criterion to see whether it satisfies about a dozen qualitative common-sense conditions such as (1) Transitivity: if D_1 is preferred to D_2, and D_2 preferred to D_3, then D_1 must be preferred to D_3, and (2) Strong Domination: if for all states of nature θ_j we have $L_{ij} < L_{kj}$, then D_i should always be preferred to D_k. This analysis, although straightforward, can become tedious. I will not follow it any further, because the final result is that there is only one class of decision criteria which passes all the tests, and this class is obtained more easily by a different line of reasoning.

A full decision theory, of course, cannot concern itself merely with the θ_j, D_i, L_{ij}. We also, in typical problems, have additional evidence E, which we recognize as relevant to the decision problem, and we have to learn how to incorporate E into the theory. In the quality-control example, E consisted of the results of the previous tests.

At this point, current decision theory takes a long, and I think unnecessary, mathematical detour. One defines a "strategy", which is a set of rules of the form, "If I receive new evidence E_i, then I will make decision D_k." In principle one first enumerates all conceivable strategies (whose number is, however, astronomical even in quite simple problems), and then tries to eliminate the undesirable ones by application of various common-sense conditions. This leads to defining a class of "admissible" strategies, which consists, crudely speaking, of all those any sane person would ever consider adopting; a strategy is admissible if no other exists which is as good or better for all states of nature.

A principal object of the theory is then to characterize the class of admissible strategies in mathematical terms, so that any such strategy can be found by carrying out a definite procedure. The fundamental theorem bearing on this is Wald's Complete Class Theorem which establishes a result already mentioned in Lecture 5. Instead of following this rather difficult argument,

I would like to make a few more remarks about the nature of the problem, and then give a different line of reasoning which leads to the same result by elementary mathematics.

What is it that makes a decision process difficult? Well, if we knew which state of nature was the correct one, there would be no problem at all; if θ_3 is the true state of nature, then the best decision D_i is the one which renders L_{i3} a minimum. In other words, once the loss function has been specified, our uncertainty as to the best decision arises solely from our uncertainty as to the state of nature. Whether the decision minimizing L_{i3} is or is not best depends entirely on this: How strongly do we believe that θ_3 is the true state of nature? How plausible is θ_3?

To a physicist or engineer it seems like a very small step--really only a rephrasing of the question--to ask next, "Conditional on all the available evidence, what is the probability P_3 that θ_3 is the true state of nature?" Not so to the orthodox statistician, who regards the word "probability" as synonomous with "long-run relative frequency in some random experiment". On this definition it is meaningless to speak of the probability of θ_3, because the state of nature is not a "random variable". Thus, if we adhere consistently to the orthodox view of probability, we will have to conclude that probability theory cannot be applied to the decision problem, at least not in this direct way.

It was just this kind of reasoning which led statisticians, in the early part of this century, to relegate problems of parameter estimation and hypothesis testing (which are really decision problems and as such are included in our general formulation) to a new field, Statistical Inference, which was regarded as distinct from probability theory. But let us look in detail at a typical problem of this type, using the loss function criterion, from the orthodox viewpoint. I want to show that a rather simple extension of the usual

orthodox arguments leads us to the same conclusion that Wald's much deeper analysis forced him to (very much against his will): that the original methods proposed by Laplace and Daniel Bernoulli are, in fact, the unique solution of the decision problem.

13.3. Parameter Estimation for Minimum Loss.

One of the situations considered in the discussion of particle counters (Lecture 8) was that of Mr. B, who knew that there was a constant, but unknown, source strength s. By observing the number of counts $\{c_1,\ldots,c_n\}$ in several different seconds, he could make an estimate of the numerical value of s, which presumably became more and more accurate with increasing n. This is a typical example of the general problem of parameter estimation.

More generally, suppose that there is ome unknown parameter α, and we make repeated observations of some quantity, obtaining an observed "sample", $x = \{x_1,\ldots,x_n\}$. We can interpret the symbol x, without subscripts, as standing for a vector in an n-dimensional "sample space". We will suppose that the possible results x_i of individual observations are real numbers. From observation of the sample x, what can we say about the unknown parameter α?

To state the problem more drastically, suppose that we are compelled to choose one specific numerical value as our "best" estimate of α, on the basis of the observed sample x, and any other prior information we might have. This is the decision situation which we all face daily, both in our capacity as scientists and engineers, and in everyday life. The driver approaching a blind intersection cannot know with certainty whether he will have enough time to cross it safely; but still he is compelled to make a decision based on what he can see, and act on it.

Now it is clear that in estimating α, the observed sample x is of no use to us unless α exerts some kind of influence on x. In other words, if we

knew α, but not x, then the probabilities $(x|\alpha) = (x_1 \ldots x_n|\alpha)$ which we would assign to various samples must depend in some way on the value of α. If we consider the different observations as independent, as is almost always done in the orthodox theory of parameter estimation, then the distribution factors:

$$(x|\alpha) = (x_1|\alpha)\ldots(x_n|\alpha) \tag{13-6}$$

However, this very restrictive assumption is not necessary (and in fact doesn't lead to any formal simplification) in discussing the general principles of parameter estimation from the decision theory standpoint.

Let $\beta = \beta(x_1 \ldots x_n)$ be an "estimator", i.e. any function of the sample values, proposed as an estimate of α. Also, let $L(\alpha,\beta)$ be the "loss" incurred by guessing the value β when α is in fact the true value. Then for any given estimator the expected loss <u>for a person who already knows the true value of α</u>, is

$$L_\alpha = \int L(\alpha,\beta)\,(x|\alpha)\; dx \tag{13-7}$$

Call this the α-expected loss. By $\int(\;)\,dx$ we mean the n-fold integration

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\quad)\;dx_1\ldots dx_n \tag{13-8}$$

There is no need to specify different limits of integration for different problems, since if certain ranges of the x_i are impossible, the factor $(x|\alpha)$ will be zero and remove contributions from those ranges. Also, this notation includes both the continuous and discrete case, since in the latter $(x|\alpha)$ is a sum of delta-functions.

On the view of one who uses the frequency definition of probability, the above phrase, "for a person who already knows the true value of α" is misleading and unwanted. The notion of the probability of sample x <u>for a person with a certain state of knowledge</u> is entirely foreign to him; he regards $(x|\alpha)$ not as a description of a mere state of knowledge about the sample, but as an objective statement of fact, giving the relative frequencies with which dif-

ferent samples are observed "in the long run". Thus the "frequentist" believes that L_α is not merely the "mathematical expectation" of loss in the present situation, but is also, with probability 1, the limit of the average of actual losses which would be incurred by using the estimator β an indefinitely large number of times. Furthermore, the idea of finding the estimator which is "best in the present specific case" is quite foreign to his outlook; because he regards the notion of probability as meaningful only in the sense of limiting frequencies, he is forced to speak instead of finding that estimator "which will prove best in the long run".

On the frequentist view, therefore, it would appear that the best estimator will be the one that minimizes L_α. Is this a variational problem? A change $\delta\beta(x)$ in the estimator produces a change of L_α of

$$\delta L_\alpha = \int \frac{\partial L}{\partial \beta}\,(x|\alpha)\;\delta\beta(x)\;dx. \qquad (13\text{-}9)$$

If we were to require this to vanish for all $\delta\beta(x)$, this would mean

$$\frac{\partial L}{\partial \beta} = 0 \quad \text{for all possible values of } \beta. \qquad (13\text{-}10)$$

Thus the problem as stated has no truly stationary solution except in the trivial case where the loss function is independent of the estimated value β; the best estimator by the criterion of minimum α-expected loss cannot be found by variational methods. Nevertheless, we can get some understanding of the problem by considering (13-7) for some specific choices of loss function. Suppose we take the quadratic loss function $L(\alpha,\beta) = (\alpha - \beta)^2$. Then (13-7) reduces to

$$L_\alpha = \alpha^2 - 2\alpha\langle\beta\rangle + \langle\beta^2\rangle \qquad (13\text{-}11)$$

or,

$$L_\alpha = (\alpha - \langle\beta\rangle)^2 + \text{var}\,(\beta) \qquad (13\text{-}12)$$

where $\text{var}\,(\beta) = \langle\beta^2\rangle - \langle\beta\rangle^2$ is the variance, and the n'th moment

$$\langle \beta^n \rangle = \int [\beta(x)]^n \, (x|\alpha) \, dx \qquad (13\text{-}13)$$

is the α-expected value of β^n. The α-expected loss is the sum of two positive terms, and a good estimator by the criterion of minimum α-expected loss has two properties:

(1) $\langle \beta \rangle = \alpha$

(2) var (β) is a minimum. (13-14)

These are just the two conditions which orthodox statistics has considered most important. An estimator with property (1) is called an unbiased estimate [more generally, the function $b(\alpha) = \langle \beta \rangle - \alpha$ is called the bias of the estimator $\beta(x)$], and one with both properties (1) and (2) was called efficient by R. A. Fisher (although this last condition is ambiguous until we specify the class of functions $\beta(x)$ to be taken into consideration). Nowadays, it is often called an unbiased minimum variance (UMV) estimator.

It has always seemed to me that the above reasoning amounts to looking at the problem backwards. We are describing the situation as it appears to a person who already knows the correct value of α, but does not know which specific sample has been observed. The above equations really refer to only one value of α, but involve many different possible values of x. But this is just the opposite of the state of knowledge which we have when we estimate a parameter; we know x, but not α. Our equations should involve only one sample, namely the one actually observed; but sholld take into account many different possible values of α.

Our job is always to do the best reasoning we can about the single situation that exists here and now, on the basis of the knowledge which we do in fact have; consideration of how things might seem to a person whose state of knowledge is different, or what might happen in some other situation that we are not reasoning about (if some different sample were observed) is

not relevant to our problem. So, we ought to do it the other way around; it is the expected value of $L(\alpha,\beta)$ over the posterior distribution $(\alpha|x)$ of α, conditional on knowledge of the sample, that should logically be minimized. Call this the x-expected loss:

$$L_x(\beta) = \int L(\alpha,\beta)(\alpha|x)\, d\alpha \tag{13-15}$$

where $(\alpha|x)$ is obtained by applying Bayes' theorem. Thus, having observed the sample x, we should calculate $L_x(\beta)$ and take as our estimate that value of β which minimizes $L_x(\beta)$. In the continuous case, subject to some elementary regularity conditions, we would use the estimator $\beta(x_1,\ldots,x_n)$ determined by

$$\frac{\partial L_x(\beta)}{\partial \beta} = 0 \tag{13-16}$$

$$\frac{\partial^2 L_x(\beta)}{\partial \beta^2} > 0 \tag{13-17}$$

These equations make no reference to any sample other than the specific one that has been observed.

But most of the prominent workers in statistics would raise strong objections to this procedure on philosophical grounds [you guessed it--that $(\alpha|x)$ is meaningless because α is not a "random variable"]. So, let's go back and take a closer look at the orthodox formulation of the problem--is there some way we could improve it without conflicting with orthodox principles?

We have already seen a practical difficulty faced in the first formulation; the criterion of minimum α-expected loss does not lead to a variational problem, and therefore even in the simplest case of a quadratic loss function, it gives us no analytical method for constructing the "best" estimator $\beta(x_1\ldots x_n)$. In fact, it is clear from (13-14) that the only really correct solution of the mathematical problem as stated, is $\beta(x_1\ldots x_n) = \alpha$, independent of the observed sample. This shows again that the criterion of minimum α-expected loss essentially describes the reasoning of a person who already

knows the correct value of α. However, the stubborn fact remains that the statistician using this criterion does not know α, and so he cannot use the correct solution of the problem. His estimator must be some function of the sample values only. Once an estimator has been suggested, it can be tested by calculating (13-12). But, except for one special class of sampling distributions $(x_1 \ldots x_n|\alpha)$, which I will consider later, the frequentist has no general principle like (13-16), only his judgment and common sense, to tell him which ones to try out in the first place.

13.4. Should We Use an Unbiased Estimate?*

What is the relative importance of removing bias and minimizing the variance? Well, from (13-12) it would appear that they are of exactly equal importance; there is no advantage in removing the bias $(\langle\beta\rangle - \alpha)$ if in so doing we increase var(β) more than enough to compensate. Yet that is just what the orthodox statistician usually does! Let me give you a specific example of this. Cramér (1946, p. 351) considers the problem of estimating the variance μ_2 of a distribution $(x_1|\mu_2)$:

$$\mu_2 = \langle x_1^2\rangle - \langle x_1\rangle^2 = \langle x_1^2\rangle \tag{13-18}$$

from n independent observations $\{x_1 \ldots x_n\}$. We assume, in (13-18) and in what follows, that $\langle x_1\rangle = 0$ since a trivial change of variables would in any event accomplish this. An elementary calculation shows that the sample variance

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^2 \tag{13-19}$$

has expectation value, over the distribution $(x_1 \ldots x_n|\mu_2) = (x_1|\mu_2)\ldots(x_n|\mu_2)$, of

$$\langle m_2\rangle = \frac{n-1}{n}\mu_2 \tag{13-20}$$

---

*This section is a digression in response to a question from the audience. It may be skipped without losing the main line of argument; however, it does contain an illustration of an important point.

and thus, as an estimator of μ_2 it has a negative bias. So, goes the argument, we should correct this by using the unbiased estimator

$$M_2 = \frac{n}{n-1}\, m_2 \tag{13-21}$$

Now, of course, the only thing that really matters here is the <u>total</u> error of our estimate; the particular way in which you or I separate error into two abstractions labelled "bias" and "variance" is a purely academic matter with no bearing on the actual quality of the estimator. So, let's look at the mean square error criterion. Replacement of m_2 by M_2 removes a term $(\langle m_2\rangle - \mu_2)^2 = \mu_2^2/n^2$ in (13-12); but it also increases the term $\mathrm{var}(m_2)$ by a factor $[n/(n-1)]^2$, so it seems obvious that, at least for large n, this has made things worse instead of better. Let's check this more carefully. Suppose we replace m_2 by the estimator,

$$\gamma_\delta = (1 + \delta)\, m_2 \tag{13-22}$$

What is the best choice of δ? The μ_2-expected loss (13-12) is now

$$\langle(\gamma_\delta - \mu_2)^2\rangle = \mu_2^2 - 2(1+\delta)\mu_2\langle m_2\rangle + (1+\delta)^2\langle m_2^2\rangle$$

$$= [(\langle m_2\rangle - \mu_2)^2 + \mathrm{var}(m_2)] - \langle m_2^2\rangle q^2 + \langle m_2^2\rangle(\delta - q)^2 \tag{13-23}$$

where

$$q \equiv \frac{\langle m_2^2\rangle - \mu_2\langle m_2\rangle}{\langle m_2^2\rangle} \tag{13-24}$$

Evidently, the best estimator in the class γ_δ is the one with $\delta = q$, and the term $-\langle m_2^2\rangle q^2$ in (13-23) represents the decrease in mean-square error obtainable by using γ_q instead of m_2. From Cramér's result (loc. cit., Eq. 27.4.2)

$$\mathrm{var}(m_2) = \langle m_2^2\rangle - \langle m_2\rangle^2 = n^{-3}\,(n-1)[(n-1)\mu_4 - (n-3)\mu_2^2] \tag{13-25}$$

where

$$\mu_4 = \langle(x_1 - \langle x_1\rangle)^4\rangle = \langle x_1^4\rangle$$

is the fourth central moment of $(x_1|\mu_2)$, we find

$$n^3\langle m_2^2\rangle = (n-1)\ [(n^2-n+2)\mu_2^2 + (n-1)\ \mathrm{var}(x^2)] \tag{13-26}$$

$$n^3\langle m_2^2\rangle q = (n-1)\ [(n-2)\mu_2^2 - (n-1)\ \mathrm{var}(x^2)] \tag{13-27}$$

where $\mathrm{var}(x^2) = \mu_4 - \mu_2^2 \geq 0.$ (13-28)

We must understand $n > 1$ in all this, for if $n = 1$, we have $m_2 = 0$; a single observation gives no information at all about the variance of $(x_1|\mu_2)$. But if $n = 2$, we have $q \leq 0$; instead of removing the bias, we should always <u>increase</u> it in order to minimize the mean square error! More generally, if $\mathrm{var}(x^2) = K\mu_2^2$ we have from (13-26), (13-27):

$$q = \frac{(n-2) - (n-1)\ K}{(n^2-n+2) + (n-1)\ K} \tag{13-29}$$

and therefore, if $K < 1$,

$$q \lessgtr 0 \quad \text{if } n \lessgtr \frac{2-K}{1-K} \tag{13-30}$$

while if $K \geq 1$, $q < 0$ for all n.

In the case of a Gaussian distribution,

$$(x_1|\mu_2) = A \exp\left[-\frac{x_1^2}{2\mu_2}\right] \tag{13-31}$$

we have

$$K = \frac{\langle x_1^4\rangle - \langle x_1^2\rangle^2}{\langle x_1^2\rangle^2} = 2 \tag{13-32}$$

We will seldom have $K < 2$, for this would imply that $(x_1|\mu_2)$ cuts off even more rapidly than Gaussian for large x_1. If $K = 2$, (13-29) reduces to

$$q = -\frac{1}{n+1} \tag{13-33}$$

which again says that rather than removing the bias we should approximately double it, in order to minimize the expected square of the error. How much better is the estimator γ_q than M_2? In the Gaussian case the mean square error of the estimator γ_q is

$$\langle(\gamma_q - \mu_2)^2\rangle = \frac{2}{n+1}\mu_2^2 \tag{13-34}$$

For a general choice of δ, it is

$$\langle(\gamma_\delta - \mu_2)^2\rangle = \mu_2^2\left[\frac{2}{n+1} + \frac{n^2-1}{n^2}\left(\delta + \frac{1}{n+1}\right)^2\right] \qquad (13\text{-}35)$$

The unbiased estimator M_2 corresponds to the choice

$$\delta = \frac{1}{n-1} \quad , \qquad (13\text{-}36)$$

and thus to the mean square error

$$\langle(M_2 - \mu_2)^2\rangle = \mu_2^2\left[\frac{2}{n+1} + \frac{2}{n}\right] \qquad (13\text{-}37)$$

which is over twóce the amount incurred by use of γ_q.

Most distributions which arise in practice, if not gaussian, have wider "tails" than gaussian so that $K > 2$. In this case, the difference will be even greater.

Up to this point, it may have seemed that I was quibbling over a very minor thing--changes in the estimator of one or two parts out of n. But now you see that the difference between (13-34) and (13-37) is not at all trivial. For example, with Cramér's unbiased estimator M_2 you will need n = 203 observations in order to get as small a mean-square error as the biased estimator γ_q gives you with only 100 observations.

There is a fantastic example in a recent book on econometrics (Valavanis, 1959; p. 60) where the author attaches such great importance to removing bias that he advocates throwing away practically all the data from the sample, if necessary, to achieve this. One reason for such an undue emphasis on bias is the belief that if we draw N successive samples of n observations each and calculate the estimators $\beta_1 \ldots \beta_N$, the average $\bar{\beta} = N^{-1}\sum\beta_i$ of these estimates will converge in probability to $\langle\beta\rangle$ as $N \to \infty$, and thus an unbiased estimator will, on sufficiently prolonged sampling, give an arbitrarily accurate estimate of α. Such a belief is almost never justified even for the fairly well controlled measurements of the physicist or engineer, not only because of

unknown systematic error, but because successive measurements lack the independence required for these limit theorems to apply. In such uncontrolled situations as economics, the situation is far worse.

But unbiased estimators are, even if we accept these limit theorems, not the only ones which appraach perfect accuracy with indefinitely prolonged sampling. Many other estimators approach the true value of α in this limit, and do it more rapidly. Our γ_q is a specific example. Furthermore, asymptotic behavior of an estimator is not really relevant, because the practical problem is always to do the best we can with a finite sample; therefore the important question is not whether an estimator tends to the true value, but how rapidly it does so.

I have a dark suspicion that a still more important reason for attaching such an undeserved importance to bias is simply that we have been caught in a psycho-semantic trap. It is well known to politicians that our thought processes are influenced to a rather alarming degree by the particular choice of words we use. When we call the quantity $\langle\beta\rangle-\alpha$ the "bias", that makes it sound like something awfully reprehensible, which we must get rid of at all costs. If we had called it instead the "component of error orthogonal to the variance", as suggested by the Pythagorean form of Eq. (13-12), then it would be clear to all that these two contributions to the error are on an exactly equal footing; and that it is folly to decrease one at the expense of increasing the other.

In the book of Chernoff and Moses (1959) these points are clearly recognized, and an even more forceful example is given showing what can be wrong with the criterion of an unbiased estimate. A company is laying a telephone cable across the ocean. They cannot know in advance exactly how much cable will be required, and so they must estimate. If they overestimate, the loss will presumably be proportional to the amount of excess cable to be disposed

of; but if they underestimate and the cable end falls off into the water, the result may be financial disaster. Use of an unbiased estimate here could only be described as foolhardy.

Note, however, that after all this argument, nothing in the above entitles us to conclude that γ_q is the best estimator of μ_2 by the mean-square criterion! For we have considered only the class (13-22) of estimators constructed by multiplying the sample variance (13-19) by some preassigned number; we can say only that γ_q is the best one in that class. The question whether some other function of the sample values, not a multiple of (13-19), might be still better by the mean-square error criterion, remains completely open. This weakness of the orthodox approach to parameter estimation--that it does not tell us how to find the best estimator, but only how to compare different guesses--is due to our having "looked at the problem backwards", in the sense I explained a moment ago. Now I want to show how the trouble can be overcome.

13.5. Reformulation of the Problem.

It is easy to see why the orthodox criterion of minimum α-expected loss is bound to get us into trouble and is unable to furnish any general rule for constructing an estimator. The mathematical problem was: for given $L(\alpha,\beta)$ and $(x|\alpha)$, what function $\beta(x_1 \ldots x_n)$ will minimize

$$L_\alpha = \int L(\alpha,\beta)\,(x|\alpha)\; dx \qquad (13\text{-}38)$$

Although this is not a variational problem, it might have a unique solution; but the solution will still, in general depend on α. Of course, there may be (and in fact are) a few exceptional cases where the α-dependence drops out; but in general the criterion of minimum α-expected loss leads to an impossible situation--even if we could solve the mathematical problem (13-38) and had before us the best estimator $\beta_\alpha(x_1 \ldots x_n)$ for each value of α, we could use the result only if α were already known, in which case we would have no need

to estimate. We were indeed looking at the problem backwards!

This makes it clear that in general we cannot use the criterion (13-38), or in fact any criterion which makes reference to only a single value of α; not for philosophical reasons but because any such criterion is built on self-contradictory premises. The person who advises us to use (13-38) puts himself in exactly the position of the shoe clerk who told a customer, "You will never be able to get those new boots on until you have worn them a while."

This also makes it clear how to correct the trouble. It is of no use to ask what estimator is best for some particular value of α, even though the question might have a definite answer; the only reason for using an estimator is that α is unknown. The estimator must therefore be some compromise that allows for all possibilities within some prescribed range of α; within this range it must do the best job of protecting against loss no matter what the true value of α turns out to be.

Thus it is some weighted average of L_α,

$$\langle L \rangle = \int f(\alpha)\ L_\alpha\ d\alpha \qquad (13\text{-}39)$$

that we should really minimize, where the function $f(\alpha) \geq 0$, which will be given a fuller interpretation later, measures in some way the relative importance of minimizing L_α for various possible values of α.

Merely to recognize this, which amounts to removing a contradiction in the original formulation, already implies the solution. For the mathematical character of the problem is completely changed by adopting (13-39) instead of (13-38). We now have a solvable variational problem with a well-behaved solution. The first variation in $\langle L \rangle$ due to an arbitrary variation $\delta\beta(x_1 \ldots x_n)$ in the estimator is

$$\delta\langle L \rangle = \int f(\alpha)\ d\alpha \int \ldots \int dx_1 \ldots dx_n \frac{\partial L}{\partial \beta}\ (x_1 \ldots x_n | \alpha)\ \delta\beta(x_1 \ldots x_n)$$

which vanishes independently of $\delta\beta$ if

$$\int d\alpha\ f(\alpha)\ \frac{\partial L}{\partial \beta}\ (x_1 \ldots x_n | \alpha) \ = 0 \qquad (13\text{-}40)$$

for all possible samples $\{x_1 \ldots x_n\}$. Equation (13-40) is the fundamental integral equation which determines the best estimator.

Taking the second variation, we find the condition that (13-40) shall yield a true minimum is

$$\int d\alpha\ f(\alpha)\ \frac{\partial^2 L}{\partial \beta^2}\ (x_1 \ldots x_n | \alpha) > 0 \qquad (13\text{-}41)$$

Thus a sufficient condition for a minimum is simply

$$\frac{\partial^2 L}{\partial \beta^2} \geq 0 \qquad (13\text{-}42)$$

but this is far stronger than necessary.

If we take the quadratic loss function $L(\alpha,\beta) = K(\alpha - \beta)^2$, equation (13-40) reduces to

$$\int d\alpha\ f(\alpha)\ (\alpha - \beta)\ (x_1 \ldots x_n | \alpha) = 0$$

or, the optimal estimator for quadratic loss is

$$\beta(x_1 \ldots x_n) = \frac{\int d\alpha\ f(\alpha)\ \alpha\ (x_1 \ldots x_n | \alpha)}{\int d\alpha\ f(\alpha)\ (x_1 \ldots x_n | \alpha)} \qquad (13\text{-}43)$$

But, you see, this is just the mean value over the posterior distribution of α:

$$(\alpha | x_1 \ldots x_n) = \frac{f(\alpha)\ (x_1 \ldots x_n | \alpha)}{\int d\alpha\ f(\alpha)\ (x_1 \ldots x_n | \alpha)} \qquad (13\text{-}44)$$

given by Bayes' theorem if we interpret $f(\alpha)$ as a prior probability density!

This example shows us, perhaps more clearly than any I have given so far, why the mathematical form of Bayes' theorem is always going to be the fundamental principle behind parameter estimation, independently of all philosophical arguments about the "meaning of probability", or about "random variables".

Let's see what happens for some other loss functions. If we take as a loss function the absolute error, $L(\alpha,\beta) = |\alpha - \beta|$, then the fundamental equation (13-40) becomes

$$\int_{-\infty}^{\beta} d\alpha\ f(\alpha)\ (x_1 \ldots x_n|\alpha) = \int_{\beta}^{\infty} d\alpha\ f(\alpha)\ (x_1 \ldots x_n|\alpha)$$

which states that $\beta(x_1 \ldots x_n)$ is to be taken as the median over the posterior distribution of α:

$$\int_{-\infty}^{\beta} d\alpha\ (\alpha|x_1 \ldots x_n) = \int_{\beta}^{\infty} d\alpha\ (\alpha|x_1 \ldots x_n) = \frac{1}{2} \tag{13-45}$$

Likewise, if we take a loss function $L(\alpha,\beta) = (\alpha - \beta)^4$, equation (13-40) leads to an estimator $\beta(x_1 \ldots x_n)$ which is the real root of

$$f(\beta) = \beta^3 - 3\bar{\alpha}\beta^2 + 3\overline{\alpha^2}\beta - \overline{\alpha^3} = 0 \tag{13-46}$$

where

$$\overline{\alpha^n} = \int d\alpha\ \alpha^n\ (\alpha|x_1 \ldots x_n) \tag{13-47}$$

is the n'th moment of the posterior distribution of α. [That (13-46) has only one real root is seen on forming the discriminant; the condition $f'(\beta) \geq 0$ for all real β is just $(\overline{\alpha^2} - \bar{\alpha}^2) \geq 0$.]

If we take $L(\alpha,\beta) = |\alpha - \beta|^k$, and pass to the limit $k \to 0$, or if we just take

$$L(\alpha,\beta) = \begin{cases} 0, & \alpha = \beta \\ 1, & \text{otherwise} \end{cases} \tag{13-48}$$

Eq. (13-40) tells us that we should choose $\beta(x_1 \ldots x_n)$ as the most probable value, or mode of the posterior distribution $(\alpha|x_1 \ldots x_n)$. If $f(\alpha) = \text{const.}$, this is just Fisher's maximum likelihood estimate.

In this result we finally see just what maximum likelihood accomplishes, and under what circumstances it is the optimal method to use. The maximum likelihood criterion is the one in which we care only about the chances of being exactly right; and if we are wrong, we don't care how wrong we are. This is just the situation we have in shooting at a small target, where "a miss is as good as a mile". But it is clear that there aren't very many other situations where this would be a rational way to behave; almost always, the amount of error is of some concern to us, and so maximum likelihood is not the best estimation criterion.

Note that in all these cases it was the posterior distribution $(\alpha|x_1 \ldots x_n)$ that was involved. That this will always be the case is easily seen by noting that our "fundamental integral equation" (13-40) is not so profound after all. It can equally well be written as

$$\frac{\partial}{\partial\beta}\int d\alpha\ f(\alpha)\ L(\alpha,\beta)\ (x_1 \ldots x_n|\alpha) = 0.$$

but if we interpret $f(\alpha)$ as a prior probability density, this is identical with (13-16), which we had already derived from much simpler reasoning! Likewise the condition (13-41) for a true minimum is identical with (13-17).

13.6. "Objectivity" of Decision Theory.

Decision Theory occupies a unique position in discussion of the logical foundations of statistics, because, as we have seen in (13-16) and (13-40), its procedures can be derived from either of two diametrically opposed viewpoints about the nature of probability theory, and it thus forms a kind of bridge between them. While there appears to be universal agreement as to the actual procedures that should be followed, there remains a fundamental disagreement as to the underlying reason for them, having its origin in the old issue of frequency vs. non-frequency definitions of probability.

From a pragmatic standpoint, such considerations may seem at first to be unimportant. However, in the attempt to apply decision-theory methods in real problems one learns very quickly that these questions intrude in the initial stage of setting up the problem in mathematical terms. In particular, our judgment as to the generality and range of validity of decision-theory methods depends on how these conceptual problems are resolved. My aim is to expound the viewpoint according to which these methods have the greatest possible range of application. Now we find that the main source of controversy here is on the issue of prior probabilities; on the orthodox viewpoint, if the problem involves use of Bayes' theorem then these methods are just not

applicable unless the prior probabilities are known frequencies. But to maintain this position consistently would imply an enormous restriction on the range of legitimate applications. Therefore, let's see whether the mathematical form of our final equations can shed any light on this issue.

Notice that only the product $f(\alpha)L(\alpha,\beta)$ is involved in (13-40) or (13-16); thus whether we interpret the problem as:

(A) Prior probability $f(\alpha)$, loss function $L(\alpha,\beta) = (\alpha - \beta)^2$ or as

(B) Uniform prior probability, loss function $L(\alpha,\beta) = f(\alpha)(\alpha - \beta)^2$ or as

(C) Prior probability $g(\alpha)$, loss function $f(\alpha)(\alpha - \beta)^2/g(\alpha)$, the solution is just the same. This is equally true for any loss function.

I emphasize this rather trivial mathematical property because of a curious psychological phenomenon. In expositions of decision theory written from the orthodox viewpoint, the writers are always very reluctant to introduce the notion of prior probability. They postpone it as long as possible, and finally give in only when the mathematics forces them to recognize that prior probabilities form the only basis for choice among the different admissible decisions. Even then, they are so unhappy about the use of prior probabilities that they feel it necessary always to invent a situation--often highly artificial--which makes the prior probabilities appear to be frequencies; and they will not use this theory for any problem where they don't see how to do this. But these same writers do not hesitate to pull a completely arbitrary loss function out of thin air without any basis at all, and proceed with the calculation!

The equations show that if your final decision depends strongly on which particular prior probability assignment you use, it is going to depend just as strongly on which particular loss function you use. If you worry about arbitrariness in the prior probabilities, then in order to be consistent, you ought to worry just as much about arbitrariness in the loss functions. If

you claim (as most writers on this subject have been doing for decades) that uncertainty as to the proper choice of prior probabilities invalidates the Laplace-Bayes theory, then in order to be consistent, you must also claim that uncertainty as to the proper choice of loss functions invalidates Wald's decision theory.

The reason for this strange lopsided attitude is closely connected with a certain philosophy variously called behavioristic, or positivistic, which wants us to restrict our statements and concepts to objectively verifiable things. Therefore the observable decision is the thing to emphasize, while the process of inductive reasoning and the judgment described by a prior probability must be swept under the rug. But I see no need to do this, because it seems to me obvious that rational action can come only as the result of rational thought.

If we refuse to consider the problem of rational thought merely on the grounds that it is not "objective", the result will not be that we obtain a more "objective" theory. The result will be that we have lost the possibility of getting any satisfactory theory at all, because we have denied ourselves any way of describing what is actually going on in the decision process. And, of course, the loss function is just the expression of a purely subjective value judgment, which can in no way be considered any more "objective" than the prior probabilities.

In fact, I claim that the prior probabilities are usually more objective than the loss function, both in the mathematical theory and in the everyday decision problems of "real life". In the mathematical theory we have two quite general formal principles--maximum entropy and transformation groups--that completely remove the arbitrariness of prior probabilities for a large class of important problems, which includes most of those discussed in statistical text books. Of course, these principles will not solve all problems,

and undoubtedly there are more such principles waiting to be discovered. I hope that one result of these talks will be to encourage others to seek them. To the best of my knowledge, there are as yet no general principles for determining loss functions--not even where the criterion is purely economic, because the utility of money remains ill-defined.

In "real life" decision problems, we have a similar situation. Each man knows, pretty well, what his prior probabilities are; and because his beliefs are based on all his past experience, they are not easily changed by one more experience, so they are fairly stable things. But in the heat of argument he may lose sight of his loss function; or he may never have bothered to reason out the consequences of his actions. Thus the labor mediator must deal with parties with violently opposing ideologies; policies considered noble by one party are regarded as reprehensible by the other. The successful labor mediator realizes that mere talk will not alter prior beliefs; and so his role must be to turn the attention of both parties away from this area, and explain clearly to each what his loss function is. In this sense, I think we can claim that in real life decision problems, the loss function is often far more "subjective" (in the sense of being less well fixed in our minds) than the prior probabilities.

Of course, we have to concede this much to the behaviorists--the final criterion by which we judge the soundness of any theory must be on the objective, pragmatic level. After a theory has been constructed, the ultimate test we apply to it is not whether its premises are philosophically satisfying, but how it works out in practice. Indeed, a major objective of these talks is to show you, in detail, just how the Laplace-Bayes theory does work out in practice and how its results compare with those of the orthodox methods; because that is something you very seldom find in any of the literature written from the orthodox viewpoint.

But in the process of constructing a theory, we must demand the right to invent and use any concepts we please, whether or not these concepts are themselves "objectively verifiable". If we deny ourselves this freedom on the grounds of some philosophical dogma, we are putting ourselves in a strait-jacket which effectively prevents further progress. In the case of physical theories, this point has been stressed repeatedly and strongly by Einstein; his own work is, of course, the perfect example of what can be accomplished through the free invention of new concepts.

Now let's see the extent to which varying loss functions lead to varying decisions, by some numerical examples.

13.7. Effect of Varying Loss Functions.

Suppose that on the basis of the observed sample x, a parameter α has the posterior distribution

$$(\alpha|x) = k\, e^{-k\alpha} \quad , \quad 0 \le \alpha < \infty \tag{13-49}$$

This has the n'th moment

$$\langle \alpha^n \rangle = \int_0^\infty \alpha^n \, (\alpha|x) \, d\alpha = n!\, k^{-n} \quad . \tag{13-50}$$

With loss function $(\alpha - \beta)^2$, the best estimator is the mean value

$$\beta = \langle \alpha \rangle = k^{-1} \quad . \tag{13-51}$$

With loss function $|\alpha - \beta|$, the estimator is the median, determined by

$$\frac{1}{2} = \int_0^\beta (\alpha|x) \, d\alpha = 1 - e^{-k\beta} \tag{13-52}$$

or

$$\beta = k^{-1} \ln 2 = 0.693 \, \langle \alpha \rangle . \tag{13-53}$$

To minimize $\langle (\alpha - \beta)^4 \rangle$, we should choose β to satisfy equation (13-46), which becomes, in this case,

$$y^3 - 3y^2 + 6y - 6 = 0 \tag{13-54}$$

with $y = k\beta$. The root of this is at $y = 1.59$, so the optimal estimator

with loss function $|\alpha - \beta|^4$ is

$$\beta = 1.59 \langle\alpha\rangle. \tag{13-55}$$

For the loss function $(\alpha - \beta)^{s+1}$ with s an odd integer, the fundamental equation (13-40) is

$$\int_0^\infty (\alpha - \beta)^s \, e^{-k\alpha} \, d\alpha = 0 \tag{13-56}$$

which reduces to

$$\sum_{m=0}^{s} \frac{(-k\beta)^m}{m!} = 0 \tag{13-57}$$

of which (13-54) is a special case with s = 3. In the case s = 5, loss function $(\alpha - \beta)^6$, we find

$$\beta = 2.025 \langle\alpha\rangle. \tag{13-58}$$

As $s \to \infty, \beta$ also increases without limit. But the maximum-likelihood estimate, which corresponds to the loss function

$$L(\alpha,\beta) = -\delta(\alpha - \beta)$$

or equally well to

$$\lim_{k\to 0} |\alpha - \beta|^k$$

is $\beta = 0$!

These numerical examples merely illustrate what was already clear intuitively; when the posterior distribution $(\alpha|x)$ is not sharply peaked, the best estimate of α depends very much on which particular loss function we use.

You might suppose that a loss function must always be a monotonically increasing function of the error $|\alpha - \beta|$. In general, of course, this will be the case, but there is nothing in this theory which restricts us to such functions. You can think of some rather frustrating situations in which, if you are going to make an error, you would rather make a large one than a small one. William Tell was in just that fix. If you study our equations for this case, you will see that there is really no very satisfactory decision

at all; and nothing can be done about it.

Our noting that the final decision depends only on the product of prior probability and loss function also helps to clear up a mystery which has long been puzzling to Bayesians. As we noted in Lecture 8, Jeffreys (1939) proposed that, in the case of a continuous parameter α known to be positive, we should express prior ignorance by assigning, not uniform prior probability, but a prior density proportional to $(1/\alpha)$. The theoretical justification of this rule was long unclear; but it yields very sensible-looking results in practice, which led Jeffreys to adopt it as fundamental in his significance tests. We saw in Lecture 12 that, in the case that α is a scale parameter, the Jeffreys prior is uniquely determined by invariance under the transformation group; but now we can see a still more general justification of it.

From the decision-theory viewpoint the thing that matters is not the prior or loss function separately; only their product enters into the final decision. If we use the absolute error loss function $|\beta - \alpha|$ when α is known to be positive, then to assign $f(\alpha) = \text{const.}$ in (13-45) amounts to saying that we demand an estimator which yields, as nearly as possible, a constant <u>absolute</u> accuracy for all values of α in $0 < \alpha < \infty$. That is clearly asking for too much in the case of large α; and we must pay the price in a poor estimate for small α. But we now see that the median of Jeffreys' posterior distribution is mathematically the same thing as the optimal estimator for uniform prior and loss function $|\beta - \alpha|/\alpha$; we ask for, as nearly as possible, a constant <u>percentage</u> accuracy over all values of α. This is, of course, exactly what we do want in most cases where we know that $0 < \alpha < \infty$. The reason for the superior performance of Jeffreys' rule is thus made apparent; and the mystery disappears if we re-interpret it as saying that the $(1/\alpha)$ factor is part of the loss function.

13.8. General Decision Theory.

In the foregoing, I have developed decision theory only in terms of one particular application; parameter estimation. But we really have the whole story already; the criterion (13-16) for constructing the optimal estimator generalizes immediately to the criterion for finding the optimal decision of any kind. The final rules are simply:

(1) Enumerate the possible states of nature θ_j, discrete or continuous, as the case may be.

(2) Assign prior probabilities $(\theta_j|X)$ which maximize the entropy subject to whatever prior information X you have.

(3) Digest any additional evidence E by application of Bayes' theorem, thus obtaining the posterior probabilities $(\theta_j|EX)$.

(4) Enumerate the possible decisions D_i.

(5) Specify the loss function $L(D_i,\theta_j)$ that tells what you want to accomplish.

(6) Make that decision D_i which minimizes the expected loss $\langle L\rangle_i = \sum_j L(D_i,\theta_j)(\theta_j|EX)$.

That is all there is to it; after all is said and done, the final rules of calculation to which the theorems of Cox, Wald, and Shannon lead us are just the ones which had already been developed by Bayes, Laplace, and Daniel Bernoulli in the 18'th century, except that the entropy principle generalizes the principle of indifference.

These rules either include, or improve upon, practically all known statistical methods for hypothesis testing and point estimation of parameters. If you have mastered them, then you have just about the entire field at your fingertips. The most outstanding thing about them is their simplicity--if we

sweep aside all the polemics and false starts that have cluttered up this field in the past and consider only the constructive arguments that lead directly to these rules, it is clear that the underlying rationale could be fully developed in a one-semester undergraduate course.

However, in spite of the utter simplicity of the rules themselves, really facile application of them involves intricate mathematics, and fine subtleties of concept; so much so that several generations of workers in this field misused them and concluded that the rules were all wrong! So, we still need a good deal of leading by the hand in order to develop facility in using them. It is a good deal like learning how to play a musical instrument--anybody can make noise with it, but you will not play this instrument well without years of practice.

As an example--although a rather trivial one--of the little tricks that help in applying this theory, note that the decision rule is invariant under any proper linear transformation of the loss function; i.e. if $L(D_i,\theta_j)$ is one loss function, then the new one

$$L'(D_i,\theta_j) = a + b\, L(D_i,\theta_j)$$

where $-\infty < a < \infty$, $0 < b < \infty$, will lead to the same decision, whatever the prior probabilities $(\theta_j|X)$ and new evidence E. Thus, in a binary decision problem, given the loss matrix

$$L_{ij} = \begin{pmatrix} 10 & 100 \\ 19 & 10 \end{pmatrix}$$

we can equally well use

$$L'_{ij} = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}$$

corresponding to a = -10/9, b = 1/9. This may simplify the calculation of expected loss quite a bit.

Lecture 14

DECISION THEORY IN SIGNAL DETECTION

In this Lecture, I want to examine in detail one of the simplest applications of the general decision theory just formulated. As I pointed out in Lecture 6, the problem of detection of signals in noise is really exactly the same as Laplace's old problem of detecting the presence of unknown systematic influences in celestial mechanics, and Shewhart's (1931) more recent problem of detecting a systematic drift in machine characteristics, in industrial quality control. It is unfortunate that the basic identity of all these problems hasn't been more widely recognized, because it has forced workers in several different fields to rediscover the same things, with varying degrees of success, over and over again.

As you know by now, all we really have to do to solve this problem is to take the principles developed in Lectures 3, 10, and 12; and supplement them with the loss function criterion for converting final probabilities into decisions. However, the literature of this field has been largely created from the standpoint of the original decision theory before it was realized that it was mathematically identical with the original Laplace methods; or at least before the full implications of this fact had "sunk in." The existing literature therefore uses a different sort of vocabulary and set of concepts than I have been using up to now. Since it exists, we have no choice but to learn these terms and viewpoints if we want to read the literature of the field. So, I want to give you a very rapid, condensed review of the

literature of the 1950's on these problems. My aim is to expose what is really essential, stripped of all unnecessary details. This material is also given in the papers of Middleton and van Meter (1955, 1956) and the treatise of Middleton (1960), in an enormously expanded form where a beginner can get lost for months without ever finding the real underlying principles. Just to have a complete, self-contained summary, I'll repeat a little bit from previous lectures.

14.1. Definitions and Preliminaries.

Notation:

$(A|B)$ = Conditional probability of A, given B

$(AB|CD)$ = Joint conditional probability of A and B, given D and C, . . .,

etc.

For our purposes, everything follows from the single fundamental rule of calculation, which we have called Rule 1:

$$(AB|C) = (A|BC)(B|C) = (B|AC)(A|C) \tag{14-1}$$

If the propositions B, C are not mutually contradictory, this may be rearranged to give the rule of "learning by experience," Bayes' theorem:

$$(A|BC) = (A|C)\frac{(B|AC)}{(B|C)} = (A|B)\frac{(C|AB)}{(C|B)} \quad . \tag{14-2}$$

If there are several mutually exclusive and exhaustive propositions B_i, then by summing (14-1) over them, we obtain the chain rule

$$(A|C) = \sum_i (A|B_iC)(B_i|C) \tag{14-3a}$$

or, for a simpler notation,

$$(A|C) = \sum_B (A|BC)(B|C) \tag{14-3b}$$

Now let

X = prior knowledge, of any kind whatsoever

S = signal

N = noise

V = V(S,N) = observed voltage

D = decision about the nature of the signal

Any probabilities conditional on X alone are called prior probabilities. Thus we have

$(S|X)$ = prior probability of the particular signal S

$(N|X) = W(N)$ = prior probability of the particular sample of noise N.

In a linear system, $V = S + N$, and

$$(V|S) = (V|SX) = W(V - S) \ . \qquad (14\text{-}4)$$

You may be disturbed by the absence of density functions, dS's, dN's, etc., which might be expected in the case of continuous S, N. Note, however, that our equations are homogeneous in these quantities, so they cancel out anyway. By $\sum_A$ I mean ordinary summation over some previously agreed set of possible values if A is discrete, integration with appropriate density functions if A is continuous.

A <u>decision rule</u> $(D_i|V_j)$, or for brevity just $(D|V)$, represents the process of drawing inferences about the signal from the observed voltage. If it is always made in a definite way, then $(D|V)$ has only the values 0, 1 for any choice of D and V; however we may also have a "randomized" decision rule according to which $(D|V)$ is a true probability distribution. Maintaining this more general view turns out to be a help in formulating the theory.

The essence of any decision rule, and in particular, any one which can be built into automatic equipment, is that the decision must be made on the basis of V alone; V is, by definition, the quantity which contains all the information actually used (in addition to the ever-present X) in arriving at the decision. Thus, if $Y \neq D$ is any other proposition, we have

$$(D|V) = (D|VY) \ . \qquad (14\text{-}5)$$

An equivalent statement is that D depends on any proposition Y only through the intermediate influence of Y on V:

$$(D|Y) = \sum_V (D|V)(V|Y) \qquad (14\text{-}6)$$

14.2. Sufficiency and Information.

Equation (14-5) has interesting consequences; suppose we wish to judge the plausibility of some proposition Y, on the basis of knowledge of V and D. From (14-1),

$$(DY|V) = (Y|VD)(D|V) = (D|VY)(Y|V)$$

and using (14-5), this reduces to

$$(Y|VD) = (Y|V) \qquad (14\text{-}7)$$

Thus, if V is known, knowledge of D is redundant and cannot help us in estimating any other quantity. The reverse is not true, however; we could equally well use (17-1) in another way:

$$(VY|D) = (Y|VD)(V|D) = (Y|D)(V|YD).$$

Combining this with (14-7), there results the

Theorem: Let D be a possible decision, given V. Then $(V|D) \neq 0$, and

$$(Y|V) = (Y|D) \quad \text{if and only if} \quad (V|D) = (V|YD) \qquad (14\text{-}8)$$

In words: knowledge of D is as good as knowledge of V for judgments about Y if and only if Y is irrelevant for judgments about V, given D. Stated differently: in the "environment" produced by knowledge of D, the propositions Y and V appear to be independent, i.e.

$$(YV|D) = (Y|D)(V|D) \qquad (14\text{-}9)$$

In this case, D is said to be a sufficient statistic for judgments about Y. In the next lecture, we will study the notion of sufficiency from a different point of view. Evidently, a decision rule which makes D a sufficient statistic for judgments about the signal S is in some sense superior to one without this property. However, such a rule does not necessarily exist. Equation

(14-9) is a very restrictive condition, since it must be satisfied for all values of Y, V, and all D for which $(D|V) \neq 0$.

As you might guess from this, the concept of sufficiency is closely related to that of information. The definition of sufficiency could equally well be stated as: D is a sufficient statistic for judgments about Y if it contains all the information about Y which V contains. Since D is determined from V, if it is not a sufficient statistic, it necessarily contains less information about Y than does V. In this statement, the term "information" was used in a loose, intuitive sense; does it remain true if we adopt Shannon's measure of information? Imagine that there are several mutually exclusive propositions Y_i, one of which must be true. For brevity we use, as above, the notation $\sum_Y f(Y) \equiv \sum_i f(Y_i)$. Then the entropy of Y with a specific value of D given is

$$H_D(Y) = -\sum_Y (Y|D) \log (Y|D) \tag{14-10}$$

and its average over all values of D is

$$\bar{H}_D(Y) = \sum_D (D|X) H_D(Y) \tag{14-11}$$

If

$$\bar{H}_C(Y) < \bar{H}_D(Y)$$

we say that C contains, on the average, more information about Y than does D. Note, however, that it may be otherwise for specific values of C and D.

Acquisition of new information can never increase $\bar{H}$; let D, V, Y be, for the moment, any three quantities and form the expression

$$\begin{aligned}\bar{H}_V(Y) - \bar{H}_{DV}(Y) &= \sum_{DVY} (DV|X)(Y|DV) \log (Y|DV) \\ &\quad - \sum_{VY} (V|X)(Y|V) \log (Y|V) \\ &= \sum_{DVY} (DV|X)(Y|DV) \log [(Y|DV)/(Y|V)]\end{aligned}$$

Using the by now familiar fact that $\log x \geq (1 - x^{-1})$, with equality if and only if $x = 1$, this becomes

$$\overline{H}_V(Y) - \overline{H}_{DV}(Y) \geq \sum_{DVY} (DV|X)[(Y|DV) - (Y|V)] = 0 \qquad (14\text{-}13)$$

Thus, $\overline{H}_{DV}(Y) \leq \overline{H}_V(Y)$, with equality if and only if Eq. (14-7) holds for all D, V, and Y for which $(DV|X) \neq 0$. Since (14-13) holds regardless of the meaning of D and V, we can equally well conclude that for all D, V, Y,

$$\overline{H}_D(Y) \geq \overline{H}_{DV}(Y) \leq \overline{H}_V(Y) \ .$$

Now letting D, V, Y resume their original meanings, we have in consequence of (14-7) $H_V(Y) = H_{DV}(Y)$, so that

$$\overline{H}_V(Y) \leq \overline{H}_D(Y) \qquad (14\text{-}14)$$

with equality if and only if Eq. (14-9) holds, i.e. if and only if D is a sufficient statistic. Thus, if by "information" we mean minus the average entropy of Y over the prior distribution of D or V, zero information loss in going from V to D is equivalent to sufficiency of D. Note that inequalities of the form (14-13) hold only for the averages $\overline{H}$, not for the H. Acquisition of a specific piece of information (that an event previously considered improbable had in fact occurred) may in some cases increase the entropy of Y. However, this is an improbable situation and on the average the entropy can only be lowered by additional information. This shows again that the term "information" is not a happy choice of word to describe entropy expressions. In spite of the entropy increase, the situation just described could hardly be called one of less information, but rather one of less certainty.

14.3. Loss Functions and Criteria of Optimum Performance.

In order to say that one decision rule is better than another, we need some specific criterion of what we want our detection system to accomplish. The criterion will vary with the application, and obviously no single decision

rule can be best for all purposes. A very general type of criterion is obtained by assigning a loss function $L(D,S)$ which represents our judgment of how serious it is to make decision D when signal S is in fact present. In case there are only two possible signals; $S_o = 0$ (i.e. no signal), and $S_1 \neq 0$, and consequently two possible decisions D_o, D_1, there are two types of error, the false alarm $A = (D_1,S_o)$ and the false rest $R = (D_o,S_1)$. In some applications, one type of error might be much more serious than the other. Suppose that a false rest is considered ten times as serious as is a false alarm, while a correct decision of either type represents no "loss." We could then take $L(D_o,S_o) = L(D_1,S_1) = 0$, $L(D_o,S_1) = 10$, $L(D_1,S_o) = 1$. Whenever the possible signals and the possible decisions form discrete sets, the loss function becomes a loss matrix. In the above example,

$$L_{ij} = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}$$

The loss matrix plays approximately the same role in detection theory as does the payoff matrix in game theory. A player in a game may choose that strategy which maximizes his expected gain, and correspondingly we may choose that decision rule $(D|V)$ which minimizes the expected loss.

Instead of assigning arbitrarily a certain loss value to each possible type of detection error, we may consider information loss by the assignment $L(D,S) = -\log (S|D)$. This is somewhat more difficult to manipulate, because now $L(D,S)$ depends on the decision rule. A decision rule which minimizes information loss is one which makes the decision in some sense as close as possible to being a sufficient statistic for judgments about the signal. In exactly what sense seems never to have been clarified.

The conditional loss $L(S)$ is the average loss incurred when the specific signal S is present

$$L(S) = \sum_D L(D,S)(D|S) \qquad (14\text{-}15)$$

which may in turn be expressed in terms of the decision rule and the properties of the noise by using (14-6). The average loss is the expected value of this over all possible signals:

$$\langle L \rangle = \sum_S L(S)(S|X) \qquad (14\text{-}16)$$

Two different criteria of optimum performance now suggest themselves:

The Minimax Criterion. For a given decision rule $(D|V)$, consider the conditional loss $L(S)$ for all possible signals, and let $[L(S)]_{max}$ be the maximum value attained by $L(S)$. We seek that decision rule for which $[L(S)]_{max}$ is as small as possible. As we noted in the last lecture, this criterion concentrates attention on the worst possible case regardless of the probability of occurrence of this case, and it is thus in a sense the most conservative one. If the worst possible case is extremely unlikely to arise, one would call it too conservative. It has, however, the practical advantage that it does not involve the prior probabilities of the different signals, $(S|X)$, and therefore it can be applied in cases where the available information about the signal is of such an indefinite type that we do not know what prior probabilities to assign.

The Bayes Criterion. We seek that decision rule for which the average loss $\langle L \rangle$ is minimized. In order to apply this, a prior distribution $(S|X)$ must be available.

Other criteria were proposed before the days of Decision Theory. In the Neyman-Pearson criterion, we fix the probability of occurrence of one type of error at some small value δ, and then minimize the probability of another type of error subject to this constraint. Siegert's "Ideal Observer" minimizes the total probability of error regardless of type. However, we will see below that these are both special cases of the Bayes criterion, for particular loss functions $L(D,S)$. The minimax criterion may also be considered a special

case of the Bayes, in which we choose the worst possible $(S|X)$, after having found the decision rule which minimizes $\langle L \rangle$ for a given $(S|X)$. The basic identity of all these criteria came as quite a surprise to the early workers in this field.

Substituting in succession equations (14-15), (14-6), and (14-3) into (14-16), we obtain for the average loss

$$\langle L \rangle = \sum_{DV} \left[\sum_{S} L(D,S)(VS|X) \right] (D|V) \qquad (14\text{-}17)$$

If $L(D,S)$ is a definite function independent of $(D|V)$ (this assumption excludes for the moment the information loss function), there is no function $(D|V)$ for which this expression is stationary in the sense of calculus of variations. We then minimize $\langle L \rangle$ merely by choosing for each possible V that decision $D_1(V)$ for which

$$K(D_1,V) \equiv \sum_{S} L(D_1,S)(VS|X) \qquad (14\text{-}18)$$

is a minimum. Thus, we adopt the decision rule

$$(D|V) = \delta(D,D_1). \qquad (14\text{-}19)$$

In general there will be only one such D_1, and the best decision rule is nonrandom. However, in case of "degeneracy," $K(D_1,V) = K(D_2,V)$, any randomized rule of the form

$$(D|V) = a\,\delta(D,D_1) + b\,\delta(D,D_2) \quad , \quad a + b = 1 \qquad (14\text{-}20)$$

is just as good. This degeneracy occurs at "threshold" values of V, where we change from one decision to another.

14.4. A Discrete Example.

Consider the case already mentioned, where there are two possible signals S_0, S_1, and a loss matrix

$$L_{ij} = \begin{pmatrix} L_{00} & L_{01} \\ L_{10} & L_{11} \end{pmatrix} = \begin{pmatrix} 0 & L_r \\ L_a & 0 \end{pmatrix}$$

where L_a, L_r are the losses incurred by a false alarm and a false rest, respectively. Then

$$\begin{aligned} K(D_0,V) &= L_{01}(VS_1|X) = L_r(VS_1|X) \\ K(D_1,V) &= L_{10}(VS_0|X) = L_a(VS_0|X) \end{aligned} \tag{14-21}$$

and the decision rule that minimizes <L> is

$$\left.\begin{aligned} &\text{Choose } D_1 \text{ if } \frac{(VS_1|X)}{(VS_0|X)} > \frac{L_a}{L_r} \\ &\text{Choose } D_0 \text{ if } \frac{(VS_1|X)}{(VS_0|X)} < \frac{L_a}{L_r} \\ &\text{Choose either at random in case of equality.} \end{aligned}\right\} \tag{14-22}$$

In words: if the prior probability that the observed voltage is due to the signal exceeds the probability that it is due to noise alone by a factor greater than the ratio of false alarm loss to false rest loss, we decide that the signal is present. If the prior probabilities of signal and no signal are

$$(S_1|X) = p, \qquad (S_0|X) = q = 1 - p \tag{14-23}$$

respectively, we have $(VS_1|X) = (V|S_1)(S_1|X) = p(V|S_1)$, etc., and the decision rule becomes

$$\text{Choose } D_1 \text{ if } \frac{(V|S_1)}{(V|S_0)} > \frac{qL_a}{pL_r}, \text{ etc.} \tag{14-24}$$

The left-hand side of (14-24) is called a likelihood ratio. It depends only on the statistical properties of the noise, and is the quantity which should be computed by the optimum receiver according to the Bayes criterion. The same quantity is the essential one regardless of the assumed loss function and regardless of the probability of occurrence of the signal; these affect only the threshold of detection. Furthermore, if the receiver merely computes this likelihood ratio and delivers it at the output without making any decision, it provides us with all the information we need to make optimum decisions

in the Bayes sense. Note particularly the generality of this result, which is one of the most important ones for our applications; no assumptions are needed as to the type of signal, linearity of the system, or statistical properties of the noise.

We now work out, for purposes of illustration, the decision rules and their degree of reliability, for several of the above criteria, in the simplest possible problem that I mentioned back in Lecture 4, to illustrate the principle of maximum likelihood. We have a linear system in which the voltage is observed at a single instant, and we are to decide whether a signal, which can have only amplitude S_1, is present in noise, which is gaussian with mean square value $\langle N^2 \rangle$:

$$W(N) = \frac{1}{\sqrt{2\pi\langle N^2\rangle}} \exp\left[- \frac{N^2}{2\langle N^2\rangle} \right] \tag{14-25}$$

The likelihood ratio in (14-24) then becomes

$$\frac{(V|S_1)}{(V|S_0)} = \frac{W(V-S_1)}{W(V)} = \exp\left[\frac{2VS_1 - S_1^2}{2\langle N^2\rangle} \right] \tag{14-26}$$

and since this is a monotonic function of V, the decision rule can be written as

$$\text{choose} \quad \begin{Bmatrix} D_1 \\ D_0 \end{Bmatrix} \quad \text{when } V \begin{Bmatrix} > \\ < \end{Bmatrix} V_b \tag{14-27}$$

with

$$\frac{V_b}{\sqrt{\langle N^2\rangle}} = \frac{1}{2s} \left[2 \log\left(\frac{qL_a}{pL_r} \right) + s^2 \right] = v_b \tag{14-28}$$

in which

$s \equiv \dfrac{S_1}{\sqrt{\langle N^2\rangle}}$ is the voltage signal-to-noise ratio, and

$v \equiv \dfrac{V}{\sqrt{\langle N^2\rangle}}$ is the normalized voltage.

Now we find for the probability of a false rest:

$$(R|X) = (D_0 S_1|X) = p \sum_V (D_0|V)(V|S_1) = p \int_{-\infty}^{V_b} dV\ W(V-S_1)$$

$$= p\ \Phi(v_b - s) \qquad (14\text{-}29)$$

and for a false alarm,

$$(A|X) = (D_1 S_0|X) = q \sum_V (D_1|V)(V|S_0) = q \int_{V_b}^{\infty} dV\ W(V)$$

$$= q[1 - \Phi(v_b)] \quad . \qquad (14\text{-}30)$$

Here $\Phi(x)$ is the cumulative normal distribution

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\ dt \qquad (14\text{-}31)$$

numerical values of which are given in most mathematical tables. For $x > 2$, a good approximation is

$$1 - \Phi(x) \simeq \frac{e^{-x^2/2}}{x\sqrt{2\pi}} \quad . \qquad (14\text{-}32)$$

As a numerical example, if $L_r = 10\ L_a$, $q = 10\ p$, these expressions reduce to

$$(A|X) = 10\ (R|X) = \frac{10}{11}\ [1 - \Phi(\tfrac{1}{2}\ s)] \qquad (14\text{-}33)$$

The probability of a false alarm is less than 0.027, and of a false rest less than 0.0027 for $s > 4$. For $s > 6$, these numbers become 1.48×10^{-3}, 1.48×10^{-4} respectively.

Let us see what the minimax criterion would give in this problem. The conditional losses are

$$L(S_0) = L_a \sum_V (D_1|V)(V|S_0) = L_a \int_{-\infty}^{\infty} (D_1|V)\ W(V)\ dV$$

$$L(S_1) = L_r \sum_V (D_0|V)(V|S_1) = L_r \int_{-\infty}^{\infty} (D_0|V)\ W(V-S_1)\ dV \qquad (14\text{-}34)$$

Writing $f(V) \equiv (D_1|V) = 1 - (D_0|V)$, the only restriction on $f(V)$ is $0 \le f(V) \le 1$. Since L_a, L_r, and $W(V)$ are all positive, a change $\delta f(V)$ in the neighborhood of any given point V will always increase one of the quantities (14-34) and decrease the other. Thus when the maximum $L(S)$ has been

made as small as possible, we will certainly have $L(S_0) = L(S_1)$, and the problem is thus to minimize $L(S_0)$ subject to this constraint. Suppose that for some particular $(S|X)$ the Bayes decision rule happened to give $L(S_0) = L(S_1)$. Then this particular solution must be identical with the minimax solution, for with the above constraint, $\langle L \rangle = [L(S)]_{max}$, and if the Bayes solution minimizes $\langle L \rangle$ with respect to all admissible variations $\delta f(V)$ in the decision rule, it a fortiori minimizes it with respect to the smaller class of variations which keep $L(S_0) = L(S_1)$. Therefore our optimum decision rule will have the same form as before: There is some threshold V_m such that

$$f(V) = \begin{cases} 0, & V < V_m \\ 1, & V > V_m \end{cases} \tag{14-36}$$

Any change in V_m from the value which makes $L(S_0) = L(S_1)$ necessarily increases one or the other of these quantities. The equation determining V_m is therefore

$$L_a \int_{V_m}^{\infty} W(V)\ dV = L_r \int_{-\infty}^{V_m} W(V-S_1)\ dV$$

or, in terms of normalized quantities,

$$L_a[1 - \Phi(v_m)] = L_r\ \Phi(v_m - s) \tag{14-37}$$

Note that (14-30), (14-31) give the conditional probabilities of false rest and false alarm for any decision rule of type (14-36), regardless of whether the threshold was determined from (14-28) or not; for the arbitrary threshold V_0

$$\begin{aligned} (R|S_1) &= (V < V_0|S_1) = \Phi(v_0 - s) \\ (A|S_0) &= (V > V_0|S_0) = \frac{1}{2}\,[1 - \Phi(v_0)] \end{aligned} \tag{14-38}$$

From (14-28) we see that there is always a particular ratio (p/q) which makes the Bayes threshold V_b equal to the minimax threshold V_m. For values of (p/q) other than this worst value, the Bayes criterion gives a lower average loss than does the minimax, although one of the conditional losses $L(S_0)$,

$L(S_1)$ will be greater than the minimax value.

These relations and several previous remarks are illustrated in Figure (14.1), in which we plot the conditional losses $L(S_0)$, $L(S_1)$ and the average loss $\langle L \rangle$ as functions of the threshold V_0, for the case $L_a = \frac{3}{2} L_r$, $p = q = \frac{1}{2}$. The minimax threshold is at the common crossing-point of these curves, while the Bayes threshold occurs at the lowest point of the $\langle L \rangle$ curve. One sees how the Bayes threshold moves as the ratio (p/q) is varied, and in particular that the value of (p/q) which makes $V_b = V_m$ also leads to the maximum values of the $\langle L \rangle_{min}$ obtained by the Bayes criterion. Thus we could also define a "maximin" criterion; first find the Bayes decision rule which gives minimum $\langle L \rangle$ for a given $(S|X)$, then vary the prior probabilities $(S|X)$ until the maximum value of $\langle L \rangle_{min}$ is attained. This is the worst possible (in the Bayes sense) prior probability, and the decision rule thus obtained is identical with the one resulting from the minimax criterion.

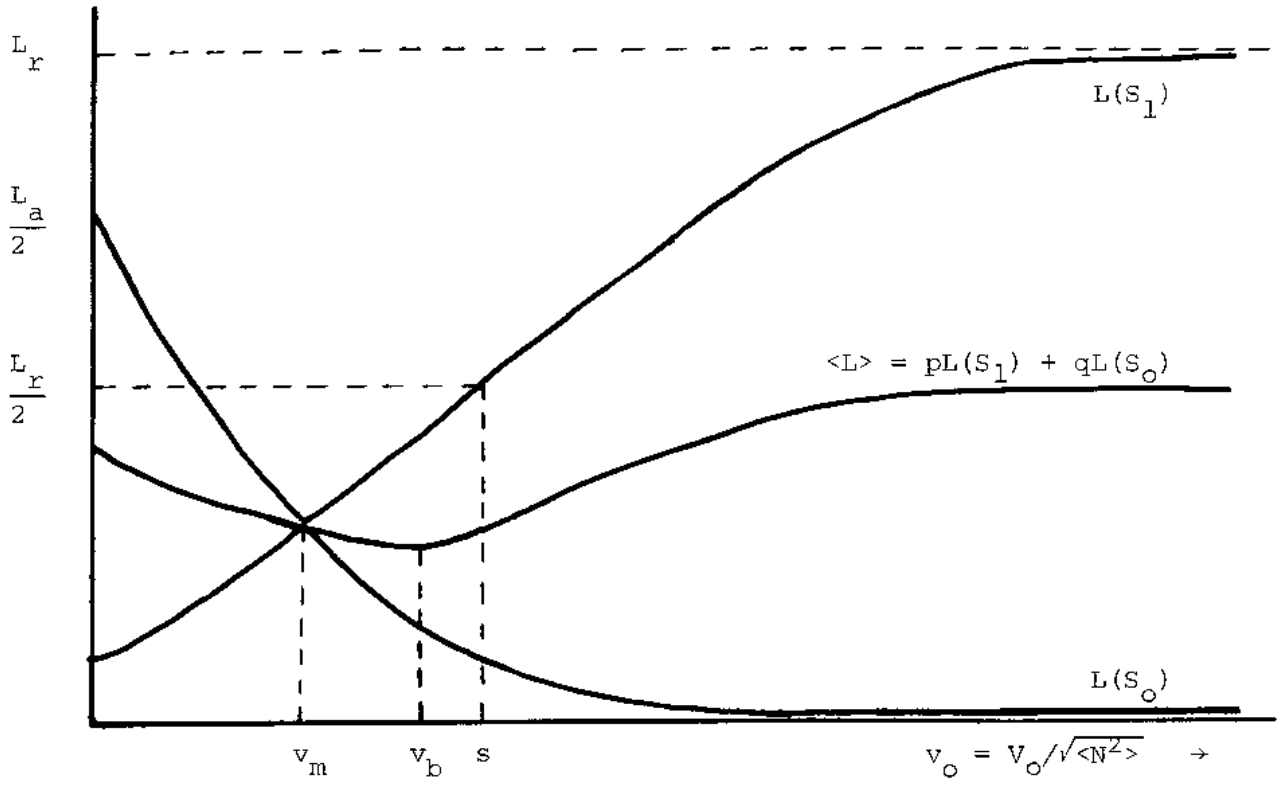

Figure 14.1. Conditional and Average Losses as functions of the detection threshold V_o. The $L(S_1)$ curve is symmetric about the point $\{s, L_r/2\}$.

The Neyman-Pearson criterion is easily discussed in this example: Suppose the conditional probability of a false alarm $(D_1|S_0)$ is held fixed at some small value ε, and we wish to minimize the conditional probability $(D_0|S_1)$ of a false rest, subject to this constraint. Now the Bayes criterion minimizes the average loss

$$\langle L \rangle = pL_r(D_0|S_1) + qL_a(D_1|S_0)$$

with respect to any admissible variation $\delta(D|V)$ in the decision rule. In particular, therefore, it minimizes it with respect to the smaller class of variations which hold $(D_1|S_0)$ constant at the value finally obtained. Thus it minimizes $(D_0|S_1)$ with respect to these variations and solves the Neyman-Pearson problem; we need only choose the particular value of the ratio (qL_a/pL_r) which results in the assumed value of ε according to equations (14-28), (14-30).

We find for the Neyman-Pearson threshold, from (14-38)

$$\Phi(v_{np}) = 1 - \varepsilon \tag{14-39}$$

and the conditional probability of detection is

$$(D_1|S_1) = 1 - (D_0|S_1) = \Phi(s - v_{np}) \tag{14-40}$$

This is the cumulative normal distribution, plotted in Appendix . First finding from the graph, v_{np} for given ε, we find that if $\varepsilon = 10^{-3}$, a detection probability of 99 per cent or better is attained for $s > 6$.

It is important to note that these numerical examples depend critically on our assumption of gaussian noise. If the noise is not gaussian, the actual situation may be either more or less favorable than indicated by the above relations. It is well known that in one sense gaussian noise is the worse possible kind; because of its maximum entropy properties, gaussian noise can obscure a weak signal more completely than can any other noise of the same average power. On the other hand, gaussian noise is a very favorable kind from which to extract a fairly strong signal, because the probability that

the noise will exceed a few times the RMS value $\sqrt{\langle N^2 \rangle}$ becomes vanishingly small. Consequently, the probability of making an incorrect decision on the presence or absence of a signal goes to zero very rapidly as the signal strength is increased. The high reliability of operation found above for $s > 6$ would not be found for noise possessing a probability distribution with wider "tails".

The type of noise distribution to be expected in any particular case depends, of course, on the physical mechanism which gives rise to the noise. When the noise is the resultant of a large number of small, independent effects, the central limit theorem of probability theory tells us that the gaussian distribution will be our best bet regardless of the nature of the individual sources.

Well, as the BBC announcers say, that is the end of my summary. All of these apparently different criteria lead, when worked out, to a probability ratio test. In the case of a binary decision, it took the simple form (14-22). Of course, any decision process can be broken down into successive binary decisions, so this case really has the whole story in it. All the different criteria amounted, in the final analysis, only to different philosophies about how you choose the threshold value at which you change your decision.

14.5. How Would Our Robot Do It?

Now let's see how this problem appears from the viewpoint of our robot. The rather long arguments we had to go through above (and even they are very highly condensed, I assure you!) to get the result are due only to the orthodox view which insists on looking at the problem backwards, i.e. on concentrating attention on the final decision rather than on the inductive reasoning process which logically has to precede it. To the robot, if our job is to make the best possible decision as to whether the signal is present, the obvious first thing we must do is calculate the probability that the signal is present.

If there are to be only two possibilities, S_0, S_1, taken into account, then after we have seen voltage V, the odds are from (5-5)

$$O(S_1|VX) = O(S_1|X)\,\frac{(V|S_1)}{(V|S_0)} \tag{14-41}$$

If we give the robot the loss function (14-21) and ask him to make the decision which minimizes the expected loss, he will evidently use the decision rule

$$\text{choose } D_1 \text{ if } \quad O(S_1|V) \equiv \frac{(S_1|V)}{(S_0|V)} > \frac{L_a}{L_r} \tag{14-42}$$

etc. But from Rule 1, $(VS_1|X) = (S_1|V)(V|X)$, $(VS_0|X) = (S_0|V)(V|X)$, and (14-42) is identical with (14-22). So, just from looking at this problem the other way around, our robot derives the same final result in exactly two lines!

You see that all this discussion of strategies, admissibility, conditional losses, etc., was completely unnecessary. Except for the introduction of the loss function at the end, there's nothing in decision theory that isn't already contained in basic probability theory, if we can only free ourselves from the dogma that "probability statements can be made only about random variables," and use the theory in the full generality given to it by Laplace.

This comparison shows why the development of decision theory has, more than any other single factor, led to this revolution in statistical thought. For about thirty years, Jeffreys tried valiantly to explain the Laplace methods to statisticians, and his efforts met only with a steady torrent of denials and ridicule. The quotation about Bayes' theorem applied to quality-control testing that I gave you back in Lecture 5 is a relatively mild example; if you have a taste for such things, you can find, particularly in the works of Fisher and von Mises, some attacks on the viewpoint of Laplace and Jeffreys which make my polemics seem rather tame. It is really astonishing how much emotional fervor can be generated by something that outsiders might consider

a rather dry and dull branch of mathematics.

It is real poetic justice that the work of one of the most respected of the "orthodox" statisticians, which was hailed, very properly, as perhaps the greatest advance in statistical practice yet produced, turned out to give, after very long and complicated arguments, exactly the same final results that the despised Laplace methods give you immediately. The only proper conclusion, it seems to me, is that the supposed distinction between statistical inference and probability theory was entirely artificial--a tragic error of judgment which has wasted perhaps a thousand man-years of our best mathematical talent in the pursuit of false goals. There is no longer any justification for trying to make this non-existent distinction.

Suppose that, in the above case of a linear system with gaussian noise, we apply Bayes' theorem in the logarithmic form of Lecture 5. If now we let S_0 and S_1 stand for numerical values giving the amplitudes of the two possible signals, the evidence for the signal is increased by

$$\log \frac{(V|S_1)}{(V|S_0)} = \frac{(V - S_0)^2 - (V - S_1)^2}{2\langle N^2 \rangle}$$

$$= \text{const.} + \frac{S_1 - S_0}{\langle N^2 \rangle} V \qquad (14\text{-}43)$$

so, the observed voltage is just a linear function of the number of db evidence for S_1.

A funny thing happened in the history of this subject. You know that electrical engineers started out not knowing anything whatsoever about statistics. They knew about signal to noise ratios. Receiver input circuits were designed for many years on the basis that signal to noise ratio was maximized. More specifically, it turned out that if you take the ratio of (peak signal)^2 to mean square noise, and find the design of input stages of the receiver which will maximize this quantity, this turned out to be a very useful thing. This

leads to the solution which is now called the classical matched filter. It has been discovered independently by at least a dozen people. I believe the first person to work out this matched filter theory was the late Professor W. W. Hansen, in about 1941. I was working with him, beginning in 1942, on problems of radar detection. He circulated a little memorandum at the time in which he gave this solution for the design of the optimum response curve of an IF strip. Years later I was thinking about an entirely different problem (an optimum antenna pattern), and when I finally got the solution, I recognized it as exactly the same thing that Bill Hansen had worked out many years before. I'll give you this theory in a later lecture. Since then I see, almost every time I open a journal concerned with these problems, that somebody else has a paper with the same solution in it.

Now, in the 1950's, people got more sophisticated about the way they handled their detection problems, and they started using this wonderful new tool, statistical decision theory, to see if there were still better ways of handling these design problems. The strange thing happened that in the case of a linear system with gaussian noise, the optimum solution which decision theory leads you to, turns out to be exactly the same old classical matched filter. When I first saw this, I was very surprised that two approaches so entirely different should lead to the same solution. But, note that our robot represents a viewpoint from which it is not at all surprising that the two lines of argument would have to give the same result. The best statistical analysis you can make of the problem will always be one in which you calculate the probability that the various signals are present by means of Bayes' theorem. But, in the case of a linear system with gaussian noise, the observed voltage is itself just a linear function of the posterior probability measured in db. So, they are essentially just two different ways of formulating the same problem.

The different approaches to the theory simply amount to different philosophies of how you choose that value of probability at which you will change your decision. Because of the fact that they all lead to the same probability ratio test, they must necessarily all be derivable from Bayes' theorem, in agreement with out robot's prediction back in Lecture 4.

The problem just examined by several different decision criteria is, of course, the simplest possible one. In a more realistic problem we will observe the voltage v(t) as a function of time, perhaps several voltages $v_1(t)$, $v_2(t)$, ... in several different channels. We may have many different possible signals $S_a(t)$, $S_b(t)$ ... to distinguish, or we may need not only to decide whether a given signal is present, but also to make the best estimates of one or more signal parameters (such as intensity, starting time, frequency, phase, rate of frequency modulation, etc.). Therefore, just as in the problem of quality control discussed in Lectures 5, 6, the details can become arbitrarily complicated. But these extensions are, from the Bayesian viewpoint, straightforward in that they require no new principles beyond those already given.

I want to come back to some of these more complicated problems of detection and filtering toward the end of these lectures; but for now let's look at another elementary kind of decision problem. In the ones discussed so far, we used Bayes' theorem, but not maximum entropy. Now I want to show you a kind of problem where we need maximum entropy. but not Bayes' theorem.

14.6. The Widget Problem.

This problem was first propounded at a symposium held at Purdue University in November, 1960--at which time, however, the full solution was not known. This was worked out later (Jaynes, 1963c), and some numerical approximations were improved in the computer work of Tribus and Fitts (1968).

The widget problem has proved to be interesting in more respects than originally realized. It is a decision problem in which there is no occasion to use Bayes' theorem, because no "new" information is acquired. Thus it would be termed a "no data" decision problem in the sense of Chernoff and Moses (1959). However, at successive stages of the problem we have more and more prior information; and digesting it by maximum entropy leads to a sequence of prior probability assignments, which lead to different decisions. Thus it is an example of the "pure" use of maximum entropy, as in statistical mechanics. It is hard to see how the problem could be formulated mathematically at all without use of maximum entropy, or some other device [like the one considered in Lecture 10 (Sec. 10.8)] which turns out in the end to be mathematically equivalent to maximum entropy.

The problem is interesting also in that we can see a continuous gradation from decision problems so simple that common sense tells us the answer instantly with no need for any mathematical theory, through problems more and more involved so that common sense has more and more difficulty in making a decision, until finally we reach a point where nobody has yet claimed to be able to see the right decision intuitively, and we require the mathematics to tell us what to do.

Finally, it turned out to be very close to an important real problem faced by oil prospectors. The details of the real problem are shrouded in proprietary caution; but I'm not giving away any secrets if I tell you that, a few years ago, I spent a week at the research laboratories of one of our large oil companies, lecturing for over 20 hours on the widget problem. They made me go through every part of the calculation in excruciating detail--much more than we have time for here--with a room full of engineers armed with slide-rules, checking up on every stage of the numerical work. I've often wondered since how far they have extended the theory beyond the original

problem, and how much it helped them; but I don't expect to find out.

Well, here is the problem. Mr. A is in charge of a Widget factory, which proudly advertises that it can make delivery in 24 hours on any size order. This, of course, is not really true, and Mr. A's job is to protect, as best he can, the Advertising Manager's reputation for veracity. This means that each morning he must decide whether the day's run of 200 Widgets will be painted red, yellow, or green. (For complex technological reasons, not relevant to the present problem, only one color can be produced per day.) We follow his problem of decision through several stages of increasing knowledge.

Stage 1. When he arrives at work, Mr. A checks with the stock room and finds that they now have in stock 100 red widgets, 150 yellow, and 50 green. His ignorance lies in the fact that he does not know how many orders for each type will come in during the day. Clearly, in this state of ignorance, Mr. A will attach the highest significance to any tiny scrap of information about orders likely to come in today; and if no such scraps are to be had, we do not envy Mr. A his job. Still, if a decision has to be made on no more information than this, his common sense will probably tell him that he had better build up that stock of green widgets.

Stage 2. Mr. A, feeling the need for more information, calls up the front office and asks, "Can you give me some idea of how many orders for red, yellow, and green widgets are likely to come in today?" They reply, "Well, we don't have the breakdown of what has been happening each day, and it would take us a week to compile that information from our files. But we do have a summary of total sales last year. Over the last year, we sold a total of 13,000 red, 26,000 yellow, and 2600 green. Figuring 260 working days, this means that last year we sold an average of 50 red, 100 yellow, 10 green each day." If Mr. A ponders this new information for a few seconds, I think he

will change his mind, and decide to make yellow ones today.

Stage 3. The man in the front office calls Mr. A back to say, "It just occurred to me that we do have a little more information that might possibly help you. We have at hand not only the total number of widgets sold last year, but also the total number of orders we processed. Last year we got a total of 173 orders for red, 2600 for yellow, and 130 for green. This means that customers who use red widgets ordered, on the average, 13000/173 = 75 widgets per order, while the average orders for yellow and green were 26000/2600 = 10, and 2600/130 = 20 respectively." This new data doesn't change the expected daily demand; but if Mr. A is very shrewd and ponders it very hard, I think he may change his mind again, and decide to make red ones today.

Stage 4. Mr. A is just about to give the order to make red ones when the front office calls him again to say, "We just got word that a messenger is on his way here with an emergency order for 40 green widgets." Now, what should he do? Up to this point, Mr. A's decision problem has been simple enough so that reasonably good common sense will tell him what to do. But now, I think he is in trouble; qualitative common sense is just not powerful enough to solve his problem, and he needs a mathematical theory to determine a definite optimum decision.

Let's summarize all the above data in a table:

	R	Y	G	Decision
1. In Stock	100	150	50	G
2. Avg. Daily Order Total	50	100	10	Y
3. Avg. Individual Order	75	10	20	R
4. Specific Order			40	?

Table 14.1. Summary of four stages of the Widget Problem.

In the last column I give the decisions that seemed, to me, to be the best ones before I had worked out the mathematics. Do other people agree with this intuitive judgment? Professor Myron Tribus has put this to the test by giving talks about this problem, and taking votes from the audience before the solution is given. Let me quote his findings as given in their paper (M. Tribus and G. Fitts, 1968). They use D_1, D_2, D_3, D_4 to stand for the optimum decisions in stages 1, 2, 3, 4 respectively:

"Before taking up the formal solution, it may be reported that Jaynes' widget problem has been presented to many gatherings of engineers who have been asked to vote on D_1, D_2, D_3, and D_4. There is almost unanimous agreement about D_1. There is about 85 percent agreement on D_2. There is about 70 percent agreement on D_3, and almost no agreement on D_4. One conclusion stands out from these informal tests; the average engineer has remarkably good intuition in problems of this kind. The majority vote for D_1, D_2, and D_3 has always been in agreement with the formal mathematical solution. However, there has been almost universal disagreement over how to defend the intuitive solution. That is, while many engineers could agree on the best course of action, they were much less in agreement on <u>why</u> that course was the best one."

14.7. Solution For Stage 2.

Now, how are we to set up this problem mathematically? In a real life situation, evidently, the problem would be a little more complicated than indicated so far, because what Mr. A does today also affects how serious his problem will be tomorrow. Mr. A's decision each day should not depend only on orders expected for that day; they should be based on his best estimates of orders likely to come in for all future days, and on the consequences of failure to meet all orders not only today but also in the future. That would

get us into the subject of dynamic programming. But for now, just to keep the problem simple, let's solve only the truncated problem in which he makes decisions on a day to day basis with no thought of tomorrow.

We have just to carry out the steps enumerated under "General Decision Theory" at the end of the last lecture. Since Stage 1 is almost too trivial to work with, consider the problem of stage 2. First, enumerate the possible "states of nature" θ_j. These correspond to all possible order situations that could arise; if Mr. A knew in advance exactly how many red, yellow, and green widgets would be ordered today, his decision problem would be trivial. Let $n_1 = 0, 1, 2, \ldots$ be the number of red widgets that will be ordered today, and similarly n_2, n_3 for yellow and green respectively. Then any conceivable order situation is given by specifying three non-negative integers $\{n_1,n_2,n_3\}$. Conversely, every ordered triple of non-negative integers represents a conceivable order situation.

Next, we are to assign prior probabilities $(\theta_j|X) = (n_1 n_2 n_3|X)$ to the states of nature, which maximize the entropy of the distribution subject to the constraints of our prior knowledge. We solved this problem generally in Lecture 10, Equations (10-26)--(10-32); and so we just have to translate the result into our present notation. The index i on x_i in Lecture 10 now corresponds to the three integers n_1, n_2, n_3; the functions $f_k(x_i)$ also correspond to the n_i, since the prior information at this stage is that the expectations $\langle n_1\rangle$, $\langle n_2\rangle$, $\langle n_3\rangle$ of orders for red, yellow, and green widgets are given as 50, 100, 10 respectively. With three average values given, we will have three Lagrange multipliers λ_1, λ_2, λ_3, and the partition function (10-30) becomes

$$Z(\lambda_1,\lambda_2,\lambda_3) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \sum_{n_3=0}^{\infty} \exp(-\lambda_1 n_1 - \lambda_2 n_2 - \lambda_3 n_3)$$

$$= \prod_{i=1}^{3} (1 - e^{-\lambda_i})^{-1} \qquad (14\text{-}44)$$

The λ_i are determined from (10-32):

$$\langle n_i \rangle = -\frac{\partial}{\partial \lambda_i} \log Z$$

$$= \left(\frac{1}{e^{\lambda_i} - 1} \right) \tag{14-45}$$

The maximum-entropy probability assignment (10-28) for the states of nature $\theta_j = \{n_1\ n_2\ n_3\}$ therefore factors:

$$p(n_1 n_2 n_3) = p_1(n_1)\ p_2(n_2)\ p_3(n_3) \tag{14-46}$$

with

$$p_i(n_i) = (1 - e^{-\lambda_i})\ e^{-\lambda_i n_i}, \quad n_i = 0, 1, 2, \ldots$$

$$= \frac{1}{\langle n_i \rangle + 1} \left[\frac{\langle n_i \rangle}{\langle n_i \rangle + 1} \right]^{n_i} \tag{14-47}$$

Thus in stage 2, Mr. A's state of knowledge about today's orders is given by three exponential distributions:

$$p_1(n_1) = \frac{1}{51} \left(\frac{50}{51} \right)^{n_1}$$

$$p_2(n_2) = \frac{1}{101} \left(\frac{100}{101} \right)^{n_2}$$

$$p_3(n_3) = \frac{1}{11} \left(\frac{10}{11} \right)^{n_3} \tag{14-48}$$

which completes step 2. Step 3, application of Bayes' theorem to digest new evidence E, is absent because there is no new evidence. Therefore, the decision must be made directly from the prior probabilities (14-48), as is always the case in statistical mechanics. So, we now proceed to step 4, enumerate the possible decisions. These are $D_1 \equiv$ make red ones today, $D_2 \equiv$ make yellow ones, $D_3 \equiv$ make green ones. In step 5, we are to introduce a loss function $L(D_i, \theta_j)$. Mr. A's judgment is that there is no loss if all orders are filled today; otherwise the loss will be proportional to--and in

view of the invariance of the decision rule under proper linear transformations that we noted at the end of Lecture 13, we may as well take it equal to--the total number of unfilled orders.

The present stock of red, yellow, and green widgets is $S_1 = 100$, $S_2 = 150$, $S_3 = 50$ respectively. On decision D_1 (make red widgets) the available stock S_1 will be increased by the day's run of 200 widgets, and the loss will be

$$L(D_1;n_1n_2n_3) = g(n_1 - S_1 - 200) + g(n_2 - S_2) + g(n_3 - S_3) \quad (14\text{-}49)$$

where $g(x)$ is the ramp function

$$g(x) \equiv \begin{cases} x, & x \geq 0 \\ 0, & x \leq 0 \end{cases} \quad (14\text{-}50)$$

Likewise, on decision D_2, D_3 the loss will be

$$L(D_2;\ n_1n_2n_3) = g(n_1 - S_1) + g(n_2 - S_2 - 200) + g(n_3 - S_3) \quad (14\text{-}51)$$

$$L(D_3;\ n_1n_2n_3) = g(n_1 - S_1) + g(n_2 - S_2) + g(n_3 - S_3 - 200) \quad (14\text{-}52)$$

So, if decision D_1 is made, the expected loss will be

$$\langle L\rangle_1 = \sum_{n_i} p(n_1n_2n_3)\ L(D_1;n_1n_2n_3)$$

$$= \sum_{n_1=0}^{\infty} p_1(n_1)\ g(n_1-S_1-200) + \sum_{n_2=0}^{\infty} p_2(n_2)\ g(n_2-S_2)$$

$$+ \sum_{n_3=0}^{\infty} p_3(n_3)\ g(n_3-S_3) \quad (14\text{-}53)$$

and similarly for D_2, D_3. The summations are elementary, giving

$$\langle L\rangle_1 = \langle n_1\rangle e^{-\lambda_1(S_1+200)} + \langle n_2\rangle e^{-\lambda_2 S_2} + \langle n_3\rangle e^{-\lambda_3 S_3}$$

$$\langle L\rangle_2 = \langle n_1\rangle e^{-\lambda_1 S_1} + \langle n_2\rangle e^{-\lambda_2(S_2+200)} + \langle n_3\rangle e^{-\lambda_3 S_3}$$

$$\langle L\rangle_3 = \langle n_1\rangle e^{-\lambda_1 S_1} + \langle n_2\rangle e^{-\lambda_2 S_2} + \langle n_3\rangle e^{-\lambda_3(S_3+200)} \quad (14\text{-}54)$$

or, inserting numerical values

$$\langle L\rangle_1 = 0.131 + 22.480 + 0.085 = 22.70$$

$$\langle L\rangle_2 = 6.902 + \ 3.073 + 0.085 = 10.06$$

$$\langle L\rangle_3 = 6.902 + 22.480 + 4 \times 10^{-10} = 29.38 \qquad (14\text{-}55)$$

showing a strong preference for decision $D_2 \equiv$ "make yellow ones today," as common sense had already anticipated.

You will recognize that Stage 2 of Mr. A's decision problem is mathematically the same as the theory of the harmonic oscillator in quantum statistical mechanics. There is still another engineering application of the harmonic oscillator equations, in some problems of message encoding, that we'll see when we take up communication theory. I'm trying to emphasize the generality of this theory, which is mathematically quite old and well known, but which has been applied in the past only in some specialized problems in physics. This general applicability can be seen only after we are emancipated from the orthodox view that all probability distributions must be justified in the frequency sense. Historically, this made it appear to most workers in statistical mechanics that the methods of Gibbs could be justified only via unproved "ergodic hypotheses" (in spite of the fact that Gibbs himself never mentioned them). But if we interpret Gibbs' equations not as assertions about frequencies but as examples of inductive reasoning based on the principle of maximum entropy, it is clear that the reasoning doesn't depend on ergodic properties or any other aspect of the laws of physics--ergo, the canonical ensemble formalism of Gibbs can be applied to any problem of inductive reasoning where the given information can be stated in the form of mean values.

14.8. Solution For Stage 3.

In Stage 3 of Mr. A's problem we have some additional pieces of information giving the average individual orders for red, yellow, and green widgets.

To take account of this new information, we need to set up a more detailed enumeration of the states of nature, in which we take into account not only the total orders for each type, but also the breakdown into individual orders. We could have done this also in stage 2, but since at that stage there was no information available bearing on this breakdown, it would have added nothing to the problem. However, in the interest of checking the consistency of this theory, you may find it amusing to retrace stage 2 on this basis and see how it would have led to exactly the same results given above.

In stage 3, a possible state of nature can be described as follows. We receive u_1 individual orders for 1 red widget each, u_2 orders for 2 red widgets each, ..., u_r individual orders for r red widgets each. Also, we receive v_y orders for y yellow widgets each, and w_g orders for g green widgets each. Thus a state of nature is specified by an infinite number of non-negative integers

$$\theta = \{u_1 u_2 \ldots ; v_1 v_2 \ldots ; w_1 w_2 \ldots\} \tag{14-56}$$

and conversely every such set of integers represents a conceivable state of nature, to which we assign a probability $p(u_1 u_2 \ldots ; v_1 v_2 \ldots ; w_1 w_2 \ldots)$.

Today's total demand for red, yellow and green widgets is, respectively

$$\begin{aligned} n_1 &= \sum_{r=1}^{\infty} r\, u_r \\ n_2 &= \sum_{y=1}^{\infty} y\, v_y \\ n_3 &= \sum_{g=1}^{\infty} g\, w_g \end{aligned} \tag{14-57}$$

the expectations of which were given in stage 2 as $\langle n_1 \rangle = 50$, $\langle n_2 \rangle = 100$, $\langle n_3 \rangle = 10$. The total number of individual orders for red, yellow, and green widgets are respectively

$$m_1 = \sum_{r=1}^{\infty} u_r$$

$$m_2 = \sum_{y=1}^{\infty} v_y$$

$$m_3 = \sum_{g=1}^{\infty} w_g \tag{14-58}$$

And the new feature of stage 3 is that $\langle m_1 \rangle$, $\langle m_2 \rangle$, $\langle m_3 \rangle$ are also known. For example, the statement that the average individual order for red widgets is 75 means that $\langle n_1 \rangle = 75 \langle m_1 \rangle$.

With six average values given, we will have six Lagrange multipliers $\{\lambda_1\mu_1;\lambda_2\mu_2;\lambda_3\mu_3\}$. The maximum-entropy probability assignment will have the form

$$p(u_1u_2\ldots;v_1v_2\ldots;w_1w_2\ldots) = \exp(-\lambda_0-\lambda_1 n_1-\mu_1 m_1-\lambda_2 n_2-\mu_2 m_2-\lambda_3 n_3-\mu_3 m_3)$$

which factors:

$$p(u_1u_2\ldots;v_1v_2\ldots;w_1w_2\ldots) = p_1(u_1u_2\ldots)p_2(v_1v_2\ldots)p_3(w_1w_2\ldots) \tag{14-59}$$

The partition function also factors:

$$Z = Z_1(\lambda_1\mu_1)\ Z_2(\lambda_2\mu_2)\ Z_3(\lambda_3\mu_3) \tag{14-60}$$

with

$$Z_1(\lambda_1\mu_1) = \sum_{u_1=1}^{\infty} \sum_{u_2=1}^{\infty} \cdots \exp[-\lambda_1(u_1+2u_2+3u_3+\ldots)-\mu_1(u_1+u_2+u_3+\ldots)]$$

$$= \prod_{r=1}^{\infty} \frac{1}{1 - e^{-r\lambda_1-\mu_1}} \tag{14-61}$$

with similar expressions for Z_2, Z_3. To find λ_1, μ_1 we apply the general rule, Equation (10-32):

$$\langle n_1 \rangle = \frac{\partial}{\partial \lambda_1} \sum_{r=1}^{\infty} \log(1 - e^{-r\lambda_1-\mu_1}) = \sum_{r=1}^{\infty} \frac{r}{e^{r\lambda_1+\mu_1} - 1} \tag{14-62}$$

$$\langle m_1 \rangle = \frac{\partial}{\partial \mu_1} \sum_{r=1}^{\infty} \log(1 - e^{-r\lambda_1-\mu_1}) = \sum_{r=1}^{\infty} \frac{1}{e^{r\lambda_1+\mu_1} - 1} \tag{14-63}$$

Comparing with equations (14-57), (14-58), we see that

$$\langle u_r \rangle = \frac{1}{e^{r\lambda_1+\mu_1} - 1} \tag{14-64}$$

and now the secret is out--Stage 3 of Mr. A's decision problem is just the theory of an ideal Bose-Einstein gas in quantum statistical mechanics!

If we treat the ideal Bose-Einstein gas by the method of the grand canonical ensemble, we obtain just these equations, in which the number r corresponds to the r'th single-particle energy level, u_r to the number of particles in the r'th state, λ_1 and μ_1 to the temperature and chemical potential.

In the present problem it is clear that for all r, $\langle u_r \rangle \ll 1$, and that $\langle u_r \rangle$ cannot decrease appreciably below $\langle u_1 \rangle$ until r is of the order of 75, the average individual order. Therefore, μ_1 will be numerically large, and λ_1 numerically small, compared to unity. This means that the series (14-62), (14-63) converge very slowly and are useless for numerical work unless you have a big computer. However, we can transform them into rapidly converging sums as follows:

$$\sum_{r=1}^{\infty} \frac{1}{e^{\lambda r+\mu} - 1} = \sum_{r=1}^{\infty} \sum_{n=1}^{\infty} e^{-n(\lambda r+\mu)}$$

$$= \sum_{n=1}^{\infty} \frac{e^{-n\mu}}{1 - e^{-n\lambda}} \tag{14-65}$$

The first term is already an excellent approximation. Similarly,

$$\sum_{r=1}^{\infty} \frac{r}{e^{\lambda r+\mu} - 1} = \sum_{n=1}^{\infty} \frac{e^{-n(\lambda+\mu)}}{(1 - e^{-n\lambda})^2} \tag{14-66}$$

and so (14-62) and (14-63) become

$$\langle n_1 \rangle \simeq \frac{e^{-\mu_1}}{\lambda_1^2} \tag{14-67}$$

$$\langle m_1 \rangle \simeq \frac{e^{-\mu_1}}{\lambda_1} \tag{14-68}$$

or

$$\lambda_1 \simeq \frac{\langle m_1 \rangle}{\langle n_1 \rangle} = \frac{1}{75} \simeq 0.0133 \tag{14-69}$$

$$e^{\mu_1} \simeq \frac{\langle n_1 \rangle}{\langle m_1 \rangle^2} = 112.5 \tag{14-70}$$

$$\mu_1 = 4.722 \tag{14-71}$$

Tribus and Fitts, evaluating the sums exactly by computer, get $\lambda_1 = 0.0131$, $\mu_1 = 4.727$; so our approximations (14-67), (14-68) are very good, at least in the case of red widgets.

The probability that u_r has a particular value is, from (14-59) or (14-61),

$$p(u_r) = (1 - e^{-r\lambda_1 - \mu_1})\, e^{-(r\lambda_1 + \mu_1)u_r} \tag{14-72}$$

which has the mean value (14-64) and the variance

$$\mathrm{var}(u_r) = \langle u_r^2 \rangle - \langle u_r \rangle^2 = \frac{e^{r\lambda_1 + \mu_1}}{(e^{r\lambda_1 + \mu_1} - 1)^2} \tag{14-73}$$

The total demand for red widgets

$$n_1 = \sum_{r=1}^{\infty} r u_r \tag{14-74}$$

is expressed as the sum of a large number of independent "random variables". The probability distribution for n_1 will have the mean value (14-67) and the variance

$$\mathrm{var}(n_1) = \sum_{r=1}^{\infty} r^2\, \mathrm{var}(u_r) = \sum_{r=1}^{\infty} \frac{r^2\, e^{r\lambda_1 + \mu_1}}{(e^{r\lambda_1 + \mu_1} - 1)^2} \tag{14-75}$$

which we convert into the rapidly convergent sum

$$\sum_{r,n=1}^{\infty} n r^2\, e^{-n(r\lambda + \mu)} = \sum_{n=1}^{\infty} n\, \frac{e^{-n(\lambda+\mu)} + e^{-n(2\lambda+\mu)}}{(1 - e^{-n\lambda})^3} \tag{14-76}$$

or, approximately,

$$\mathrm{var}(n_1) \simeq \frac{2e^{-\mu_1}}{\lambda_1^{\,3}} = \frac{2}{\lambda_1}\langle n_1\rangle \ . \qquad (14\text{-}77)$$

At this point I have to anticipate some mathematical facts concerning the Central Limit Theorem, that we'll study later. Because n_1 is the sum of a large number of small terms, the probability distribution for n_1 will be very nearly gaussian:

$$p(n_1) \simeq A \exp\left\{-\frac{\lambda_1(n_1-\langle n_1\rangle)^2}{4\langle n_1\rangle}\right\} \qquad (14\text{-}78)$$

for those values of n_1 which can arise in many different ways. For example, the case $n_1 = 2$ can arise in only two ways: $u_1 = 2$, or $u_2 = 1$, all other u_k being zero. On the other hand, the case $n_1 = 150$ can arise in an enormous number of different ways, and the "smoothing" mechanism of the central limit theorem can operate. Thus, Equation (14-78) will be a good enough approximation for the large values of n_1 of interest to us, but it may not be for small n_1.

Lecture 15

SURVEY OF ORTHODOX PRINCIPLES

Now I want to turn to a few other topics which come under the heading of clearing up various questions that were left dangling in previous lectures. We need to have an understanding of the terminology and the various concepts and principles of orthodox statistics in order to make comparisons and refer easily to the existing literature. We have already examined the principle of maximum likelihood in Lecture 9, and in the last two lectures we saw something of the orthodox principles for point estimation of parameters, and the orthodox approach to decision theory. This seems like as good a time as any to extend the list.

The methods to be described are now obsolete, in the sense that Bayesian methods either include them as special cases, or improve on them. Nevertheless, they exist, the literature is full of them, and they will continue to appear in the literature throughout our lifetimes, because many Statistics Departments are still teaching them to their students as if Bayesian methods didn't exist. So, we have no choice but to learn the terminology of orthodox statistics.

However, don't get the impression that there exists any definite monolithic "orthodox theory." In fact, orthodox statistics is a mish-mash of mutually contradictory ad hoc principles, and there are just as many--and just as bitter-controversies between different workers within the orthodox school as between orthodox and Bayesian advocates.

15.1. Sufficient Statistics.

Given a sampling distribution function $(x_1 \ldots x_n|\alpha)$ and a proposed estimator $\beta(x_1 \ldots x_n)$ of α, let us carry out a change of variables $(x_1 \ldots x_n) \to (y_1 \ldots y_n)$ such that $y_1 = \beta(x_1 \ldots x_n)$ and the jacobian $J = \partial(y_1 \ldots y_n)/\partial(x_1 \ldots x_n)$ is finite and not identically zero. Then the sampling distribution function of the y_i is

$$(y_1 \ldots y_n|\alpha) = (x_1 \ldots x_n|\alpha)\,|J|^{-1} \tag{15-1}$$

By our Rule 1, this can be factored:

$$(y_1 \ldots y_n|\alpha) = (\beta y_2 \ldots y_n|\alpha) = (\beta|\alpha)(y_2 \ldots y_n|\beta\alpha) \tag{15-2}$$

Suppose now that $(y_2 \ldots y_n|\beta\alpha)$ turns out to be independent of α. This is equivalent to saying that the original sampling distribution can be factored in the form

$$(x_1 \ldots x_n|\alpha) = g(x_1 \ldots x_n)\,(\beta|\alpha) \tag{15-3}$$

where $g(x_1 \ldots x_n) = (y_2 \ldots y_n|\beta)|J|$ can be expressed as a function of the x_i, not involving α. Therefore, if β is known, knowing the value of α would give us no more information about the sample. Conversely, it seems intuitively that if β is known, then knowledge of $(y_2 \ldots y_n)$ could give us no further information about α; i.e. <u>all</u> the information in the sample, that is relevant for inference about α, is contained summarized in the single function $\beta(x_1 \ldots x_n)$.

Let us check whether this is true. The ultimate criterion is, of course, whether the conjectured property can be derived from Bayes' theorem; i.e. whether the posterior distribution $(\alpha|x_1 \ldots x_n)$ depends on the sample only through the function $\beta(x_1 \ldots x_n)$.

This distribution has the form

$$(\alpha|x_1 \ldots x_n) = \frac{(x_1 \ldots x_n|\alpha)\ f(\alpha)}{\int (x_1 \ldots x_n|\alpha)\ f(\alpha)\ d\alpha} \tag{15-4}$$

where $f(\alpha)$ is a prior probability density.

Substituting (15-3) into this, we obtain

$$(\alpha|x_1 \ldots x_n) = \frac{(\beta|\alpha)\ g(x_1 \ldots x_n)\ f(\alpha)}{\int (\beta|\alpha)\ g(x_1 \ldots x_n)\ f(\alpha)\ d\alpha} \tag{15-5}$$

Since $g(x_1 \ldots x_n)$ does not depend on α, it cancels out, leaving us with

$$(\alpha|x_1 \ldots x_n) = \frac{(\beta|\alpha)\ f(\alpha)}{\int (\beta|\alpha)\ f(\alpha)\ d\alpha} \tag{15-6}$$

which says, as conjectured, that the posterior probability distribution of α depends only on the particular function $\beta(x_1 \ldots x_n)$ of the sample values. All other properties of the sample are irrelevant for inference about α. In this case, β is said to be a sufficient statistic for α, a terminology introduced by Fisher. More generally, any function $f(x_1 \ldots x_n)$ of the sample values is called a "statistic."

For example, let α be the unknown mean value of a gaussian distribution of known variance σ^2. Then

$$(x_1 \ldots x_n|\alpha) = A \exp\left[- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \alpha)^2 \right] \tag{15-7}$$

where A is a normalizing constant. Rearranging, we have

$$\begin{aligned}(x_1 \ldots x_n|\alpha) &= A \exp\left[- \frac{n}{2\sigma^2} (\overline{x^2} - 2\alpha\overline{x} + \alpha^2) \right] \\ &= A \exp\left[- \frac{ns^2}{2\sigma^2} \right] \exp\left[- \frac{n}{2\sigma^2} (\overline{x} - \alpha)^2 \right]\end{aligned} \tag{15-8}$$

where

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{15-9}$$

$$\overline{x^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \tag{15-10}$$

$$s^2 = \overline{x^2} - \overline{x}^2 \tag{15-11}$$

are the sample mean, mean square, and variance respectively.

Suppose we propose the sample mean as our estimator; i.e. we take

$$\beta(x_1 \ldots x_n) \equiv \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{15-12}$$

The sampling distribution of β is

$$(\beta|\alpha) = \int \cdots \int dx_1 \ldots dx_n \; \delta\left(\beta - \frac{1}{n}\sum x_i\right) (x_1 \ldots x_n|\alpha) \qquad (15\text{-}13)$$

To evaluate this, it is easier to take first its Fourier transform, or <u>characteristic function</u>:

$$\phi(k) \equiv \left\langle e^{ik\beta} \right\rangle = \int_{-\infty}^{\infty} (\beta|\alpha)\; e^{ik\beta}\, d\beta$$

$$= \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_n \; e^{i\frac{k}{n}(x_1+\ldots+x_n)} \; (x_1 \ldots x_n|\alpha)$$

The integration is elementary, and we find

$$\phi(k) = \exp\left[\, ik\alpha - \frac{k^2\sigma^2}{n}\right] \qquad (15\text{-}14)$$

Then, inverting the Fourier integral, we have

$$(\beta|\alpha) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(k)\; e^{-ik\beta}\, dk$$

$$= \sqrt{\frac{n}{2\pi\sigma^2}}\, \exp\left[\, - \frac{n}{2\sigma^2}\, (\beta - \alpha)^2\right] \qquad (15\text{-}15)$$

But, comparing with (14-8), we see that the factorization property (15-3) does hold for this estimator, and consequently β is a sufficient statistic for estimation of α.

Conversely, applying Bayes' theorem (14-4), we find

$$(\alpha|x_1 \ldots x_n) = \frac{\exp\left[\, - \frac{n}{2\sigma^2}\, (\alpha - \overline{x})^2\right] f(\alpha)}{\int \exp\left[\, - \frac{n}{2\sigma^2}\, (\alpha - \overline{x})^2\right] f(\alpha)\; d\alpha} \qquad (15\text{-}16)$$

which says again that the sample mean $\overline{x}$ is a sufficient statistic for estimation of α. The parameter α would be termed the "population mean" by the statistician. However, this underlying "population" is entirely fictitious in most real problems.

If the mean α and standard deviation σ are both unknown, we can apply Bayes' theorem to find their joint posterior probability density $(\alpha\sigma|x_1 \ldots x_n)$.

In this case we need the correct normalization constant for the sample distribution function:

$$(x_1 \ldots x_n | \alpha,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \alpha)^2}{2\sigma^2}\right\} \qquad (15\text{-}17)$$

Bayes' theorem then yields, with prior probability density $f(\alpha,\sigma)$:

$$(\alpha\sigma | x_1 \ldots x_n) = \frac{A\, f(\alpha,\sigma)}{\sigma^n} \exp\left\{-\frac{n}{2\sigma^2}\,[s^2 + (\alpha - \bar{x})^2]\right\} \qquad (15\text{-}18)$$

where A is a normalizing constant independent of α and σ. Since only the sample mean and variance $\bar{x}$, s^2 appear here, $\bar{x}$ and s^2 are <u>jointly sufficient</u> for α and σ, a fact that I mentioned briefly at the end of Lecture 6. In general (15-18) will show some correlation between α and σ; but if we just want the best estimates of each independently of the other, we get them from the <u>marginal distributions</u> obtained by integrating out the unwanted parameter:

$$(\alpha | x_1 \ldots x_n) = \int (\alpha\sigma | x_1 \ldots x_n)\, d\sigma \qquad (15\text{-}19)$$

$$(\sigma | x_1 \ldots x_n) = \int (\alpha\sigma | x_1 \ldots x_n)\, d\alpha \qquad (15\text{-}20)$$

Similarly, let $0 < \alpha < 1$, $0 \le x_i < \infty$, and consider the distribution

$$(x_1 \ldots x_n | \alpha) = A \prod_{i=1}^{n} x_i^p\, \alpha^{x_i} \qquad (15\text{-}21)$$

Since this factors:

$$(x_1 \ldots x_n | \alpha) = A\, (x_1 \ldots x_n)^p\, \alpha^{n\bar{x}} \qquad (15\text{-}22)$$

we find as before that the sample mean $\bar{x}$ is a sufficient statistic for α, and the best estimate of α, by the criterion of any loss function, will be some function $\beta(\bar{x})$ of the sample mean only.

Likewise, consider the rectangular distribution

$$(x_1 \ldots x_n | \alpha) = \prod_{i=1}^{n} f_\alpha(x_i) \qquad (15\text{-}23)$$

where

$$f_\alpha(x) \equiv \begin{cases} 0, & x < 0 \\ \alpha^{-1}, & 0 \le x \le \alpha \\ 0, & \alpha < x \end{cases} \tag{15-24}$$

Thus,

$$(x_1 \ldots x_n | \alpha) = \begin{cases} 0, & x_{min} < 0 \\ \alpha^{-n}, & 0 \le x_{min} \le x_{max} \le \alpha \\ 0, & \alpha < x_{max} \end{cases} \tag{15-25}$$

where x_{min}, x_{max} are the minimum and maximum observed sample values. The posterior distribution $(\alpha | x_1 \ldots x_n)$ depends on the x_i only through x_{max}, (and, of course, on the number n of observations). Consequently x_{max} is a sufficient statistic for estimation of α; or, in a little different terminology often found in the literature, x_{max} and n are <u>jointly sufficient</u>.

Evidently, the condition for existence of a sufficient statistic is that a single function $\gamma(x_1 \ldots x_n)$ of the sample values must exist such that $(x_1 \ldots x_n | \alpha)$ factors into the form

$$(x_1 \ldots x_n | \alpha) = g(x_1 \ldots x_n)\ h(\gamma, \alpha). \tag{15-26}$$

For the rectangular distribution, this is the case with $\gamma(x_1 \ldots x_n) = x_{max}$, $g(x_1 \ldots x_n) = 1$, and

$$h(\gamma, \alpha) = \begin{cases} \alpha^{-n}, & \alpha \ge \gamma \\ 0, & \alpha < \gamma \end{cases} \tag{15-27}$$

A sufficient statistic does not always exist. For example, the Cauchy distribution $(x_1 \ldots x_n | \alpha) = A \prod_{i=1}^{n} [1 + (x_i - \alpha)^2]^{-1}$ does not admit any factorization of the form (15-26), nor does the truncated exponential distribution $(x_1 \ldots x_n | \alpha) = A \exp[-\alpha(x_1 + \ldots + x_n)]$, $0 \le x_1 \ldots x_n \le \alpha$. But in the latter case $\bar{x}$ and x_{max} are jointly sufficient for α.

15.2. Efficient Estimates.

I have already pointed out [Eq. (13-38)] that the criterion of minimum α-expected loss does not in general lead to any specific "best" estimator $\beta(x_1 \ldots x_n)$, but it may do so in some special cases. We can now exhibit one such special case. Consider a quadratic loss function $L(\alpha,\beta) = (\beta - \alpha)^2$, and independent sampling so that

$$(x_1 \ldots x_n | \alpha) = f(x_1,\alpha)\ f(x_2,\alpha) \ \ldots\ f(x_n,\alpha) \tag{15-28}$$

An estimator β which minimizes the α-expected loss was called "efficient" by R. A. Fisher. In some of the later literature, however, the term "efficient" is taken to mean only that this condition is approached asymptotically, in the limit of large samples. This is the condition called "asymptotic efficiency" by Cramér (1946). A famous inequality associated with the names of Fréchet, Darmois, Rao, Cramér, and others, places a lower limit on the α-expected loss with any estimator $\beta(x_1 \ldots x_n)$:

$$\langle(\beta - \alpha)^2\rangle \geq \frac{\left(\frac{d\langle\beta\rangle}{d\alpha}\right)^2}{n\int\left(\frac{\partial \log f}{\partial\alpha}\right)^2 f(x,\alpha)\ dx} \tag{15-29}$$

with equality when and only when the following two conditions are met:

(1) β is a sufficient statistic for estimation of α, i.e.

$$(x_1 \ldots x_n | \alpha) = g(x_1 \ldots x_n)\ h(\beta,\alpha) \tag{15-30}$$

(2) the function $h(\beta,\alpha)$ satisfies

$$\frac{\partial \log h}{\partial\alpha} = k(\alpha)(\beta - \alpha) \tag{15-31}$$

for some function $k(\alpha)$. A simple proof of this theorem is given by Cramér (1946, Sec. 32.3). From (15-30) and (15-31) it is seen that the sampling distribution function must also satisfy

$$\frac{\partial \log (x_1 \ldots x_n | \alpha)}{\partial \alpha} = k(\alpha)(\beta - \alpha) \tag{15-32}$$

or, on integration, it must have the form

$$(x_1 \ldots x_n | \alpha) = \frac{m(x_1 \ldots x_n) \exp[-\lambda(\alpha)\ \beta(x_1 \ldots x_n)]}{Z(\lambda)} \tag{15-33}$$

where λ depends only on α, and

$$Z(\lambda) \equiv \int m(x_1 \ldots x_n) \exp[-\lambda \beta(x_1 \ldots x_n)]\ dx_1 \ldots dx_n \tag{15-34}$$

Since this is just the canonical distribution of statistical mechanics, we may restate the theorem as follows: The best estimator $\beta(x_1 \ldots x_n)$ by the criterion of minimum α-expected loss, which achieves equality in (15-29), exists when and only when the sampling distribution function has the canonical form with maximum entropy, relative to some weighting function $m(x_1 \ldots x_n)$, for a given expectation value $<\beta>$.

Thus, for example, the energy of a system at thermal equilibrium is always a sufficient and efficient statistic for estimation of the temperature of the heat-bath surrounding it, all other details of its state being irrelevant for that purpose.

We examined the notion of sufficiency in Lecture 14, from the standpoint of "information" in the sense of entropy, and saw in Eq. (14-14) the exact sense in which the colloquial term "information" is related to entropy. Although "sufficiency" was introduced by R. A. Fisher within the context of orthodox statistics, we saw in Eq. (15-6) that it is exactly derivable from Bayes' theorem. Therefore, it remains a valid and useful notion in Bayesian statistics; any problem of inference in which a single sufficient statistic exists, will be vastly simpler mathematically, and will lead to much shorter calculations in applications. Generally, in nontrivial real problems where a sufficient statistic does not exist, we will be driven to approximations

in reducing data.

The notion of efficiency, however, is not of any particular value in Bayesian statistics, because Bayes' theorem automatically gives us the best estimator by the criterion of any loss function. Thus the need to compare different estimators doesn't arise unless the equations are so complicated that we have to resort to approximations. But then it is the x-expected loss [as defined in Eq. (13-15)] rather than the α-expected loss that provides our criterion of good approximation.

Furthermore, the notion of efficiency doesn't really have any "objective" meaning, because it depends on the particular way you or I choose to define our parameters. For example, instead of the parameter α, there is no reason why we couldn't use just as well, the parameter $\gamma \equiv \alpha^2$, or $\delta \equiv \log \alpha$, etc., and of course any satisfactory statistical methods ought to lead us to the same final conclusions however we have defined our parameters. But the Fisher definition of efficiency is so parameter-dependent that if an efficient estimator of α exists, then an efficient estimator of α^2 does <u>not</u> exist! For these reasons, we will have no further use for the concept of efficiency.

15.3. Tests of Goodness of Fit.

Back in Lecture 7, when we discussed the application of Bayes' theorem to such problems as the validity of Newtonian celestial mechanics, we noted this: Bayes' theorem tells us that we cannot say how the observed facts affect the probability of some hypothesis H, until we state some specific alternatives against which H is to be tested. For example. suppose there are only two possible hypotheses, H and H', to be considered. Then, on any data D, we must always have $(H|D) + (H'|D) = 1$, and in terms of our logarithmic measure of plausibility in decibels, Bayes' theorem becomes

$$e(H|D) = e(H|X) + 10 \log_{10} \frac{(D|H)}{(D|H')} \qquad (15\text{-}35)$$

which we might describe in words by saying that, "data D supports hypothesis H relative to H', by 10 $\log_{10}$ $(D|H)/(D|H')$ decibels." The phrase relative to H' is essential here, since with some other alternative H", the change in evidence for H, $[e(H|D) - e(H|X)]$ might be entirely different; it does not make sense to ask how much the observed facts tend "in themselves" to support or refute H (except, of course, in the case where D is absolutely impossible on hypothesis H, so deductive reasoning can take over).

Now as long as we talk only in these generalities, our common sense readily assents to this. But if we consider specific problems, we may have some doubts. For example, in the particle counter problem of Lecture 8 we had a case (known source strength s and known counter efficiency a) where the probability of getting c counts in any one second is a Poisson distribution (8-5) with mean value $\bar{c} = sa$:

$$(c|s,a) = e^{-sa}\frac{(sa)^c}{c!} \tag{15-36}$$

Although it wasn't necessary for the problem we were considering then, we can still ask: what can we infer from this about the relative frequencies with which we would see c counts if we repeat the measurement in many different seconds, with the result $\{c_1 c_2 \ldots c_n\}$? If the probability of any particular event (say the event c = 12) is independently equal to

$$p = e^{-sa}\frac{(sa)^{12}}{12!} \tag{15-37}$$

at each trial, then the probability that the event will occur exactly r times in n trials is the binomial distribution

$$(r|n) = \binom{n}{r} p^r (1-p)^{n-r} \tag{15-38}$$

or, the probability that it will occur with frequency f = r/n, is

$$(f|n) = \frac{n!}{(nf)!\,(n-nf)!}\, p^{nf}(1-p)^{n-nf} \tag{15-39}$$

When n is very large, we can use the Stirling approximation (10-16) to get

$$L \equiv \frac{1}{n} \log (f/n)$$

$$\simeq - f \log f - (1-f) \log (1-f) + f \log p + (1-f) \log (1-p) \qquad (15\text{-}40)$$

Treating f as a continuous variable,

$$\frac{\partial L}{\partial f} = \log \frac{1-f}{f} - \log \frac{1-p}{p}$$

$$\frac{\partial^2 L}{\partial f^2} = - \frac{1}{f(1-f)}$$

So L reaches a maximum at $f = p$, and we have the Taylor series expansion about that point:

$$L(f) = L(p) - \frac{(f-p)^2}{2p(1-p)} + \ldots$$

Therefore, an approximation (which is actually much better than you might guess from this simple derivation) to (15-39) is

$$(f|n) \simeq (\text{const.}) \cdot \exp\left\{ - \frac{n(f-p)^2}{2p(1-p)} \right\} \qquad (15\text{-}41)$$

Thus the most likely frequency to be observed is numerically equal to the probability; and the (mean ± standard deviation) estimate of the frequency is

$$(f)_{\text{est}} = p \pm \sqrt{\frac{p(1-p)}{n}} \qquad (15\text{-}42)$$

Here is another connection between probability and frequency which common sense could have anticipated, except that it would hardly give us a quantitative interval of reasonable "error." The result (15-42) will be generalized to a wider class of probability models in the next two lectures.

In the long run, therefore, we expect that the actual frequencies of various counts will be distributed in a manner approximating the Poisson distribution (15-36). Now we can perform the experiment, and the experimental frequencies either will or will not be a reasonable approximation to the predicted values. If, by the time we have observed a few thousand counts, the observed frequencies are wildly different from a Poisson distribution,

our common sense will tell us that the theory which led to Poisson prediction must be wrong. Yet we have not said anything about any alternatives! Is our common sense wrong here, or is there some way we can reconcile the theory with common sense?

Let's look again at equation (15-35). No matter what H' is, we must have $(D|H') \geq 1$, and therefore a statement which is independent of any alternative hypotheses is

$$e(H|D) \geq e(H|X) + 10 \log_{10}(D|H) = e(H|X) - \psi_\infty \qquad (15\text{-}43)$$

where

$$\psi_\infty = -10 \log_{10}(D|H) \geq 0. \qquad (15\text{-}44)$$

Thus, <u>there is no possible alternative which data D could support, relative to H, by more than ψ_∞ decibels</u>.

This suggests a solution to our paradox: in judging the amount of agreement between theory and observations, the proper question to ask is not, "How well does data D support hypothesis H?" A much better question is, "Are there any alternatives H' which data D would support relative to H, and how much support is possible?" Probability theory can give no meaningful answer to the first question, but it can give a very definite answer to the second.

We might be tempted to conclude that the proper criterion of "goodness of fit" is simply ψ_∞, or what is the same thing, the probability $(D|H)$. This is not so, however, as the following argument shows. <u>After</u> we have obtained data D, it is always possible to invent a strange, "sure thing" hypothesis H_s, according to which D was inevitable: $(D|H_s) = 1$, and H_s will always be supported relative to H by exactly ψ_∞ decibels. Let us see what this implies. Suppose I toss a die N = 10,000 times, and record the result of each toss. Then, on the hypothesis H ≡ "the die is honest," each of the

6^N possible results has probability 6^{-N}, or $\psi_\infty = 10 \log_{10}(6^N) = 77{,}815$ decibels! No matter what I observe in the 10,000 tosses, there is always an hypothesis H_s that will be supported relative to H by this enormous amount. If, after performing this experiment, I continue to believe that the die is honest, it can be only because I considered the prior probability of H_s to be very much lower than minus 77,815 decibels. Otherwise, I am reasoning inconsistently.

This is, I think, all perfectly correct and we have to accept the conclusion. The prior probability of H_s was indeed much lower than 6^{-N}, simply because there were 6^N different "sure thing" hypotheses which were all on the same footing before I observed D. But it is obvious that in practice we don't want to bother with this kind of hypothesis; even though it is supported by the data more than any other, its prior probability is so low that we are not going to accept it anyway.

In practice we are not interested in comparing H to all conceivable alternatives, but only to all those in some restricted class Ω, consisting of hypotheses which we consider to be in some sense "reasonable" a priori. Let me give one example (by far the most common and useful one) of a test relative to such a restricted class of hypotheses.

We consider some experiment, which has r possible outcomes, $A_1, A_2, \ldots, A_r$. Define the quantities

$$x_n \equiv m, \text{ if } A_m \text{ is true on the n'th trial} \tag{15-45}$$

Thus each x_n can take on the values $x_n = 1, 2, \ldots, r$. If the experiment consists of tossing a die, then $r = 6$, and x_n is the number of spots up on the n'th toss. Suppose now we wish to take into account only the hypotheses belonging to the "Bernoulli class" B_r, in which the probabilities of the A_m on successive repetitions of the experiment are considered independent and stationary; thus, when H is in B_r, the probability, conditional on H, of any specific sequence $\{x_1 \ldots x_N\}$ of observations has the form

$$(x_1 \ldots x_N | H) = p_1^{n_1} \ldots p_r^{n_r} \tag{15-46}$$

where p_m is the probability of result A_m in any trial, and n_m is the number of times A_m was true in the sequence. Of course, $\sum_m n_m = N$. To every hypothesis in B_r there corresponds a set of numbers $\{p_1 \ldots p_r\}$ such that $p_m \geq 0$, $\sum p_m = 1$, and for our present purposes these numbers completely characterize the hypothesis. Conversely, every such set of numbers defines an hypothesis belonging to the Bernoulli class B_r.

Now let's note an important lemma, which we have used before to establish some properties of entropy. Using the fact that $\log x \geq (1 - x^{-1})$, with equality if and only if $x = 1$, we find at once that

$$\sum_{i=1}^{r} n_i \log \left(\frac{n_i}{Np_i} \right) \geq 0, \tag{15-47}$$

with equality if and only if $p_i = n_i/N$ for all i. This inequality is the same as

$$\log (x_1 \ldots x_N | H) \leq N \sum f_i \log f_i \tag{15-48}$$

where $f_i = n_i/N$ is the observed frequency of result A_i. The righthand side of (15-48) depends only on the observed sample, so if we consider various hypotheses H_1, H_2, ... in B_r in the light of this particular sample, the quantity (15-47) gives us a measure of how well the different hypotheses fit the data; the nearer to equality, the better the fit.

For convenience in numerical work, let's express the quantity (15-47) in decibel units:

$$\psi_B \equiv 10 \sum_{i=1}^{r} n_i \log_{10} \left(\frac{n_i}{Np_i} \right) \tag{15-49}$$

To see the exact significance of ψ_B, suppose we apply Bayes' theorem in the form of Equation (15-35). There are only two hypotheses, $H = \{p_1 \ldots p_r\}$, and

$H' = \{p_1', \ldots, p_r'\}$ to be considered. Let the values of (15-49) computed according to H and H' be ψ_B, ψ_B' respectively. Then Bayes' theorem reads

$$e(H|x_1 \ldots x_N) = e(H|X) + 10 \log_{10} \frac{(x_1 \ldots x_N|H)}{(x_1 \ldots x_N|H')}$$

$$= e(H|X) + \psi_B' - \psi_B \qquad (15\text{-}50)$$

Now we can always find an hypothesis H' in B_r, for which $p_i' = n_i/N$, and $\psi_B' = 0$; therefore ψ_B has the following meaning:

> Given an hypothesis H and the observed data $\{x_1 \ldots x_N\}$, compute ψ_B from (15-49). Then given any $D \leq \psi_B$, it is possible to find an alternative hypothesis H' in B_r such that the data will support H' relative to H by D decibels. There is no H' in B_r which is supported relative to H by more than ψ_B decibels.

Thus, ψ_B is exactly the appropriate measure of "goodness of fit" relative to the class of Bernoulli alternatives.

We can also interpret ψ_B in this manner: we may regard the observed results $\{x_1 \ldots x_N\}$ as a "message" consisting of N symbols chosen from an alphabet of r letters. On each repetition of the experiment, Nature transmits to us one more letter of the message. How much information is transmitted by this message, under the Bernoulli probability assignment with independence of successive symbols? Note that

$$\psi_B/N = 10 \sum_{i=1}^{r} f_i \log_{10}(f_i/p_i) \qquad (15\text{-}51)$$

with $f_i = n_i/N$. Thus, $(-\psi_B/N)$ is the entropy per symbol of the observed message distribution $\{f_1 \ldots f_r\}$ relative to the "expected distribution" $\{p_1 \ldots p_r\}$. This shows that the notion of entropy is, in a sense, "inherent" in probability theory. Independently of Shannon's theorem, entropy or some monotonic function of entropy will appear automatically in the equations of

anyone who is willing to use Bayes' theorem for hypothesis testing.

Historically, a slightly different test was introduced by Karl Pearson. We expect that, if hypothesis H is true, then n_i will be close to Np_i, in the sense that the difference $|n_i - Np_i|$ will grow with N only as $\sqrt{N}$. Call this "condition A." Using the expansion $\log x = (x-1) - (x-1)^2/2 + \ldots$, we easily find that

$$\sum_{i=1}^{r} n_i \log \frac{n_i}{Np_i} = \frac{1}{2} \sum_{i=1}^{r} \frac{(n_i - Np_i)^2}{Np_i} + O\left(\frac{1}{\sqrt{N}}\right) \qquad (15\text{-}52)$$

the quantity designated as $O(1/\sqrt{N})$ tending to zero as indicated <u>provided</u> that the observed sample does in fact satisfy condition A. The quantity

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - Np_i)^2}{Np_i} = N \sum_{i=1}^{r} \frac{(f_i - p_i)^2}{p_i} \qquad (15\text{-}53)$$

is thus very nearly proportional to ψ_B, if the sample frequencies are close to the expected values:

$$\psi_B = (10 \log_{10} e) \frac{1}{2} \chi^2 + O(1/\sqrt{N}) = 2.1715 \chi^2 + O(1/\sqrt{N}) \qquad (15\text{-}54)$$

Pearson suggested that the quantity χ^2 be used as a criterion of goodness of fit, and this has led to the "Chi-squared" test, one of the most used techniques of orthodox statistics. Before describing the test, let's examine first its theoretical basis and suitability as a criterion. Evidently, $\chi^2 \geq 0$, and $\chi^2 = 0$ only if the observed frequencies agree exactly with those expected if the hypothesis is true. So, larger values of χ^2 correspond in some way to greater deviations between prediction and observation, and too large a value of χ^2 should lead us to doubt the truth of the hypothesis. But these qualitative properties are possessed also by ψ--and by any number of other quantities we could define. We have seen the theoretical basis, and precise significance, of ψ; so we ask (noting the comments of Pratt and Bross, as quoted in the preface) whether there exists any "connected argument" leading to χ^2.

The results of a search for this connected argument are disappointing. Scanning a number of orthodox textbooks, we find that χ^2 is often introduced as a straight deux ex machina; but Cramér (1946) does attempt to prepare the way for the idea, in these words: "It will then be in conformity with the general principle of least squares to adopt as measure of deviation an expression of the form $\sum c_i(n_i/N - p_i)^2$ where the coefficients c_i may be chosen more or less arbitrarily. It was shown by K. Pearson that if we take $c_i = N/p_i$, we shall obtain a deviation measure with particularly simple properties." In other words, χ^2 is adopted, not because of any connected argument but because it has, in Pratt's words, "some pleasant properties."

We have seen that in some cases χ^2 is nearly a multiple of ψ and in such cases they will of course lead to essentially the same conclusions. But let's try to understand the quantitative difference in these criteria by a technique that I want to use a lot from now on, in comparing orthodox and Bayesian methods. As discussed in the preface, we often find a small quantitative difference between Bayesian and orthodox results, which would be of no consequence in most practical problems, and is so small that our common sense is unable to pass judgment on which result is preferable. But when this happens, we can understand the difference by "magnifying" it--by finding some extreme problem where the difference is so great that our common sense can tell us which theory is giving sensible results, and which is not.

As our first example of this magnification technique, let's compare ψ and χ^2 to see which is the more reasonable criterion of goodness of fit.

15.4. Comparison of ψ and Chi-squared.

A coin toss can give three different outcomes: (1) heads, (2) tails, (3) it may stand on edge. Suppose that Mr. A's knowledge of coins is such that he assigns probabilities $p_1 = p_2 = 0.499$, $p_3 = 0.002$ to these cases.

We are in communication with Mr. B on the planet Mars, who has never seen a coin and doesn't have the slightest idea what a coin is. So, when told that there are three possible outcomes at each trial, and nothing more, he can only assign equal probabilities, $p_1' = p_2' = p_3' = 1/3$.

Now we want to test Mr. A's hypothesis against Mr. B's by doing a "random" experiment. We toss the coin 29 times and observe the outcomes: $n_1 = n_2 = 14$; $n_3 = 1$. So, we have for the two hypotheses:

$$\psi_A = 10\left[28 \log_{10}\left(\frac{14}{29 \times .499}\right) + \log_{10}\left(\frac{1}{29 \times .002}\right)\right] = 8.34 \text{ db}$$

$$\psi_B = 10\left[28 \log_{10}\left(\frac{14 \times 3}{29}\right) + \log_{10}\left(\frac{3}{29}\right)\right] = 35.19 \text{ db.}$$

From this experiment the man on Mars thus learns that (a) there is another hypothesis about the coin that is 35.2 db better than his (35.2 db corresponds to odds of over 3,300:1) and so unless he can justify an extremely low prior probability for that alternative, he cannot reasonably adhere to his first theory. (b) Mr. A's hypothesis is better than his by some 26.8 db, and in fact is within about 8 db of the best hypothesis that could be made, under our assumption of independent Bernoulli trials B_3. Here the ψ-test tells us pretty much what our common sense does.

But suppose that the man on Mars knew only about "orthodox" statistical principles as usually taught; and therefore believed that χ^2 was the proper criterion of goodness of fit. He would find that

$$\chi_A^2 = 2\,\frac{(14 - 29 \times .499)^2}{29 \times .499} + \frac{(1 - 29 \times .002)^2}{29 \times .002} = 15.33$$

$$\chi_B^2 = 2\,\frac{(14 - 29 \times .333)^2}{29 \times .333} + \frac{(1 - 29 \times .333)^2}{29 \times .333} = 11.65$$

and he would report back delightedly: "My hypothesis, by the accepted statistical test, is shown to be slightly preferable to yours!"

I think that many persons trained to use χ^2 will find this comparison startling, and will immediately try to find the error in my numerical work

above. We have here still another fulfillment of our robot's prediction back in Lecture 4. The ψ criterion is exactly derivable from Bayes' theorem; therefore any criterion which is only an approximation to it must contain either an inconsistency or a qualitative violation of common sense, which can be exhibited by producing special cases.

We can learn an important lesson about the practical use of χ^2 by looking more closely at what is happening here. On hypothesis A, the "expected" number of heads or tails in 29 tosses was $Np_1 = 14.471$. The actual observed number must be an integer; and we supposed above that in each case it was the closest possible integer, namely 14. This certainly seems a mild assumption, not harmful to hypothesis A. Yet this small discrepancy between expected and observed sample numbers, in a sense the smallest it could possibly be, nevertheless had an enormous effect on χ^2. The spook lies entirely in the fact that ${\chi_A}^2$ turned out so much larger than seems reasonable; there is nothing surprising about the other numerical values. Evidently, it is the last term in ${\chi_A}^2$, which refers to the fact that the coin stood on edge once in 29 tosses, that is causing the trouble. On hypothesis A, the probability that this would happen exactly n times in 29 tosses is our binomial distribution

$$(n|N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

with N = 29, p = 0.002. From this, we find that the probability of seeing the coin on edge one or more times in 29 trials is about (1/18); i.e. the fact that we saw it even once is a bit unexpected, and constitutes some evidence against A, that contributes 8 db to the value of ψ_A. But this amount of evidence is certainly not overwhelming; if our travel guide tells us that London has fog, on the average, one day in 18, we are hardly astonished to see fog on the day we arrive.

It is the $(1/p_i)$ weighting factor in the summand of χ^2 that causes this anomaly. Because of it, the χ^2 criterion essentially concentrates its attention on the extremely unlikely possibilities, if the hypothesis contains them; and the slightest discrepancy between expected and observed sample numbers for the unlikely events severely penalizes the hypothesis. The ψ-test also contains this effect, but in a much milder form, the $(1/p_i)$ factor appearing only in the logarithm.

To see this effect more clearly, suppose now that the experiment had yielded the results $n_1 = 14$, $n_2 = 15$, $n_3 = 0$. Evidently, by either the χ^2 or ψ criterion, this ought to make hypothesis A look better, B worse, than in the first example. Repeating the calculations, we now find

$$\psi_A = 0.30 \text{ db} \qquad \chi_A^2 = 0.0925$$

$$\psi_B = 51.2 \text{ db} \qquad \chi_B^2 = 14.55$$

You see that by far the greatest relative change was in χ_A^2; both criteria now agree that hypothesis A is far superior to B.

This shows what can happen through uncritical use of χ^2. Professor Q believes in extrasensory perception, and undertakes to prove it to us poor benighted, intransigent doubters. So he plays card games. On the "null hypothesis" that only chance is operating, it is extremely unlikely that the subject will guess many cards correctly.

The first few hundred times he plays, the results are disappointing; but these are readily explained away on the ground that the subject is not in a "receptive" mood. [The literature of parapsychology abounds with wistful complaints about the difficulty of reproducing the phenomenon; indeed, just the kind of difficulty one would expect if the phenomenon did not exist!]

But one day providence smiles on Mr. Q; the subject comes through handsomely. Immediately he calls in the statisticians, the mathematicians,

the notary publics, and the newspaper reporters. An extremely improbable event has at last occurred; and χ^2 is enormous. Now he can publish the results and assert: "The validity of the data is certified by reputable, disinterested persons, the statistical analysis has been under the supervision of recognized statisticians, the calculations have been checked by competent mathematicians. By the accepted statistical test, the null hypothesis has been decisively rejected." And everything he has said is absolutely true!

Moral: For testing hypotheses involving moderately large probabilities, which agree moderately well with observation, it won't make much difference whether we use ψ or χ^2. But for testing hypotheses involving extremely unlikely events, we had better use ψ; or life might become too exciting for us.

Now let's describe briefly the Chi-squared test as done in practice. We have the so-called "null hypothesis" H to be tested, and no alternative is stated. The null hypothesis predicts certain relative frequencies $\{p_1 \ldots p_r\}$ and corresponding sample numbers $\{Np_1, \ldots, Np_r\}$ where N is the number of trials. We observe the actual sample numbers $\{n_1, \ldots, n_r\}$. If some of the n_i are very small, we group categories together so that each n_i is at least, say, five. For example, in a case with r = 6, if the observed sample numbers were {6, 11, 14, 7, 3, 2} we would group the last two categories together, making it equivalent to a problem with r' = 5 distinguishable outcomes per trial, with sample numbers {6, 11, 14, 7, 5}, and null hypothesis H' which predicts frequencies $\{p_1, p_2, p_3, p_4, p_5+p_6\}$.

We then calculate the observed value of χ^2:

$$\chi^2_{obs} = \sum_{i=1}^{r'} \frac{(n_i - Np_i)^2}{Np_i} \tag{15-55}$$

as our measure of deviation of observation from prediction. Evidently, it is very unlikely that we would find $\chi^2_{obs} = 0$ even if the hypothesis is true.

So, goes the orthodox reasoning, we should calculate the probability that χ^2 would have various values, given H', and reject H if the probability of a deviation as great or greater than χ^2_{obs} is sufficiently small; usually one takes 5 per cent as the threshold of rejection.

Now the n_i are integers, so χ^2 is capable of taking on only a discrete set of numerical values, at most $(N+r-1)!/N!(r-1)!$ different values, if the p_i are all different and incommensurable. Therefore, the exact χ^2 distribution is necessarily discrete and defined at only a finite number of points. However, for sufficiently large N, the number and density of points becomes so large that we may approximate the χ^2 distribution by a continuous one. The "pleasant property" referred to by Cramér and Pratt, is then the fact, at first glance surprising, that in the limit of large N, we obtain a universal distribution law: the probability that χ^2 lies in the interval $d(\chi^2)$ is

$$g(\chi^2)\, d(\chi^2) = \frac{\chi^{f-2}}{2^{f/2}\left(\frac{f-2}{2}\right)!} \exp\left\{-\frac{1}{2}\chi^2\right\} d(\chi^2) \qquad (15\text{-}56)$$

where f is called the "number of degrees of freedom" of the χ^2-distribution. If the null hypothesis H was completely specified (i.e. if it contained no variable parameters), then $f = r' - 1$, where r' is the number of categories used in the sum of (15-55). But if H contains unspecified parameters which must be estimated from the data, we take $f = r' - 1 - m$, where m is the number of parameters estimated.

We readily calculate the expectation and variance of χ^2 over this distribution: $\langle\chi^2\rangle = f$, $\mathrm{var}(\chi^2) = 2f$; so if we were given H but didn't have the data of the experiment, the (mean ± standard deviation) estimate of the χ^2 we expect to see, would be just

$$(\chi^2)_{est} = f \pm \sqrt{2f} \qquad (15\text{-}57)$$

The reason usually given for grouping categories for which the sample numbers

are small, is that the approximation (15-56) would otherwise be bad. But grouping inevitably throws away some of the relevant evidence of the sample, and there is never any reason to do this when using ψ.

The probability that we would see a deviation as great or greater than χ^2_{obs} is then

$$P(\chi^2_{obs}) = \int_{\chi^2_{obs}}^{\infty} g(\chi^2)\, d(\chi^2)$$

$$= \int_{q_{obs}}^{\infty} \frac{q^k}{k!} e^{-q}\, dq \qquad (15\text{-}58)$$

where $q \equiv \frac{1}{2}\chi^2$, $k \equiv (f-2)/2$. If $P(\chi^2_{obs}) < 0.05$, we reject the null hypothesis at the 5% "significance level" (sometimes called the 95% level). Tables of χ^2_{obs} for which P = 0.01, 0.05, 0.10, 0.50 for various numbers of degrees of freedom, are given in most orthodox textbooks and collections of statistical tables.

Note the traditional procedure here: we choose some basically arbitrary significance level first, then report only whether the null hypothesis was or was not rejected at this level. Evidently, this doesn't tell us very much about the real import of the data; if you tell me that the hypothesis was rejected at the 5% level, then I don't know from this whether it would have been rejected at the 2%, or 1%, level. If you tell me it was not rejected at the 5% level, then I don't know whether is would have been rejected at the 10%, or 20%, level. The orthodox statistician would tell us far more about what the data really indicates if he would report instead the significance level $P(\chi^2_{obs})$ at which the null hypothesis is just barely rejected; for then we know what the verdict would be at all levels. But, for reasons totally incomprehensible to me, orthodox practice never does this, on the Chi-squared or any other significance test. In fact, the orthodox χ^2 and other tables are so constructed that you can't report the conclusions in

this more informative way, because they give numerical values only at such widely separated values of the significance level that interpolation isn't possible.

So, let me show you how to find numerical values of $P(\chi^2_{obs})$ from (15-58) without using the Chi-squared tables. Writing $q = q_o + t$, we have

$$P = \int_{q_o}^{\infty} \frac{q^k}{k!} e^{-q}\, dq = \int_0^{\infty} \frac{(q_o+t)^k}{k!} e^{-(q_o+t)}\, dt$$

$$= \frac{1}{k!} \sum_{m=0}^{k} \binom{k}{m} \int_0^{\infty} q_o^{\,m}\, t^{k-m}\, e^{-(q_o+t)}\, dt$$

$$= \sum_{m=0}^{k} e^{-q_o} \frac{q_o^{\,m}}{m!} \qquad (15\text{-}59)$$

But this is just the cumulative Poisson distribution; i.e. the probability

$$(m \le k | q_o) = \sum_{m=0}^{k} (m | q_o)$$

that $m \le k$, if m has a Poisson distribution with mean value $\langle m \rangle = q_o$:

$$(m | q_o) = e^{-q_o} \frac{q_o^{\,m}}{m!} \qquad (15\text{-}60)$$

Numerical values of (15-59) for all values of q_o, k of usual interest are given in the graph of the cumulative Poisson distribution in Appendix C. Use of this will somewhat improve the value of the Chi-squared test.

But if you use the ψ-test instead, you don't need any tables or graphs at all. The evidential meaning of the sample is then described simply by the numerical value of ψ; and not by a further arbitrary constraint such as tail areas. Of course, the numerical value of ψ doesn't in itself tell you whether to reject the hypothesis (although we could, with just as much justification as in the Chi-squared test, prescribe some definite "level" at which to reject). From the Bayesian point of view, there is simply no use in "rejecting" any hypothesis unless we can replace it with a definite alternative

known to be better; and whether this is justified must obviously depend not only on ψ, but also on the prior probability of the alternative (recall our quotation from E. L. Lehmann on p. 90), and on the consequences of making wrong decisions.

In spite of the difference in viewpoints, there is often not much difference in the actual conclusions reached. For example, as the number of degrees of freedom f increases, the orthodox statistician will accept a higher value of χ^2 [roughly proportional to f, as (15-57) indicates] before rejecting the hypothesis, on the grounds that such a high value is quite likely to occur if the hypothesis is true; but the Bayesian who will reject it only in favor of a definite alternative, must also accept a proportionally higher value of ψ, because the number of reasonable alternatives is increasing exponentially with f, and the prior probability of any one of them is correspondingly decreasing. So, in either case we reject the hypothesis if ψ or χ^2 exceeds some limit, with an enormous difference in the philosophy of how we choose that limit, but not necessarily a big difference in its actual location.

Although the point isn't made in the orthodox literature which just doesn't mention alternatives at all, we see from the above that χ^2 is <u>not</u> a measure of goodness of fit relative to all conceivable alternatives; but only relative to those in the same Bernoulli class. More generally, given any well-defined class C of alternatives, if we can write Bayes' theorem (describing the effect of new data D on the plausibility of two hypotheses H_1, H_2) in the form

$$e(H_1|DX) - e(H_1|X) = \psi_1 - \psi_2$$

where ψ_i depends only on the sample and H_i, is non-negative over C, and vanishes for some H_i in C, then we have constructed the appropriate ψ which

measures goodness of fit relative to the class of alternatives C.

In a recent article, Anscombe (1963) holds it to be a weakness of the Bayesian method that we had to introduce a specific class of alternatives. It seems to me, however, that it is entirely meaningless to speak of "goodness of fit" without reference to definite alternatives. For example, if you ask a scientist, "How well did the Zilch experiment fit the Bong theory?" you may get this reply: "Well, if you had asked me last week, I would have said it fits the Bong theory very handsomely; the experimental points lie much closer to Bong's curve than to the old Smith theory curve. But just yesterday I learned that this fellow Jones has worked out a new theory based on entirely different assumptions; and his curve goes right through the experimental points. So, now I'm afraid I have to say that the Zilch experiment pretty well demolishes the Bong theory."

Whether given data support or refute an hypothesis depends entirely on which alternatives we have in mind; if we fail to specify any alternatives we cannot hope to get a meaningful significance test, because we have not asked a well-posed question. The question when we should seek new alternatives must involve our knowledge about the "mechanism" being studied, and the line of reasoning which led to formulation of the null hypothesis in the first place; it cannot be answered merely from examining the null hypothesis and the sample. I would hold it to be a great merit of the Bayesian approach that it forces us to recognize these things, which have apparently not been obvious to statisticians (although qualitatively they are part of the elementary common sense which any scientist uses constantly in judging his theories).

This is a good example of what, I suggest, is the general situation; the Bayesian approach to statistics supplies the missing theoretical basis for, and often improvements on, orthodox methods which had long been, just as Pratt says, "ad hoc procedures with some pleasant properties."

15.5. An Acceptance Test.

Here is another very interesting example of a useful significance test. The probability that a certain machine will operate without failure for a time t is, by hypothesis, $\exp(-\lambda t)$. We test n units for a time t, and observe r failures; what assurance do we then have that the mean life $\theta = \lambda^{-1}$ exceeds a preassigned value θ_o? Let us examine the orthodox solution based on the same kind of philosophy that we just saw in the Chi-squared test (i.e. it is taboo to speak of the probability that θ has various values, because θ isn't a "random variable"; so we can use only the probability of getting various sample values, or the probability distribution of some "statistic"); and also give the Bayesian solution.

Sobel and Tischendorf (1959) (hereafter denoted ST) give an orthodox solution with tables that are reproduced in Roberts (1963). The test is to have a critical number C (i.e. we accept only if $r \leq C$). On the hypothesis that we have the maximum tolerable failure rate, $\lambda_o = \theta_o^{-1}$, the probability that we shall see r or fewer failures is the binomial sum

$$W(n,r) = \sum_{k=0}^{r} \binom{n}{k} e^{-(n-k)\lambda_o t} (1 - e^{-\lambda_o t})^k \qquad (15\text{-}61)$$

and so, setting $W(n,C) \leq 1 - P$ gives us the sample size n required in order that this test will assure $\theta \geq \theta_o$ at the 100 P per cent significance level. From the ST tables we find, for example, that if we wish to test only for a time $t = 0.01\ \theta_o$ with C = 3, then at the 90 per cent significance level we shall require a test sample of n = 668 units; while if we are willing to test for a time $t = \theta_o$ with C = 1, we need test only 5 units.

The amount of testing called for is appalling if $t \ll \theta_o$; and out of the question if the units are complete systems. For example, if we want to have 95 per cent confidence (synonymous with significance) that a space vehicle has $\theta_o \geq 10$ years, but the test must be made in six months, then

with C = 1, the ST tables say that we must build and test 97 vehicles! Suppose that, nevertheless, it had been decreed on the highest policy level that this degree of confidence must be attained, and you were in charge of the testing program. If a more careful analysis of the statistical problem, requiring a few man-years of statisticians' time, could reduce the test sample by only one or two units, it would be well justified economically. Scrutinizing the test more closely, we note four points:

(1) We know from the experiment not only the total number r of failures, but also the particular times $\{t_1 \ldots t_r\}$ at which failure occurred. This information is clearly relevant to the question being asked; but the ST test makes no use of it.

(2) The test has a "quasi-sequential" feature; if we adopt an acceptance number C = 3, then as soon as the fourth failure occurs, we know that the units are going to be rejected. If no failures occur, the required degree of confidence will be built up long before the time t specified in the ST tables. In fact, t is the maximum possible testing time, which is actually required only in the marginal case where we observe exactly C failures. A test which is "quasi-sequential" in the sense that it terminates when a clear rejection or the required confidence is attained, will have an expected length less than t; conversely, such a test with the expected length set at t will require fewer units tested.

(3) We have relevant prior information; after all, the engineers who designed the space vehicle knew in advance what degree of reliability was needed. They have chosen the quality of materials and components, and the construction methods, with this in mind. Each sub-unit has had its own tests. The vehicles would never have reached the final testing stage unless the engineers knew that they were operating satisfactorily. In other words, we are not testing a completely unknown entity. These facts constitute prior

information about the reliability, just as cogent as anything we can learn from a random experiment.

(4) In practice, we are usually concerned with a different question than the one the ST test answers. An astronaut starting a five-year flight to Mars would not be particularly comforted to be told, "We are 95 per cent confident that the average life of an imaginary population of space vehicles like yours, is at least ten years." He would much rather hear, "There is 95 per cent probability that this vehicle will operate without breakdown for ten years." Such a statement might appear meaningless to an orthodox statistician who holds that (probability) $\equiv$ (frequency). But such a statement would be very meaningful indeed to the astronaut. This is hardly a trivial point; for if it were known that $\lambda^{-1} = 10$ years, the probability that a particular vehicle will actually run for 10 years would be only $1/e = 0.368$; and the period for which we are 95 per cent sure of success would be only $-10 \ln(0.95)$ years, or 6.2 months. Reports which concern only the "mean life" can be rather misleading!

Let us first compare the ST test with a Bayesian test which makes use of exactly the same information; i.e. we are allowed to use only the total number of failures, not the actual failure times. On the hypothesis that the failure rate is λ, the probability that exactly r units fail in time t is

$$p(r|n,\lambda,t) = \binom{n}{r} e^{-(n-r)\lambda t} (1 - e^{-\lambda t})^r \qquad (15\text{-}62)$$

I want to defer discussion of nonuniform priors to a later section; for the time being suppose we assign a uniform prior to λ. This amounts to saying that, before the test, we consider it extremely unlikely that our space vehicles have a mean life as long as a microsecond; nevertheless it will be of interest to see the result of using this prior. The posterior distribution of λ is then

$$p(d\lambda|n,r,t) = \frac{n!}{(n-r-1)!\ r!}\, e^{-(n-r)\lambda t}\, (1 - e^{-\lambda t})^r\, d(\lambda t) \qquad (15\text{-}63)$$

The Bayesian acceptance criterion, which ensures $\theta \geq \lambda_o^{-1}$ with 100 P per cent probability, is then

$$\int_{\lambda_o}^{\infty} p(d\lambda|n,r,t) \leq 1 - P \qquad (15\text{-}64)$$

But the left-hand side of (15-64) is identical with W(n,r) given by (15-61); this is just the well-known identity of the incomplete Beta function and the incomplete binomial sum, given already in the original memoir of Bayes (1762). In this first comparison we therefore find that the ST test is mathematically identical with a Bayesian test in which (1) we are denied use of the actual failure times; (2) because of this it is not possible to take advantage of the quasi-sequential feature; (3) we assign a ridiculously pessimistic prior to λ; (4) we still are not answering the question of real interest for most applications.

Of these shortcomings, (2) is readily corrected, and (1) undoubtedly could be corrected, without departing from orthodox principles. On the hypothesis that the failure rate is λ, the probability that r specified units fail in the time intervals $\{dt_1 \ldots dt_r\}$ respectively, and the remaining (n-r) units do not fail in time t, is

$$p(dt_1 \ldots dt_r|n,\lambda,t) = [\lambda^r\, e^{-\lambda r\bar{t}}\, dt_1 \ldots dt_r][e^{-(n-r)\lambda t}] \qquad (15\text{-}65)$$

where $\bar{t} \equiv r^{-1} \sum t_i$ is the mean life of the units which failed. There is no single "statistic" which conveys all the relevant information; but r and $\bar{t}$ are jointly sufficient, and so an optimal orthodox test must somehow make use of both. When we seek their joint sampling distribution $p(r,d\bar{t}|n,\lambda,t)$ we find, to our dismay, that for given r the interval $0 < \bar{t} < t$ is broken up into r equal intervals, with a different analytical expression for each. Evidently a decrease in r, or an increase in $\bar{t}$, should incline us in the

direction of acceptance; but at what rate should we trade off one against the other? To specify a definite critical region in both variables would seem to imply some postulate as to their relative importance. The problem does not appear simple, either mathematically or conceptually; and I would not presume to guess how an orthodox statistician would solve it.

The relative simplicity of the Bayesian analysis is particularly striking in this problem; for all four of the above shortcomings are corrected effortlessly. For the time being, we again assign the pessimistic uniform prior to λ; from (15-65), the posterior distribution of λ is then

$$p(d\lambda \mid n,t,t_1 \ldots t_r) = \frac{(\lambda T)^r}{r!} e^{-\lambda T} \, d(\lambda T) \tag{15-66}$$

where

$$T \equiv r\bar{t} + (n-r)t \tag{15-67}$$

is the total unit-hours of failure-free operation observed. The posterior probability that $\lambda \geq \lambda_o$ is now

$$B(n,r) = \frac{1}{r!} \int_{\lambda_o T}^{\infty} x^r e^{-x} \, dx = e^{-\lambda_o T} \sum_{k=0}^{r} \frac{(\lambda_o T)^k}{k!} \tag{15-68}$$

and so, $B(n,r) \leq 1 - P$ is the new Bayesian acceptance criterion at the 100 P per cent level; the test can terminate with acceptance as soon as this inequality is satisfied.

Numerical analysis shows little difference between this test and the ST teat in the usual range of practical interest, where we test for a time short compared to θ_o and observe only a very few failures. For, if $\lambda_o t << 1$, and $r << n$, then the Poisson approximation to (15-61) will be valid (as in Lecture 8); but this is just the expression (15-68) except for the replacement of T by nt, which is itself a good approximation. In this region the Bayesian test (15-68) with maximum possible duration t generally calls for a test sample one or two units smaller than the ST test. Our common sense readily

assents to this; for if we see only a few failures, then information about the actual failure time adds little to our state of knowledge.

Now let us magnify. The big differences between (15-61) and (15-68) will occur when we find many failures; if all n units fail, the ST test tells us to reject at all confidence levels, even though the observed mean life may have been thousands of times our preassigned θ_o. The Bayesian test (15-68) does not break down in this way; thus if we test 9 units and all fail, it tells us to accept at the 90 per cent level if the observed mean life $\bar{t} \geq 1.58\ \theta_o$. If we test 10 units and 9 fail, the ST test says we can assert with 90 per cent confidence that $\theta \geq 0.22t$; the Bayesian test (15-68) says there is 90 per cent probability that $\theta \geq 0.63\ \bar{t} + 0.07\ t$. Our common sense has no difficulty in deciding which result we should prefer; thus taking the actual failure times into account leads to a clear, although usually not spectacular, improvement in the test. The person who rejects the use of Bayes' theorem in the manner of Eq. (15-66) will be able to obtain a comparable improvement only with far greater difficulty.

But the Bayesian test (15-68) can be further improved in two respects. To correct shortcoming (4), and give a test which refers to the reliability of the individual unit instead of the mean life of an imaginary "population" of them, we note that if λ were known, then by our original hypothesis the probability that the lifetime θ of a given unit is at least θ_o, is

$$p(\theta \geq \theta_o | \lambda) = e^{-\lambda \theta_o} \tag{15-69}$$

The probability that $\theta \geq \theta_o$, conditional on the evidence of the test, is therefore

$$p(\theta \geq \theta_o | n, t_1 \ldots t_r) = \int_0^\infty e^{-\lambda \theta_o}\, p(d\lambda | n, t_1 \ldots t_r)$$

$$= \left(\frac{T}{T + \theta_o} \right)^{r+1} \tag{15-70}$$

Thus, the Bayesian test which ensures, with 100 P per cent probability, that the life of an individual unit is at least θ_o, has an acceptance criterion that the expression (15-70) is $\geq$ P; a result which is simple, sensible, and as far as I can see, utterly beyond the reach of orthodox statistics.

The Bayesian tests (15-68) and (15-70) are, however, still based on a ridiculous prior for λ; another improvement, even further beyond the reach of orthodox statistics, will be found presently, as a result of using a reasonable prior.

Lecture 16

THE A_p DISTRIBUTION AND RULE OF SUCCESSION

Up to this point we have given our robot fairly general principles by which he can convert information into numerical values of prior probabilities, and convert posterior probabilities into definite final decisions; so he is now able to solve lots of problems. But he still operates in a rather inefficient way in one respect. When we give him new information and ask him to reason about it, he has to go back into his memory (this proposition X that involves everything that has ever happened to him). He must scan his entire memory storage reels for anything relevant to the problem before he can start reasoning on it. As the robot gets older this gets to be a more and more time-consuming process.

Now, human brains don't do this. We have some machinery built into us which summarizes our past conclusions, and allows us to forget the details which led us to those conclusions. We want to see whether it's possible to give the robot a definite mechanism by which he can store conclusions rather than isolated facts.

16.1. Memory Storage for Old Robots.

Let me point out another thing, which we will see is closely related to this problem. Suppose you have a penny and you are allowed to examine it carefully, convince yourself that it's an honest coin, has a head and tail, and center of gravity where it ought to be. Then, you're asked to give the

probability that this coin will come up heads on the first toss. I'm sure you'll say 1/2. Now, suppose you are asked to assign a probability to the proposition that there is life on Mars. Well, I don't know what your opinion is there, but on the basis of all the things that I have read on the subject, I would again say about 1/2 for the probability. But, even though I have assigned the same probability to them, I have a very different state of knowledge about those propositions. To see that, imagine the effect of getting new information. Suppose we tossed the coin five times and it comes up tails every time. You ask me what's my probability for heads on the next throw; I'll still say 1/2. But if you tell me one more fact about Mars, I'm ready to change my probability assignment completely. My state of belief has a great instability in the case of Mars, but there's something which makes it very stable in the case of the penny.

Now, it seemed to me for a long time that this was a fatal objection to Laplace's form of probability theory. We need to associate with a proposition not just a single number representing plausibility, but two numbers; one representing the plausibility, and the other how stable it is in the face of new evidence. And so, a kind of two-valued theory would have to be developed before it would make any sense. In the early 1950's, I even gave a talk at one of the Berkeley Statistical Symposiums, expounding this viewpoint. This is, furthermore, just what Carnap (1952) has done; his continuum of inductive methods consists of a class of probability functions $C_\lambda(h,e)$ in which λ is the "stability parameter."

But now, I think that there's a mechanism by which we can show that our present theory automatically contains all these things. So far, all the propositions we have asked the robot to think about are ones which had to be either true or false. Suppose we bring in new propositions of a different type. It doesn't make sense to say the proposition is either true or false,

but still we are going to say the robot assigns credibility to it. Now, these propositions are sometimes hard to state verbally, and I, at least, am never able to write a verbal statement that's unambiguous. But you noticed before that we can get around that very nicely by recognizing that if I state all probabilities conditional on X for a given problem, I've told you everything about X that's relevant to the problem. So, I want to introduce a new proposition A_p, defined by

$$(A|A_pE) \equiv p \tag{16-1}$$

where E is any additional evidence. If I had to render A_p as a verbal statement, it would come out something like this:

$A_p \equiv$ "Regardless of anything else you may have been told, the probability of A is p."

Now, A_p is a strange proposition, but if we allow the robot to reason with propositions of this sort, Bayes' theorem guarantees that there's nothing to prevent him from getting an A_p worked over onto the left side in his probabilities: $(A_p|E)$. Now, what are we doing here? We're talking about the "probability of a probability." I defined A_p by writing an equation. You ask me what it means, and I reply by writing more equations. So let's write the equations; if X says nothing about A except that it is possible for A to be true, and also possible for it to be false, then as we saw in the case of the "completely ignorant population" in Lecture 12,

$$(A_p|X) = 1, \qquad 0 \le p \le 1. \tag{16-2}$$

The transformation group arguments of Lecture 12 apply to this problem. As soon as we have this, we can use Bayes' theorem to get the probability (density) of A_p, conditional on other things. In particular,

$$(A_p|E) = (A_p|X)\,\frac{(E|A_p)}{(E|X)} = \frac{(E|A_p)}{(E|X)} \tag{16-3}$$

Now,

$$(A|E) = \int_0^1 (AA_p|E)\, dp \tag{16-4}$$

The propositions A_p are mutually exclusive and exhaustive (in fact, every A_p flatly and dogmatically contradicts every other A_q), so we can do this. We're just going to apply all of our mathematical rules with total disregard of the fact that A_p is a funny kind of proposition. We believe that these rules form a consistent way of manipulating propositions; their application cannot lead to contradictions. (Of course, we haven't really proved that they are consistent; we have proved only that if we represent degrees of plausibility by real numbers and require qualitative agreement with common sense, any other rules would be inconsistent.) But consistency is a purely structural property of the rules, which could not depend on the particular semantic meaning you or I might attach to a proposition. So now we can blow up the integrand of (16-4) by our Rule 1:

$$(A|E) = \int_0^1 (A|A_pE)(A_p|E)\, dp \tag{16-5}$$

But from the definition (16-1) of A_p, the first factor is just p, and so

$$(A|E) = \int_0^1 (A_p|E)\, p\, dp \tag{16-6}$$

The probability which our robot assigns to proposition A is just the <u>first moment</u> of the distribution of A_p. Therefore, the distribution of A_p should contain an awful lot more information about the robot's state of mind concerning A, than just the probability of A. I think the introduction of propositions of this sort solves both of the problems mentioned, and also gives us a powerful analytical tool for calculating probabilities.

To see why, let's first note some lemmas about relevance. Suppose this evidence E consists of two parts; $E = E_aE_b$, where E_a is relevant to A and, given E_a, E_b is not relevant:

$$(A|E) = (A|E_aE_b) = (A|E_a) \tag{16-7}$$

By Bayes' theorem, it follows that, given E_a, A must also be irrelevant to E_b, for

$$(E_b|AE_a) = (E_b|E_a)\,\frac{(A|E_bE_a)}{(A|E_a)} = (E_b|E_a) \tag{16-8}$$

Let's call this property "weak irrelevance." Now does this imply that E_b is irrelevant to A_p? Evidently not, for (16-7) says only that the first moments of $(A_p|E_a)$ and $(A_p|E_aE_b)$ are the same. But suppose that for a given E_b, (16-7) holds independently of what E_a might be; call this "strong irrelevance." Then we have

$$(A|E) = \int_0^1 (A_p|E_aE_b)\; p\; dp = \int_0^1 (A_p|E_a)\; p\; dp. \tag{16-9}$$

If this is to hold for all $(A_p|E_a)$, the integrands must be the same

$$(A_p|E_aE_b) = (A_p|E_a) \tag{16-10}$$

and from Bayes' theorem it follows as in (16-8) that A_p is irrelevant to E_b:

$$(E_b|A_pE_a) = (E_b|E_a) \tag{16-11}$$

for all E_a.

Now, suppose our robot gets a new piece of evidence, F. How does this change his state of knowledge about A? We could expand directly by Bayes' theorem, which we have done before, but let's use our A_p this time,

$$(A|EF) = \int_0^1 (A_p|EF)\; p\; dp = \int_0^1 (A_p|E)\,\frac{(F|A_pE)}{(F|E)}\; p\; dp\;. \tag{16-12}$$

In this likelihood ratio, any part of E that is irrelevant to A_p can be struck out. Because, by Bayes' theorem, it is equal to

$$\frac{(F|A_pE_aE_b)}{(F|E_aE_b)} = \frac{(F|A_pE_a)\left[\dfrac{(E_b|FA_pE_a)}{(E_b|A_pE_a)}\right]}{(F|E_a)\left[\dfrac{(E_b|FE_a)}{(E_b|E_a)}\right]} = \frac{(F|A_pE_a)}{(F|E_a)} \tag{16-13}$$

where we have used (16-11). Now if E_a still contains a part irrelevant to A_p, we can repeat this process. Imagine this carried out as many times as

possible; the part E_{aa} of E that is left contains nothing at all that is irrelevant to A_p. E_{aa} must then be some statement only about A. But then by the definition (16-1) of A_p, we see that A_p automatically cancels out E_{aa} in the numerator: $(F|A_pE_{aa}) = (F|A_p)$. And so we have (16-12) reduced to

$$(A|EF) = \frac{1}{(F|E_{aa})} \int_0^1 (A_p|E)(F|A_p)\ p\ dp \qquad (16\text{-}14)$$

The weak point in this argument is that I haven't proved that it is possible to resolve E into a completely relevant part and completely irrelevant part. However, it is easy to show that in many applications it <u>is</u> possible. So, let's just say that the following results apply to the case where the prior information is "completely resolvable." We don't know whether it is the most general case; but we do know that it is not an empty one.

Now, $(F|E_{aa})$ is a troublesome thing which we would like to get rid of. It's really just a normalizing factor, and we can eliminate it the way we did in Equation (5-3); by calculating the odds on A instead of the probability. This is just

$$\frac{(A|EF)}{(a|FE)} = \frac{\int_0^1 (A_p|E)(F|A_p)\ p\ dp}{\int_0^1 (A_p|E)(F|A_p)\ (1\text{-}p)\ dp} = O(A|EF) \qquad (16\text{-}15)$$

The proposition E, which for this problem represents our prior evidence, now appears only in the combination $(A_p|E)$. This means that <u>the only property of E which the robot needs in order to reason out the effect of new information is this distribution</u> $(A_p|E)$. Everything that has ever happened to him which is relevant to this proposition A may consist of millions and millions of isolated separate facts. Whenever he receives new information, he does not have to go back and search his entire memory for every little detail of experience relevant to A. Everything he needs in order to reason about it is contained summarized in this one function, $(A_p|E)$. So, for each proposi-

tion about which he is going to have to reason, he can store a function like that in Figure (16.1). Whenever he receives new information, F, he will be well advised to calculate $(A_p|EF)$, and he then can erase his previous $(A_p|E)$ and for the future store only $(A_p|EF)$.

This shows that in a machine which does inductive reasoning, the memory storage problem is very much simpler than it is in a machine which does only deductive reasoning, like this one you have down at the end of the hall. This doesn't mean that the robot is able to throw away entirely all of his past experience, because there's always a possibility that some new proposition will come up which he has not had to reason about before. And whenever this happens, then, of course, he <u>will</u> have to go back to his original archives and search for every scrap of information he has relevant to this proposition.

With a little introspection, I think we would all agree that that's exactly what goes on in our minds. If you are asked how plausible you regard

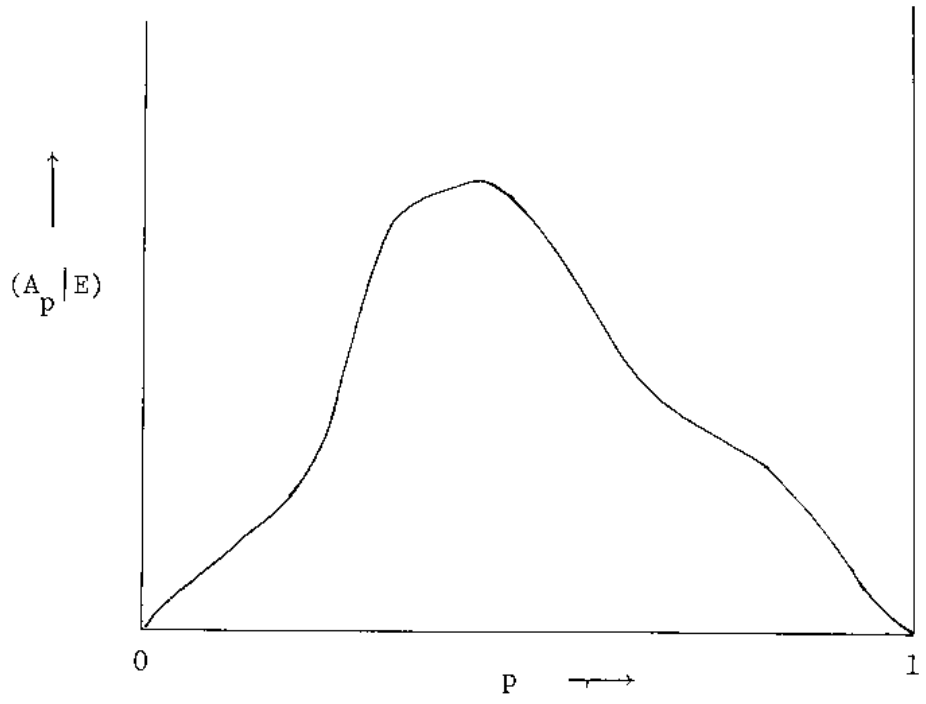

Figure 16.1

some proposition, you don't go back and recall all the details of everything that you ever learned about this proposition. You recall your previous state of mind about it. How many of us can still remember the argument that first convinced us that

$$\frac{d \sin x}{dx} = \cos x \quad ?$$

Let's look once more at Equation (16-14). If the new information F is to make any appreciable change in the probability of A, we can see from this integral what has to happen. If the distribution of $(A_p|E)$ was already very sharply peaked at one particular value of p, then $(F|A_p)$ will have to be even more sharply peaked at some other value of p, if we are going to get any appreciable change in the probability. On the other hand, if the distribution $(A_p|E)$ is a very broad one, then, of course, almost any small amount of slope in $(F|A_p)$ can make a big change in the probability which the robot assigns to A. So, the stability of the robot's state of mind is essentially the <u>width</u> of the distribution $(A_p|E)$. I don't think there's any single number which fully describes this stability. On the other hand, whenever he has accumulated enough evidence so that $(A_p|E)$ is fairly well sharply peaked at some value of p, then the variance of that distribution becomes a pretty good measure of how stable his state of mind is. The greater amount of previous information he has collected, the narrower his A_p-distribution will be, and therefore the harder it will be for any new evidence to change that state of mind.

Now we can see the difference between the penny and Mars. In the case of the penny, my distribution $(A_p|E)$, based on my prior knowledge, is represented by a curve something like Figure (16.2a). In the case of the question of life on Mars, my state of knowledge is described by an $(A_p|E)$ distribution something like Figure (16.2b), qualitatively. The first moment is the same

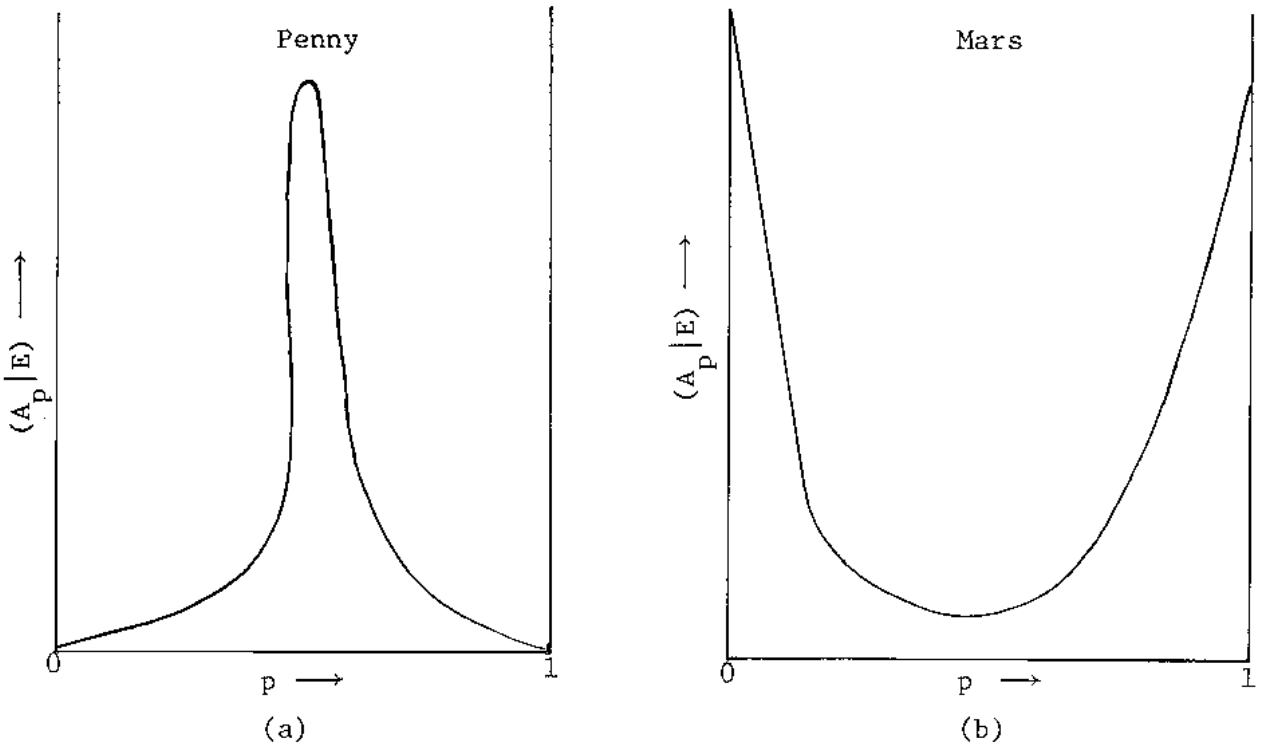

Figure 16.2

in the two cases. So, I assign probability 1/2 to either one; nevertheless, there's all the difference in the world between my state of knowledge about those two propositions, and this difference is represented in the distribution of $(A_p|E)$.

Now, incidentally, I might mention an amusing thing. While I was first working some of this out, a newspaper story showed up from which I would like to read you a few sentences. This is from the Associated Press, December 14, 1957, entitled, "Brain Stockpiles Man's Most Inner Thoughts." It starts out: "Everything you have ever thought, done, or said--a complete record of every conscious moment--is logged in the comprehensive computer of your brain. You will never be able to recall more than the tiniest fraction of it to memory, but you'll never lose it either. These are the findings of Dr. Wilder Penfield, Director of the Montreal Neurological Institute, and a leading Neurosurgeon. The brain's ability to store experiences, many

lying below consciousness, has been recognized for some time, but the extent of this function is recorded by Dr. Penfield."

Now there are several examples given, of experiments on patients suffering from epilepsy. Stimulation of a definite location in the brain recalled a definite experience from the past, which the patient had not been previously able to recall to memory. This has happened many times. I'm sure you have all read about these things. Here are the concluding sentences of this article. Dr. Penfield now says, "This is not memory as we usually use the word, although it may have a relation to it. No man can recall by voluntary effort such a wealth of detail. A man may learn a song so he can sing it perfectly, but he cannot recall in detail any one of the many times he heard it. Most things that a man is able to recall to memory are generalizations and summaries. If it were not so, we might find ourselves confused by too great a richness of detail."

16.2. An Application.

Now let's imagine that a "random" experiment is being performed. From the results of the experiment in the past, we want to do the best job we can of predicting results in the future. To make the problem a definite one, introduce the propositions:

X ≡ "For each trial we admit two prior hypotheses: A true, and A false. The underlying 'causal mechanism' is assumed the same at every trial. This means, for example, that (1) the probability assigned to A at the n'th trial does not depend on n, and (2) evidence concerning the results of past trials retains its relevance for all time; thus for predicting the outcome of trial 1,000, knowledge of the result of trial 1 is just as

relevant as knowledge of the result of trial 999.

There is no other prior evidence.

$N_n \equiv$ "A true n times in N trials in the past."

$M_m \equiv$ "A true m times in M trials in the future."

The verbal statement of X suffers from just the same ambiguities that we have found before, and which have caused so much trouble and controversy in the past. One of the important points I want to put across in these talks is that you have not given any precise description of the prior information until you have given, not verbal statements, but equations, which specify the prior probabilities to be used. In the present problem, this more precise statement of X is, as before

$$(A_p|X) = 1 \quad , \qquad 0 \le p \le 1 \tag{16-16}$$

with the additional understanding that the same A_p-distribution is to be used for calculations pertaining to all trials. What we are after is $(M_m|N_n)$. First, note that by many repetitions of our Rule 1 and Rule 2, in the same way that we found Equation (5-34), we have the binomial distributions

$$\begin{aligned} (N_n|A_p) &= \binom{N}{n} p^n (1-p)^{N-n} \\ (M_m|A_p) &= \binom{M}{m} p^m (1-p)^{M-m} \quad . \end{aligned} \tag{16-17}$$

Note that, although A_p sounds like an awfully dogmatic and indefensible statement to us the way we've introduced it, this is actually the way in which probability is introduced in almost all present testbooks. One postulates that an event possesses some intrinsic, "absolute" or "physical" probability, whose numerical value we can never determine exactly. Nevertheless, no one questions that such an "absolute" probability exists. Cramér (1946, p. 154), for example, takes it as his fundamental axiom. That is just as dogmatic a statement as our A_p; and I think it is, in fact, just our A_p. The equations you see in current textbooks are all like the two I have just

written; whenever p appears as a given number, there's an A_p hiding in the right-hand of your probability symbols.

Mathematically, the only difference between what we're doing here and what is done in current textbooks is that we recognize the existence of that right-hand side for all probabilities, and we are not afraid to use Bayes' theorem to work any proposition whatsoever back and forth from one side of our symbols to the other. I think that in refusing to make free use of Bayes' theorem, modern writers are depriving themselves of the most powerful single principle in probability theory. When a problem of statistical inference is studied long enough, sometimes for decades, one is always forced eventually to a conclusion that could have been derived in three lines from Bayes' theorem. We saw this in the quality-control example and in the case of decision theory; and we'll see several more examples in the remainder of these talks.

Now, we need to find the prior probability $(N_n|X)$. This is already determined from $(A_p|X)$, for our trick of resolving a proposition into mutually exclusive alternatives gives us

$$(N_n|X) = \int_0^1 (N_n A_p|X)\, dp = \int_0^1 (N_n|A_p)(A_p|X)\, dp = \binom{N}{n} \int_0^1 p^n (1-p)^{N-n}\, dp \quad .$$

The integral we have to evaluate is the complete Beta-function:

$$\int_0^1 x^r (1-x)^s\, dx = \frac{r!\, s!}{(r+s+1)!} \tag{16-18}$$

Thus, we have

$$(N_n|X) = \begin{cases} \frac{1}{N+1}, & 0 \le n \le N \\ 0, & N < n \end{cases} \tag{16-19}$$

i.e., just the uniform distribution of maximum entropy. $(M_m|X)$ is similarly found. Now we can turn (16-17) around by Bayes' theorem:

$$(A_p|N_n) = (A_p|X)\, \frac{(N_n|A_p)}{(N_n|X)} = (N+1)(N_n|A_p) \tag{16-20}$$

and so finally the desired probability is

$$(M_m|N_n) = \int_0^1 (M_m A_p|N_n)\, dp = \int_0^1 (M_m|A_p N_n)(A_p|N_n)\, dp \quad . \qquad (16\text{-}21)$$

Since $(M_m|A_p N_n) = (M_m|A_p)$ by the definition of A_p, we have everything in the integrand on the board. Substituting into (16-21), we have again an Eulerian integral, and our result is

$$(M_m|N_n) = \frac{\binom{n+m}{n}\binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}} \qquad (16\text{-}22)$$

Note that this is <u>not</u> the same as the hypergeometric distribution (5-23) of sampling theory. Let's look at this result first in the special case M = m = 1. It will then reduce to the probability of A being true in the next trial, given that it had been true n times in the previous N trials. The result is

$$(A|N_n) = \frac{n+1}{N+2} \quad . \qquad (16\text{-}23)$$

This is <u>Laplace's rule of succession</u>. It occupies a supreme position in probability theory; it has been easily the most misunderstood and misapplied rule in the theory, from the time Laplace first gave it in 1774. In almost any book on probability you'll find this rule mentioned very briefly, mainly in order to warn the reader not to use it. But we've got to take the trouble to understand it because in our design of this robot, Laplace's rule of succession is, like Bayes' theorem, one of the most important rules we have. It is a new rule for converting raw information into numerical values of probabilities, and it gives us one of the most important connections between probability and frequency.

<u>16.3. Laplace's Rule of Succession</u>.

Poor old Laplace has been lampooned for generations because he illustrated use of this rule by calculating the probability that the sun will rise tomorrow, given that it has risen every day for the past 5,000 years. One gets a

rather large factor in favor of the sun rising again tomorrow, of course. With no exceptions at all as far as I know, modern writers on probability have considered this a pure absurdity. Even Jeffreys and Carnap find fault with the rule of succession.

I have to confess to you that I am unable to see anything at all absurd about the rule of succession. I recommend very strongly that you do a little literature searching, and read some of the objections various writers have to it. I think you will see that in every case the same thing has happened. First, Laplace was quoted out of context, and secondly, in order to demonstrate the absurdity of the rule of succession, the author applies it to a case where it was never intended to be applied, because there is additional prior information which was not taken into account.

If you go back and read Laplace (1819) himself, you will see that in the very next sentence after this sunrise episode, he points out to the reader that this is the probability based only on the information that the event has occurred n times in N trials, and that our knowledge of celestial mechanics represents a great deal of additional information. Of course, if you have additional information beyond the numbers n and N, then you ought to take it into account. You are then considering a different problem, the rule of succession no longer applies, and you can get an entirely different answer. This theory gives the results of consistent plausible reasoning on the basis of the information which was put into it.

Let me give you three famous examples of the kind of objections to the rule of succession which you find in the literature. (1) Suppose the solidification of hydrogen to have been once accomplished. According to the rule of succession, the probability that it will solidify again if the experiment is repeated is 2/3. This does not in the least represent the state of belief of any scientist. (2) A boy is 10 years old today. According to the rule

of succession, he has the probability 11/12 of living one more year. His grandfather is 70; and so according to this rule he has the probability 71/72 of living one more year. The rule violates qualitative common sense! (3) Consider the case N = n = 0. It then says that any conjecture without any verification has the probability 1/2. Thus there is probability 1/2 that there are exactly 137 elephants on Mars. Also there is probability 1/2 that there are 138 elephants on Mars. Therefore, it is certain that there are at least 137 elephants on Mars. But the rule says also that there is probability 1/2 that there are no elephants on Mars. The rule is logically self-contradictory!

The trouble with examples (1) and (2) is obvious in view of our earlier remarks; in each case, an enormous amount of highly relevant prior information, known to all of us, was simply ignored, producing a flagrant misuse of the rule of succession. But let's look a little more closely at example (3). Wasn't the law applied correctly here? I certainly can't claim that we had prior information about elephants on Mars which was ignored, can I? And even if I could, that still wouldn't account for the self-contradiction. Evidently, if the rule of succession is going to survive example (3), there must be some very basic points about the use of probability theory which we still have to learn.

Well, now, what do we mean when we say that there's no evidence for a proposition? The question is not what you or I might mean colloquially by such a statement. The question is, what does it mean to the robot? What does it mean in terms of probability theory?

The prior information we used in derivation of the rule of succession was that the robot is told that there are only two possibilities: A true, and A false. His entire "universe of discourse" consists of only two propositions. In the case N = 0, we could solve the problem also by direct appli-

cation of the principle of indifference, Rule 4; and this will of course give the same answer $(A|X) = 1/2$, that we got from the rule of succession. But just by noting this, we see what is wrong. Merely by admitting the possibility of three different propositions being true, instead of only two, we have already specified prior information different from that used in deriving the rule of succession.

If the robot is told to consider 137 different ways in which A could be false, and only one way in which it could be true, then the prior probability of A is 1/138, not 1/2. So, we see that the example of the elephants on Mars was, again, a gross misapplication of the rule of succession.

Moral: Probability theory, like any other mathematical theory, cannot give us a definite answer unless we ask it a definite question. We should always start a problem with an explicit enumeration of the different propositions we're going to consider. That is part of the "boundary conditions" which must be specified before we have a uniquely defined mathematical problem. If you say, "I don't know what the possible propositions are," that is mathematically equivalent to saying, "I don't know what problem I want to solve." This is just the point that I have already belabored back in Lecture 7.

In this connection we have to remember that probability theory never solves problems of actual practice, because all such problems are infinitely complicated. We solve only idealizations of the real problem, and the solution is useful to the extent that the idealization is a good one. In the example of the solidification of hydrogen, the prior information which our common sense uses so easily, is actually so complicated that nobody knows how to convert it into a prior probability assignment. I don't think there is any reason to doubt that probability theory is, in principle, competent to deal with such problems; but we have not yet learned how to translate them into

mathematical language without oversimplifying so much that the solution is useless.

Laplace's rule of succession provides a definite solution to a definite problem. Everybody denounces it as nonsense because it is not also the solution to some other problem. The case where the problem can be reasonably idealized to one with only two hypotheses to be considered, a belief in a constant "causal mechanism," and no other prior information, is the only case where it applies. You can, of course, generalize it to any number of hypotheses, and let me just give you the result of doing this.

There are K different hypotheses, $\{A_1, A_2, \ldots, A_K\}$, a belief that the "causal mechanism" is constant, and no other prior information. We perform a random experiment N times, and observe A_1 true n_1 times, A_2 true n_2 times, etc. Of course, $\sum_i n_i = N$. On the basis of this evidence, what is the probability that in the next $M = \sum_i m_i$ repetitions of the experiment, A_i will be true exactly m_i times? To find the distribution $(m_1 \ldots m_K | n_1 \ldots n_K)$ that answers this, define the prior knowledge by a K-dimensional uniform prior distribution

$$(A_{p_1 \cdots p_K} | X) = C\, \delta(p_1 + \ldots + p_K - 1) \quad , \quad p_i \geq 0 \tag{16-24}$$

To find the normalization constant C, we set

$$\int_0^\infty dp_1 \ldots \int_0^\infty dp_K \; (A_{p_1 \cdots p_K} | X) = 1 = C\, I(1) \tag{16-25}$$

where

$$I(r) \equiv \int_0^\infty dp_1 \ldots \int_0^\infty dp_K \; \delta(p_1 + \ldots + p_K - r) \tag{16-26}$$

Direct evaluation of this would be rather messy, so let's use the following trick. First, take the Laplace transform of (16-26)

$$\int_0^\infty e^{-\alpha r} I(r)\, dr = \int_0^\infty dp_1 \ldots \int_0^\infty dp_K \; e^{-\alpha(p_1 + \ldots + p_K)} = \frac{1}{\alpha^K} \; . \tag{16-27}$$

Or, inverting the Laplace transform,

$$I(r) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \frac{e^{\alpha r}}{\alpha^K}\, d\alpha = \frac{1}{(K-1)!}\frac{d^{K-1}}{d\alpha^{K-1}} e^{\alpha r}\Big|_{\alpha=0}$$

$$= \frac{r^{K-1}}{(K-1)!} \tag{16-28}$$

Thus,

$$C = \frac{1}{I(1)} = (K-1)! \tag{16-29}$$

By this device, we avoided having to consider complicated details about different ranges of integration over the different p_i, that would come up if we tried to evaluate (16-26) directly.

The prior distribution $(n_1 \ldots n_K | X)$ is then, using the same trick,

$$(n_1 \ldots n_K | X) = \frac{N!}{n_1! \ldots n_K!}\int_0^\infty dp_1 \ldots \int_0^\infty dp_K\, p_1^{n_1} \ldots p_K^{n_K}\, (A_{p_1 \ldots p_K} | X)$$

$$= \frac{N!\ (K-1)!}{n_1! \ldots n_K!}\, J(1) \tag{16-30}$$

where

$$J(r) \equiv \int_0^\infty dp_1 \ldots \int_0^\infty dp_K\, p_1^{n_1} \ldots p_K^{n_K}\, \delta(p_1 + \ldots + p_K - r) \tag{16-31}$$

which we evaluate as before by taking the Laplace transform:

$$\int_0^\infty e^{-\alpha r}\, J(r)\, dr = \int_0^\infty dp_1 \ldots \int_0^\infty dp_K\, p_1^{n_1} \ldots p_K^{n_K}\, e^{-\alpha(p_1 + \ldots + p_K)}$$

$$= \prod_{i=1}^K \frac{n_i!}{\alpha^{n_i+1}} \tag{16-32}$$

So, as in (16-28), we have

$$J(r) = \frac{n_1! \ldots n_K!}{2\pi i}\int_{-i\infty}^{i\infty} \frac{e^{\alpha r}\, d\alpha}{\alpha^{N+K}} = \frac{n_1! \ldots n_K!}{(N+K-1)!}\, r^{N+K-1} \tag{16-33}$$

and

$$(n_1 \ldots n_K | X) = \frac{N!(K-1)!}{(N+K-1)!}, \qquad n_i \geq 0,\ n_1+\ldots+n_K = N \tag{16-34}$$

Therefore, by Bayes' theorem

$$(A_{p_1 \ldots p_K} | n_1 \ldots n_K) = (A_{p_1 \ldots p_K} | X) \frac{(n_1 \ldots n_K | A_{p_1 \ldots p_K})}{(n_1 \ldots n_K | X)}$$

$$= \frac{(N+K-1)!}{n_1! \ldots n_K!} p_1^{n_1} \ldots p_K^{n_K} \delta(p_1+\ldots+p_K-1) \tag{16-35}$$

and finally

$$(m_1 \ldots m_K | n_1 \ldots n_K) = \int_0^\infty dp_1 \ldots \int_0^\infty dp_K \, (m_1 \ldots m_K | A_{p_1 \ldots p_K}) (A_{p_1 \ldots p_K} | n_1 \ldots n_K)$$

$$= \frac{M!}{m_1! \ldots m_K!} \frac{(N+K-1)!}{n_1! \ldots n_K!} \int_0^\infty dp_1 \ldots \int_0^\infty dp_K \, p_1^{n_1+m_1} \ldots p_K^{n_K+m_K} \delta(p_1+\ldots+p_K-1) \tag{16-36}$$

The integral is the same as J(1) except for the replacement $n_i \to n_i+m_i$. So, from (16-33),

$$(m_1 \ldots m_K | n_1 \ldots n_K) = \frac{M!}{m_1! \ldots m_K!} \frac{(N+K-1)!}{n_1! \ldots n_K!} \frac{(n_1+m_1)! \ldots (n_K+m_K)!}{(N+M+K-1)!} \tag{16-37}$$

or, reorganizing into binomial coefficients,

$$(m_1 \ldots m_K | n_1 \ldots n_K) = \frac{\binom{n_1+m_1}{n_1} \ldots \binom{n_K+m_K}{n_K}}{\binom{N+M+K-1}{M}} . \tag{16-38}$$

In the case where we want just the probability that A_1 will be true on the next trial, we need this formula with $M = m_1 = 1$, all other $m_i = 0$. The result is the generalized law of succession:

$$(A_1 | n_1, N, K) = \frac{n_1 + 1}{N + K} . \tag{16-39}$$

You see that in the case $N = n_1 = 0$, this reduces to the answer provided by the principle of indifference, Rule 4, which it therefore contains as a

special case. If K is a power of 2, this is the same as a method of inductive reasoning proposed by Carnap in 1945, which he denotes as c*(h,e) in his "Continuum of Inductive Methods."

Now, use of the rule of succession in cases where N is very small is rather foolish, of course. Not really wrong; just foolish. Because if we have no prior evidence about A, and we make such a small number of observations that we get practically no evidence; well, that's just not a very promising basis on which to do plausible reasoning. We can't expect to get anything useful out of it. We do, of course, get definite numerical values for the probabilities, but these values are very "soft," i.e., very unstable, because the A_p distribution is still very broad for small N. Our common sense tells us that the evidence N_n for small N provides no reliable basis for further predictions, and we'll see in the next lecture that this conclusion also follows as a consequence of the theory we're developing here.

The real reason for introducing the rule of succession lies in the cases where we <u>do</u> get a significant amount of information from the random experiment; i.e., when N is a large number. In this case, fortunately, we can pretty much forget about these fine points concerning prior evidence. The particular initial assignment $(A_p|X)$ will no longer have much influence on the results, for the same reason as in the particle-counter problem. This remains true for the generalized case leading to (16-38). You see from (16-39) that as soon as the number of observations N is large compared to the number of hypotheses K, then the probability assigned to any particular hypothesis depends for all practical purposes, only on what we have observed, and not on how many prior hypotheses there are. If you contemplate this for ten seconds, I think your common sense will tell you that the criterion $N \gg K$ is exactly the right one for this to be so.

16.4. Confirmation and Weight of Evidence.

Now, I'd like to introduce a few new ideas which are suggested by our calculations involving A_p. We saw that the stability of probability assignment in the face of new evidence is essentially determined by the width of the A_p distribution. If E is prior evidence and F is new evidence, then

$$(A|EF) = \int_0^1 (A_p|EF)\ p\ dp = \frac{\int_0^1 (A_p|F)(A_p|E)\ p\ dp}{\int_0^1 (A_p|F)(A_p|E)\ dp} \tag{16-40}$$

We'll say that F is compatible with E, as far as A is concerned, if having the new evidence, F, doesn't make any appreciable change in the probability of A; i.e.,

$$(A|EF) = (A|E) \tag{16-41}$$

The new evidence can make an enormous change in the distribution of A_p without changing the first moment. It might sharpen it up very much, or broaden it. We could become either more certain or more uncertain about A, but if F doesn't change the center of gravity of the A_p distribution, we still end up assigning the same probability to A.

Now, the stronger property: the new evidence F confirms the previous probability assignment, if F is compatible with it, and at the same time, gives us more confidence in it. In other words, we exclude one of these possibilities, and with new evidence F the A_p distribution narrows. Suppose F consists of performing some random experiment and observing the frequency with which A is true. In this case $F = N_n$, and our previous result, Eq. (16-20), gives

$$(A_p|N_n) = \frac{(N+1)!}{n!(N-n)!}\, p^n (1-p)^{N-n}$$

$$\simeq (\text{constant}) \cdot \exp\left[-\frac{(p-f)^2}{2\sigma^2}\right] \tag{16-42}$$

where

$$\sigma^2 = \frac{f(1-f)}{N} \tag{16-43}$$

and $f = (n/N)$ is the observed frequency of A. The approximation is derived by expanding $\log(A_p|N_n)$ in a Taylor series about its peak value, and is valid when $n>>1$ and $(N-n)>> 1$. If these conditions are satisfied, then $(A_p|N_n)$ is very nearly symmetric about its peak value. Then, if the observed frequency f is close to the prior probability $(A|E)$, the new evidence N_n will not affect the first moment of the A_p distribution, but will sharpen it up, and that will constitute a confirmation as I defined it. This shows one more connection between probability and frequency. I defined the "confirmation" of a probability assignment according to entirely different ideas than are usually used to define it. I defined it in a way that agrees with our intuitive notion of confirmation of a previous state of mind. But it turned out that the same experimental evidence would constitute confirmation on either the frequency theory or our theory.

Now, from this we can see another useful notion; which I'll call weight of evidence.

Let's consider A_p, given two different pieces of evidence, E and F.

$$(A_p|EF) = (\text{constant})(A_p|E)(A_p|F) \tag{16-44}$$

If the distribution $(A_p|F)$ was very much sharper than the distribution $(A_p|E)$, then the product of the two would still have its peak at practically the value determined by F. In this case, we would say that the evidence F carries much greater "weight" than the evidence E. If we have F, it doesn't really matter much whether we take E into account or not. On the other hand, if we don't have F, then whatever evidence E may represent will be extremely significant, because it will represent the best we are able to do. So, acquiring one piece of evidence which carries a great amount of weight can

make it, for all practical purposes, unnecessary to continue keeping track of other pieces of evidence which carry only a small weight.

Of course, this is exactly the way our minds operate. When we receive one very significant piece of evidence, we no longer pay so much attention to vague evidence. In so doing, we are not being very inconsistent, because it wouldn't make much difference anyway. So, our intuitive notion of weight of evidence is bound up with the sharpness of this A_p distribution. Evidence concerning A that we consider very significant is not necessarily evidence that makes a big change in the probability of A. It is evidence that makes a big change in this distribution of A_p. Now seeing this, we can get a little more insight into the principle of indifference, Rule 4, and also make contact between this theory and Carnap's methods of inductive reasoning.

Before we can use the principle of indifference to assign numerical values of probabilities, there are two different conditions that have to be satisfied: (1) we have to be able to analyze the situation into mutually exclusive, exhaustive possibilities; (2) having done this, we must then find that the available information gives us no reason to prefer any of the possibilities to any other. In practice, these conditions are hardly ever met unless there's some evident element of symmetry in the problem. But there are two entirely different ways in which condition (2) might be satisfied. It might be satisfied as a result of ignorance, or it might be satisfied as a result of positive knowledge about the situation.

To illustrate this, let's suppose that a person who is known to be very dishonest is going to toss a coin and there are two people watching him. Mr. A is allowed to examine the coin. He has all the facilities of the National Bureau of Standards at his disposal. He performs thousands of experiments with scales and calipers and magnetometers and microscopes, X-rays, and neutron beams, and so on. Finally, he is convinced that the coin <u>is</u> perfectly

honest. Mr. B is not allowed to do this. All he knows is that a coin is being tossed by a shady character. He suspects the coin is biased, but he has no idea in which direction.

Condition (2) is satisfied equally well for both of these people. Each of them would start out by assigning probability one-half to each face. The same probability assignment can described a condition of complete ignorance or a condition of very great knowledge. Now, this sort of situation has seemed paradoxical for a long time. Why doesn't Mr. A's extra knowledge make any difference? Well, of course, it does make a difference. It makes a very important difference, but one that doesn't show up until we start performing this random experiment. The difference is not in the probability of A, but in the distribution of A_p.

Suppose the first toss is heads. To Mr. B, that constitutes evidence that the coin is biased to favor heads. And so, on the next toss, he would assign new probabilities to take that into account. But to Mr. A, the evidence that the coin is honest carries overwhelmingly greater weight than the evidence of one throw, and he'll continue to assign a probability of 1/2.

Well, now, you see what's going to happen. To Mr. B, every toss of the coin represents new evidence about its bias. Every time it's tossed, he will revise his assignments for the next toss; but after several tosses his assignments will get more and more stable, and in the limit $N \to \infty$ they will tend to the observed frequency of heads. To observer A, the evidence of symmetry continues to carry greater weight than the evidence of almost any number of throws, and he persists in assigning probability 1/2. Each has done consistent plausible reasoning on the basis of the information available to him, and our theory accounts for the behavior of each.

If you assumed that Mr. A had perfect knowledge of symmetry, you might conclude that his A_p distribution is a true δ-function. In that case, his

mind could never be changed by any amount of new data from the random experiment. Of course, that's a limiting case that's never reached in practice. Not even the Bureau of Standards can give us evidence that good.

16.5. Carnap's Inductive Methods.

Carnap (1952) gives an infinite family of possible "inductive methods," by which one can convert prior information and frequency data into a probability assignment and an estimate of frequencies for the future. His principle is that the final probability assignment $(A|N_n X)$ should be a weighted average of the prior probability $(A|X)$ and the observed frequency, $f = n/N$. Assigning a weight N to the "empirical factor" f, and an arbitrary weight λ to the "logical factor" $(A|X)$ leads to the method which Carnap denotes by $c_\lambda(h,e)$. Introduction of the A_p distribution accounts for this in more detail; the theory developed here includes all of Carnap's methods as special cases corresponding to different prior distributions $(A_p|X)$, and leads us to reinterpret λ as the weight of prior evidence. Thus, in the case of two hypotheses, the Carnap λ-method is the one you can calculate from the prior distribution $(A_p|X) = (\text{const.})\cdot[p(1-p)]^r$, with $2r = \lambda - 2$. The result is

$$(A|N_n X) = \frac{2n + \lambda}{2N + 2\lambda} = \frac{(n+r) + 1}{(N+2r) + 2} \quad . \qquad (16\text{-}45)$$

Greater λ thus corresponds to a more sharply peaked $(A_p|X)$ distribution.

In our coin-tossing example, the gentleman from the Bureau of Standards reasons according to a Carnap method with λ of the order of, perhaps, thousands to millions; while Mr. B, with much less prior knowledge about the coin, would use a λ of perhaps 5 or 6. (The case $\lambda = 2$, which gives Laplace's rule of succession, is much too broad to be realistic for coin tossing; for Mr. B surely knows that the center of gravity of a coin can't be moved by more than half its thickness from the geometrical center. Actually, as we will see in Lecture 19, this analysis isn't always applicable to tossing of real coins,

for reasons having to do with the laws of physics.)

From the second way I wrote Equation (16-45), you see that the Carnap λ-method corresponds to a weight of prior evidence which would be given by $(\lambda-2)$ trials, in exactly half of which A was observed to be true. Can we understand why the weighting of prior evidence is λ = (number of prior trials + 2), while that of the new evidence N_n is only (number of new trials) = N? Well, look at it this way. The appearance of the (+2) is the robot's way of telling us this: prior knowledge that it is <u>possible</u> for A to be either true or false, is equivalent to knowledge that A has been true at least once, and false at least once. This is hardly a derivation; but I think it makes excellent common sense.

But let's pursue this line of reasoning a step further. We started with the statement X: it is <u>possible</u> for A to be either true or false at any trial; but that is still a somewhat vague statement. Suppose we interpret it as meaning that A has been observed true exactly once, and false exactly once. If we grant that this state of knowledge is correctly described by Laplace's assignment $(A_p|X) = 1$, then <u>what was the "pre-prior" state of knowledge before we had the data X</u>? To answer this, we need only apply Bayes' theorem backwards, as we did at the beginning of Lecture 7. The result is: our "pre-prior" A_p-distribution must have been

$$(A_p|\ )\ dp = (\text{const.})\ \frac{dp}{p(1-p)} \tag{16-46}$$

which is the quasi-distribution representing "complete ignorance," or the "basic measure" of our parameter space, that we found by transformation groups in Lecture 12. So, here is another line of thought that could have led us to this measure.

It appears, then, that if we have definite prior evidence that it <u>is</u> possible for A to be either true or false on any one trial, then Laplace's

rule $(A_p|X) = 1$ is the appropriate one to use. But if initially we are so completely uncertain that we're not even sure whether it is possible for A to be true on some trials and false on others, then we should use the prior (16-46).

How different are the numerical results which the pre-prior assignment (16-46) gives us? Repeating the derivation of (16-20) with this pre-prior assignment we find that, provided n is not zero or N,

$$(A_p|N_n)' = \frac{(N-1)!}{(n-1)!\,(N-n-1)!}\,p^{n-1}(1-p)^{N-n-1} \tag{16-47}$$

which leads, instead of to Laplace's rule of succession, to the mean-value estimate of p:

$$(A|N_n)' = \int_0^1 (A_p|N_n)'\; p\; dp = \frac{n}{N} \tag{16-48}$$

equal to the observed frequency, and identical with the maximum-likelihood estimate of p. Likewise, provided $0 < n < N$, we find instead of (16-22) the formula

$$(M_m|N_n)' = \frac{\binom{m+n-1}{m}\binom{M-m+N-n-1}{M-m}}{\binom{N+M-1}{M}} \tag{16-49}$$

All of these results correspond to having observed one less success and one less failure.

Lecture 17

PROBABILITY AND FREQUENCY IN EXCHANGEABLE SEQUENCES

We are now in a position to say quite a bit more about connections between probability and frequency. These are of two main types: (a) given an observed frequency in a random experiment, to convert this information into a probability assignment, and (b) given a probability assignment, to predict the frequency with which some condition will be realized. We have seen, in Lectures 10 and 12, how the principles of maximum entropy and transformation groups lead to probability assignments which, if the quantity of interest happens to be the result of some "random experiment," correspond automatically to predicted frequencies, and thus solve problem (b) in some situations.

The rule of succession gives us the solution to problem (a) in a wide class of problems; if we have observed whether A was true in a very large number of trials, and the only knowledge we have about A is the result of this random experiment, and the constancy of the "causal mechanism," then it says that the probability we should assign to A at the next trial becomes practically equal to the observed frequency. Now, in fact, this is exactly what people who define probability in terms of frequency do; one postulates the existence of an unknown "absolute" probability, whose numerical value is to be found by performing random experiments. Of course, you must perform a very large number of experiments. Then the observed frequency of A is taken as the estimate of the probability. As we saw in Lecture 15, even the +1 and +2 in Laplace's formula turn up when the "frequentist" refines his

methods by taking the center of a confidence interval. So, I don't see how even the most ardent advocate of the frequency theory of probability can damn the rule of succession without thereby damning his own procedure; after all polemics, there remains the simple fact that in his own procedure, he is doing exactly what Laplace's rule of succession tells him to do. Indeed, to define probability in terms of frequency is equivalent to saying that the rule of succession is the only rule which can be used for converting observational data into probability assignments.

17.1. Prediction of Frequencies.

Now let's consider problem (b) in this situation; to reason from a probability to a frequency. This is simply a problem of parameter estimation, not different in principle from any other. Suppose that instead of asking for the probability that A will be true in the next trial, we wish to infer something about the relative frequency of A in an indefinitely large number of trials, on the basis of the evidence N_n. We must take the limit of Equation (16-22) as $M \to \infty$, $m \to \infty$, in such a way that $(m/M) \to f$. Introducing the proposition

$$A_f \equiv \text{"The frequency of A true in an indefinitely large number of trials is f,"}$$

we find in the limit that the probability density of A_f, given N_n, is

$$(A_f|N_n) = \frac{(N+1)!}{n!\ (N-n)!} f^n (1-f)^{N-n} \quad , \tag{17-1}$$

which is the same as our $(A_p|N_n)$ in (16-20), with f numerically equal to p. According to (17-1) the most probable frequency is equal to (n/N), the observed frequency in the past. But we have noted before that in parameter estimation (if you object to my calling f a "parameter," then let's just call it "prediction"), the most probable value is usually a poorer estimate than the mean value in the small sample case, where they can be appreciably

different. The mean value estimate of the frequency is

$$\overline{f} = \int_0^1 f(A_f|N_n)\, df = \frac{n+1}{N+2} \qquad (17\text{-}2)$$

i.e., just the same as the value of $(A|N_n)$ given by Laplace's rule of succession. Thus, we can interpret the rule in either way; the probability which Laplace's theory assigns to A at a single trial is numerically equal to the estimate of frequency which minimizes the expected square of the error. You see how nicely this corresponds with the relation between probability and frequency which we found in the maximum-entropy and transformation group arguments.

Note also that the distribution $(A_f|N_n)$ is quite broad for small N, confirming our expectation that no reliable predictions should be possible in this case. As a numerical example, if A has been observed true once in two trials, then $\overline{f} = (A|N_n) = 1/2$; but according to (17-1) it is still an even bet that the true frequency f lies outside the interval $0.326 < f < 0.674$. With no evidence at all ($N = n = 0$), it would be an even bet that f lies outside the interval $0.25 < f < 0.75$. More generally, the variance of (17-1) is

$$\mathrm{var}(A_f|N_n) = \overline{f^2} - \overline{f}^2 = \overline{f}(1-\overline{f})/(N+3) \qquad (17\text{-}3)$$

so that the expected error in the estimate (17-2) decreases like $N^{-1/2}$. More detailed conclusions about the reliability of predictions, which we could make from (17-2) are for all practical purposes identical with those the statistician would make by the method of confidence intervals.

All these results hold also for the generalized rule of succession. Taking the limit of (16-38) as $M \to \infty$, $m_i/M \to f_i$, we find the joint probability distribution for A_i to occur with frequency f_i to be

$$(f_1 \ldots f_k | n_1 \ldots n_k)\ df_1 \ldots df_k$$

$$= \frac{(n+k-1)!}{n_1! \ldots n_k!}\ (f_1^{n_1} \ldots f_k^{n_k})\ \delta(f_1+\ldots+f_k-1)\ df_1 \ldots df_k \qquad (17\text{-}4)$$

The probability that the frequency f_1 will be in the range df_1 is found by integrating (17-4) over all values of $f_2 \ldots f_k$ compatible with $f_i \geq 0$, $(f_2+\ldots+f_k) = 1 - f_1$. This can be carried out by application of Laplace transforms in a well known way, and the result is

$$(f_1 | n_1 \ldots n_k)\ df_1 = \frac{(N+K-1)!}{n_1!\ (N-n_1+K-2)!}\ f_1^{n_1}(1-f_1)^{N-n_1+K-2}\ df_1 \qquad (17\text{-}5)$$

from which we find the most probable and mean value estimates of f_1 to be

$$(\hat{f}_1) = \frac{n_1}{N+K-2} \qquad (17\text{-}6)$$

$$\overline{f}_1 = \frac{n_1+1}{N+K} \quad , \quad \text{compare (16-39)} \qquad (17\text{-}7)$$

Another interesting result is found by taking the limit of $(M_m | A_p)$ in (16-17) as $M \to \infty$, $(m/M) \to f$. We easily find

$$(A_f | A_p) = \delta(f-p) \qquad (17\text{-}8)$$

Likewise, taking the limit of $(A_p | N_n)$ in (16-20) as $N \to \infty$, we find

$$(A_p | A_f) = \delta(p-f) \qquad (17\text{-}9)$$

which also follows from (17-8) by application of Bayes' theorem. Therefore, if B is any proposition, we have from our standard argument,

$$(B | A_f) = \int_0^1 (BA_p | A_f)\ dp = \int_0^1 (B | A_p A_f)(A_p | A_f)\ dp$$

$$= \int_0^1 (B | A_p)\ \delta(p-f)\ dp \quad . \qquad (17\text{-}10)$$

In the last step we used the property (16-1) that A_p automatically neutralizes

any other statement about A. Thus, if f and p are numerically equal, we have $(B|A_p) = (B|A_f)$; A_p and A_f are <u>equivalent</u> statements in their implication for plausible reasoning.

To verify this equivalence in one case, note that in the limit $N \to \infty$, $(n/N) \to f$, $(M_m|N_n)$ in Equation (16-22) reduces to the binomial distribution $(M_m|A_p)$ as given by (16-17). The generalized formula (16-), in the corresponding limit, goes into the multinomial distribution,

$$(m_1 \ldots m_k | f_1 \ldots f_k) = \frac{m!}{m_1! \ldots m_k!} f_1^{m_1} \ldots f_k^{m_k} \quad . \qquad (17\text{-}11)$$

This equivalence shows why it is so easy to confuse the notions of probability and frequency, and why in many problems this confusion does no harm. Whenever the available information consists of observed frequencies in a large sample, and constancy of the "causal mechanism," Laplace's theory becomes mathematically equivalent to the frequency theory. Most of the "classical" problems of statistics (life insurance, etc.) are of just this type; and as long as one works only on such problems, all is well. The harm arises when we consider more general problems.

Today, physics and engineering offer many important applications for probability theory in which there is an absolutely essential part of the evidence which cannot be stated in terms of frequencies, and/or the quantities about which we need plausible inference have nothing to do with frequencies. The axiom (probability) $\equiv$ (frequency), if applied consistently, would prevent us from using probability theory in these problems.

17.2. One-Dimensional Neutron Multiplication.

Our discussion so far has been rather abstract; perhaps too much so. In order to make amends for this, I would like to show you a specific physical problem where these equations apply. This was first described in a short

note by Bellman, Kalaba, and Wing (1957) and further developed in the recent book of Wing (1962). Neutrons are traveling in fissionable material, and we want to estimate how many new neutrons will be produced in the long run in consequence of one incident trigger neutron. In order to have a tractable mathematical problem, we make some drastic simplifying assumptions:

(a) the neutrons travel only in the ±x-direction, at a constant velocity.

(b) each time a neutron, traveling either to the right or the left, initiates a fission reaction, the result is exactly two neutrons, one traveling to the right, one to the left. The net result is therefore that any neutron will from time to time emit a progeny neutron traveling in the opposite direction.

(c) the progeny neutrons are immediately able to produce still more progeny in the same manner.

We fire a single trigger neutron into a thickness x of fissionable material from the left, and the problem is to predict the number of neutrons that will emerge from the left and from the right, over all time, as a consequence. At least, that is what we would <u>like</u> to calculate. But of course, the number of emerging neutrons is not <u>determined</u> by any of the given data, and so the best we can do is to calculate the <u>probability</u> that exactly n neutrons will be transmitted or reflected. I want to make a detailed comparison of the Laplace theory and the frequency theory of probability, as applied to the initial formulation of this problem. I am concerned mainly with the underlying rationale by which we relate probability theory to the physical model.

Many proponents of the frequency theory berate the Laplace theory on purely philosophical grounds that have nothing to do with its success or failure in applications. There is a more defensible position, held by some, who recognize that the present state of affairs gives them no reason for smugness,

and a good reason for caution. While they believe that at present the frequency theory is superior, they also say, as one of my correspondents did to me, "I will most cheerfully renounce the frequency theory for any theory that yields me a better understanding and a more efficient formalism." The trouble is that the current statistical literature gives us no opportunity to see the Laplace theory in actual use so that valid comparisons could be made; and that is the situation I am trying to correct here.

First, let us formulate the problem as it would be done on the frequency theory. Here is the way the "frequentist" would reason:

"The exeerimentalists have measured for us the relative frequency $p = a\Delta$ of fission in a very small thickness Δ of this material. This means that they have fired N trigger neutrons at a thin film of thickness Δ, and observed fission in n cases. Since N is finite, we cannot find the exact value of p from this, but it is approximately equal to the observed frequency (n/N). More precisely, we can find confidence limits for p. In similar situations, we can expect that about k per cent of the time, the limits (Cramér, 1946; p. 515)

$$\frac{N}{N+\lambda^2}\left[\frac{2n+\lambda^2}{2N} \pm \lambda\sqrt{\frac{n(N-n)}{N^3}+\frac{\lambda^2}{4N^2}}\right] \tag{17-12}$$

will include the true value of p, where λ is the (100 - k) per cent value of a normal deviate. For example, with $\lambda = \sqrt{2}$, the range

$$\frac{n+1}{N+2} \pm \frac{N}{N+2}\sqrt{\frac{2n(N-n)}{N^3}+\frac{1}{N^2}} \approx \frac{n+1}{N+2} \pm \sqrt{\frac{2n(N-n)}{N^3}} \tag{17-13}$$

will cover the correct p in about 84 per cent of similar cases. [Again, there's that +1 and +2 of Laplace's rule of succession!] In general, the connection between λ and k is given by

$$\frac{1}{\sqrt{2\pi}}\int_{-\lambda}^{\lambda} e^{-\frac{x^2}{2}}\,dx = \frac{k}{100}$$

Equation (17-12) is an approximation valid when the numbers n and (N-n) are sufficiently large; the exact confidence limits are difficult to express analytically, and for small N one should consult the graphs of Pearson and Clopper (1934). The number p is, of course, a definite, but imperfectly known, physical constant characteristic of the fissionable material.

"Now in order to calculate the relative frequency with which n neutrons will be reflected from a thickness x of this material, we have to make some additional assumptions. We assume that the probability of fission per unit length is always the same for each neutron independently of its history. Due to the complexity of the causes operating, it seems reasonable to assume this; but the real test of whether it is a valid assumption can come only from comparison of the final results of our calculation with experiment. This assumption means that the probabilities of fission in successive slabs of thickness Δ are independent so that, for example, the probability that an incident neutron will undergo fission in the second slab of thickness Δ, but not in the first, is the product $p(1-p)$.

"At this point we turn to the mathematics and solve the problem by any one of several possible techniques, emerging with the relative frequencies $p_n(x)$, $q_n(x)$ for reflection or transmission of n neutrons, respectively. [Actually, the analytical solution has not yet been found, but the book of Wing (1962) gives the results of numerical integration, which is equally good for orr purposes.]

"We now compare these predictions with experiment. When the first trigger neutron is fired into the thickness x, we observe r_1 neutrons reflected and t_1 neutrons transmitted. This datum does not in any way affect the assignments $p_n(x)$, $q_n(x)$, since the latter have no meaning in terms of a single experiment, but are predictions only of limiting frequencies for an indefinitely large number of experiments. We therefore must repeat

the experiment many times, and record the numbers r_i, t_i for each experiment. If we find that the frequency of cases for which $r_i = n$ tends sufficiently close to $p_n(x)$ ['sufficiently close' being determined by certain significance tests such as Chi-squared], then we conclude that the theory is satisfactory; or at least that it is not rejected by the data. If, however, the observed frequencies show a wide departure from $p_n(x)$, then we know that there is something wrong with our initial set of assumptions.

"Now, of course, the theory is either right or wrong. If it is wrong, then in principle the entire theory is demolished, and we have to start all over again, trying to find the right theory. In practice, it may happen that only one minor feature of the theory has to be changed, so that most of the old calculations will still be useful in the new theory."

* *

Now let's state this same problem in terms of Laplace's theory. We regard it simply as an exercise in plausible reasoning, in which we make the best possible guesses as to the outcome of a single experiment, or of any finite number of them. We are not concerned with the prediction, or even the existence, of limiting frequencies; because any assertion about the outcome of an impossible experiment is obviously an empty statement, and cannot be relevant to any application. We reason as follows:

The experimentalists have provided us with the evidence N_n, by firing N neutrons at a thin film of thickness Δ, and observing fission in n cases. Since by hypothesis the only prior knowledge was that a neutron either will or will not undergo fission, we have just the situation where Laplace's rule of succession applies and the probability, on this evidence, of fission for the (N+1)'th neutron in thickness Δ, is

$$p \equiv (F_{N+1}|N_n) = \frac{n+1}{N+2} \tag{17-14}$$

where

$F_m \equiv$ "the m'th neutron will undergo fission."

Whether N is large or small, the question of the "accuracy" of this probability does not arise--it is exact by definition. Of course, we will prefer to have as large a value of N as possible, since this increases the weight of the evidence N_n and makes the probability p, not more <u>accurate</u>, but more <u>stable</u>. The probability p is manifestly <u>not</u> a physical property of the fissionable material, but is only a means of describing our state of knowledge about it, on the basis of the evidence N_n. For, if the preliminary experiment had yielded a different result N_n', then we would of course assign a different probability p'; but the properties of the fissionable material would remain the same.

We now fire a neutron at a thickness $x = M\Delta$. Define the propositions,

$F^n \equiv$ "The neutron will cause fission in the n'th slab of thickness Δ."

$f^n \equiv$ "The neutron will not cause fission in the n'th slab."

The probability of fission in slab 1 is then

$$p = (F^1|N_n) = \frac{n+1}{N+2} \quad . \tag{17-15}$$

But now the probability that fission will occur in the second but not the first slab, is <u>not</u> p(1-p) as in the first treatment. At this point we see one of the fundamental differences between the theories. From our Rule 1, we have

$$(F^2f^1|N_n) = (F^2|f^1N_n)(f^1|N_n) = \frac{n+1}{N+2}\left[1 - \frac{n+1}{N+2}\right]$$

$$= \frac{(n+1)(N-n+1)}{(N+2)(N+3)} \tag{17-16}$$

The difference is that in calculating the probability $(F^2|f^1N_n)$, we must take into account the evidence f^1, that a neutron has passed through one more thickness Δ without fission. This amounts to one more experiment in

addition to that leading to N_n. The evidence f^1 is fully as cogent as N_n, and it would be clearly inconsistent to take one into account and ignore the other. Continuing in this way, we find that the probability that the incident neutron will emit exactly m first-generation progeny in passing through thickness $M\Delta$ is just the expression

$$(M_m|N_n) = \binom{M}{m} \frac{(n+m)!\,(N+1)!\,(N+M-n-m)!}{n!\,(N-n)!\,(N+M+1)!} \tag{17-17}$$

which we have derived before, Eq. (16-22). Now if N is not a very large number, this may differ appreciably from the value

$$(M_m|A_p) = \binom{M}{m} p^m(1-p)^{M-m} \tag{17-18}$$

which one obtains in the frequency approach. However, note again that as the weight of the evidence N_n increases, we find $(A_{p'}|N_n) \to \delta(p' - \frac{n}{N})$, and

$$(M_m|N_n) \to (M_m|A_p)$$

in the limit $N \to \infty$, $(n/N) \to p$. The difference in the two results is negligible whenever N>>M; i.e. when the weight of the evidence N_n greatly exceeds that of M_m. Now let's study the difference between (17-17) and (17-18) more closely. From (17-17) we have for the mean value estimate of m, on the Laplace theory,

$$\overline{m} = M\,\frac{n+1}{N+2} \tag{17-19}$$

To state the accuracy of this estimate, we can calculate the variance of the distribution (17-17). This is most easily done by using the representation (16-21):

$$\begin{aligned}\overline{m^2} &= \sum_{m=0}^{M} m^2 \int_0^1 (M_m|A_p)(A_p|N_n)\,dp \\ &= \frac{(N+1)!}{n!\,(N-n)!}\int_0^1 [Mp + M(M-1)p^2]\,p^n(1-p)^{N-n}\,dp \\ &= M\,\frac{n+1}{N+2} + M(M-1)\,\frac{(n+1)(n+2)}{(N+2)(N+3)}\end{aligned} \tag{17-20}$$

which gives the variance

$$V = \overline{m^2} - \overline{m}^2 = \frac{N+M+2}{N+3} M \frac{n+1}{N+2}\left[1 - \frac{n+1}{N+2}\right] \tag{17-21}$$

while, from (17-18), the frequency theory gives

$$\overline{m}_o = Mp \tag{17-22}$$

$$V_o = (\overline{m^2} - \overline{m}^2)_o = Mp(1-p) \tag{17-23}$$

If the frequentist takes the center of the confidence interval (17-13) as his "best" estimate of p, then he will take p = (n+1)/(N+2) in these equations. So, we both obtain the same estimate, but the variance (17-21) is greater by the amount

$$V - V_o = \frac{M-1}{N+3} Mp(1-p) \tag{17-24}$$

Why this difference? Why is it that the Laplace theory seems to determine the value of m less precisely than the frequency theory? Well, appearances are deceiving here. The fact is that the Laplace theory determines the value of m <u>more</u> precisely than the frequency theory; the variance (17-23) is not the entire measure of the uncertainty as to m on the frequency theory, because there is still the uncertainty as to the "true" value of p. According to (17-23), p is uncertain by about $\pm\sqrt{2p(1-p)/N}$, so the mean value (17-22) is uncertain by about

$$\pm M\sqrt{\frac{2p(1-p)}{N}} \tag{17-25}$$

in addition to the uncertainty represented by (17-23). If we suppose that the uncertainties (17-23) and (17-25) are independent, the total mean square uncertainty as to the value of m on the frequency theory would be represented by the sum of (17-23) and

$$M^2 \frac{2p(1-p)}{N} \tag{17-26}$$

which more than wipes out the difference (17-24). The factor 2 in (17-26)

would of course be changed somewhat by adopting a different confidence level; but no reasonable choice can change it very much.

In the frequency theory, the two uncertainties (17-23), (17-26) appear as entirely separate effects which are determined by applying two different principles; one by conventional probability theory, the other by confidence intervals. In the Laplace theory no such distinction exists; both are given automatically by a single calculation. We found exactly this same situation back in our particle-counter problem [Lecture 9, Sec. 9.3.], when we compared our robot's procedure with that of the orthodox statistician.

The mechanism by which the Laplace theory is able to do this is very interesting. It is just the difference already noted; in the derivation of (17-17) we are continually taking into account additional evidence accumulated in the new experiment, such as f^1 in (17-16). In the frequency theory, the uncertainty (17-25) in p arises because only a finite amount of data was provided by the preliminary experiment given N_n. It is just for that reason that the new evidence, such as f^1, is still relevant. In thus giving a consistent treatment of all the evidence, the Laplace theory automatically includes the effect of the finiteness of the preliminary data, which the frequency theory is able to do only crudely by the introduction of confidence intervals. In the Laplace theory there is no need to decide on any arbitrary "confidence level" because probability theory, when consistently applied to the whole problem, already tells us what weight should be given to the preliminary data N_n.

What we get in return for this is not merely a more unified treatment; in yielding a smaller net uncertainty in m, the Laplace theory shows that the two sources of uncertainty (17-23) and (17-26) of the frequency theory are not independent; they have a small negative correlation, so that they tend to compensate each other. That is the reason for Laplace's smaller

probable error. If you think about this very hard, you will be able to see intuitively why this negative correlation has to be there--I won't deprive you of the pleasure of figuring it out for yourself. All this subtlety is completely lost in the frequency theory.

"But," someone will object, "you are ignoring a very practical consideration which was the original reason for introducing confidence intervals. While I grant that in principle it is better to treat the whole problem in a single calculation, in practice we usually have to break it up into two different ones. After all, the preliminary data N_n was obtained by one group of people, who had to communicate their results to another group, who then carried out the second calculation applying this data. It is a practical necessity that the first group be able to state their conclusions in a way that tells honestly what they found, and how reliable it was. Their data can also be used in many other ways than in your second calculation, and the introduction of confidence intervals thus filled a very important practical need for communication between different workers."

Of course, if you have followed everything in these lectures so far, you know the answer to this. The memory storage problem was our original point of departure, and the problem just discussed is a specific example of just what I pointed out more abstractly in Eq. (16-15). You see from (16-21). and also in our derivation of (17-21), that the only property of the preliminary data which we needed in order to analyze the whole problem was the A_p-distribution $(A_p|N_n)$ that resulted from the preliminary experiment. The principle of confidence intervals was introduced to fill a very practical need. But there was no need to introduce any new principle for this purpose; it is already contained in probability theory, which shows that the exact way of communicating what you have learned is not by specifying confidence intervals, but by specifying your final A_p-distribution.

As a further point of comparison, note that in the Laplace theory there was no need to introduce any "statistical assumption" about independence of events in successive slabs of thickness Δ. In fact, the theory told us, as in Eq. (17-16), that these probabilities are not independent when we have only a finite amount of preliminary data; and it was just this fact that enabled the Laplace theory to take account of the uncertainty which the frequency theory describes by means of confidence intervals.

Now this brings up a very fundamental point about probability theory, which the frequency theory fails to recognize; but which is essential for applications to both communication theory and statistical mechanics, as I will show in later lectures. What do we mean by saying that two events are "independent?"

In the frequency theory, the only kind of independence recognized is causal independence; i.e. the fact that one event occurred does not in itself exert any physical influence on the occurrence of the other. Thus, in the coin-tossing example that I discussed in Lecture 16, the fact that the coin comes up heads on one toss, of course, doesn't physically affect the result of the next toss, and so on the frequency theory one would call the coin-tossing experiment a typical case of "independent repetitions of a random experiment;" the probability of a heads at both tosses must be the product of the separate probabilities . But then, you lose any way of describing the difference between the reasoning of Mr. A and Mr. B in that example!

In Laplace's theory, "independence" means something entirely different, which we see from a glance at our Rule 1: $(AB|C) = (B|C)(A|BC)$. Independence means that $(A|BC) = (A|C)$; i.e. knowledge that B is true does not affect the probability we assign to A. Thus, independence means not mere causal independence, but logical independence. Even though heads at one toss does not physically predispose the coin to give heads at the next, the knowledge

that we got heads may have a very great influence on our predictions as to the next toss.

The importance of this is that the various limit theorems, which I'll say more about later, require independence in their derivations. Consequently, even though there may be strict causal independence, if there is not also logical independence, these limit theorems will not hold. Writers of the frequency school of thought, who deny that probability theory has anything to do with inductive reasoning, recognize the existence only of causal connections, and as a consequence, they have long been applying these limit theorems to physical and communication processes where, I claim, they are incorrect and completely misleading. This was noted long ago by Keynes (1921), who stressed exactly this same point.

I think these comparisons make it very clear that, at least in this kind of problem, the Laplace theory does provide the "better understanding and more efficient formalism" that my colleague asked for.

17.3. The de Finetti Theorem.

So far we have considered the notion of an A_p-distribution and derived a certain class of probability distributions from it, under the restriction that the same A_p-distribution is to be used for all trials. Intuitively, this means that we have assumed the underlying "mechanism" as constant, but unknown. It is clear that this is a very restrictive assumption, and the question arises, how general is the class of probability functions that we can obtain in this way? In order to state the problem clearly, let us define

$$x_n \equiv \begin{cases} 1, & \text{if A is true on the n'th trial} \\ 0, & \text{if A is false on the n'th trial} \end{cases}$$

Then a state of knowledge about N trials is described in the most general

way by a probability function $p(x_1 \ldots x_N)$ which could, in principle, be defined arbitrarily (except for normalization) at each of the 2^N points.

We now ask; what is a necessary and sufficient condition on $p(x_1 \ldots x_N)$ for it to be derivable from an A_p-distribution? What test could we apply to a given distribution $p(x_1 \ldots x_N)$ to tell whether it is included in our theory as given above? A necessary condition is clear from our previous equations; any distribution obtainable in the way we have derived them necessarily has the property that the probability that A is true in n specified trials, and false in the remaining (N-n) trials, depends only on the numbers n and N; i.e., not on which trials in $1 \leq n \leq N$ were specified. If this is so, we say that $p(x_1 \ldots x_N)$ defines an exchangeable sequence.

An important theorem of de Finetti (1937) asserts that the converse is also true: any exchangeable probability function $p(x_1 \ldots x_N)$ can be generated by an A_p-distribution. Thus there is a function $(A_p|X) = g(p)$ such that $g(p) \geq 0$, $\int_0^1 g(p)\, dp = 1$, and the probability that in N trials A is true in n specified trials and false in the remaining (N-n), is given by

$$P_N(n) = \int_0^1 p^n (1-p)^{N-n}\, g(p)\, dp \tag{17-27}$$

This can be proved as follows. Note that $p^n(1-p)^{N-n}$ is a polynomial of degree N:

$$p^n(1-p)^{N-n} = p^n \sum_{m=0}^{N-n} \binom{N-n}{m} (-p)^m = \sum_{k=0}^{N} \alpha_k(N,n)\, p^k \tag{17-28}$$

which defines $\alpha_k(N,n)$. Therefore, if (17-27) holds, we would have

$$P_N(n) = \sum_{k=0}^{N} \alpha_k(N,n)\, \beta_k \tag{17-29}$$

where

$$\beta_n = \int_0^1 p^n\, g(p)\, dp \tag{17-30}$$

is the n'th moment of g(p). Thus, specifying $\beta_0, \beta_1, \beta_2, \ldots, \beta_N$ is equivalent to specifying all the $P_N(n)$ for $n = 0, 1, 2, \ldots, N$. Conversely, for given N,

specifying $P_N(n)$, $0 \le n \le N$, is equivalent to specifying $\{\beta_0 \ldots \beta_N\}$. In fact, β_N is the probability that $x_1 = x_2 = \ldots = x_N = 1$, regardless of what happens in later trials, and its relation to $P_N(n)$ can be established directly without reference to any function $g(p)$.

So, the problem reduces to this: if the numbers β_0, β_1, β_2, ... are specified, under what conditions does a function $g(p) \ge 0$ exist such that (17-30) holds? This is just the well-known Hausdorff moment problem, whose solution can be found many places; for example in the book of Widder (1941; Chap. 3). Translated into our notation, the main theorem is this: A necessary and sufficient condition that a function $g(p) \ge 0$ exists satisfying (17-30) [and therefore also (17-27)] is that there exist a number B such that

$$\sum_{n=0}^{N} \binom{N}{n} P_N(n) \le B, \qquad N = 0, 1, 2, \ldots. \tag{17-31}$$

But, from the interpretation of $P_N(n)$ as probabilities, we see that the equality sign always holds in (17-31) with $B = 1$, and the proof is completed.

Here is another way of looking at it, which might be made into a proof with a little more work, and perhaps discloses more clearly the intuitive reason for the de Finetti theorem, as well as showing immediately just how much we have said about $g(p)$ when we specify the $P_N(n)$. Imagine $g(p)$ expanded in the form

$$g(p) = \sum_{n=0}^{\infty} a_n \phi_n(p) \tag{17-32}$$

where $\phi_n(p)$ are the complete orthonormal set of polynomials in $0 \le p \le 1$, essentially the Legendre functions:

$$\phi_n(p) = \frac{\sqrt{2n+1}}{n!} \frac{d^n}{dp^n} [p(1-p)]^n$$

$$= (-)^n \sqrt{2n+1}\, P_n(2p-1) \quad . \tag{17-33}$$

$\phi_n(p)$ is a polynomial of degree n, and satisfies

$$\int_0^1 \phi_m(p)\ \phi_n(p)\ dp = \delta_{mn} \qquad (17\text{-}34)$$

If we substitute (17-34) into (17-27), only a finite number of terms will survive, because $\phi_k(p)$ is orthogonal to all polynomials of degree $N < k$. Then, it is easily seen that for given N, specifying the values of $P_N(n)$, $0 \le n \le N$, is equivalent to specifying the first (N+1) expansion coefficients $\{a_0, a_1, a_2, \ldots, a_N\}$. Thus, as $N \to \infty$, a function $g(p)$, defined by (17-32), becomes uniquely determined to the same extent that a fourier series uniquely determines its generating function; i.e., "almost everywhere." The main trouble with this argument is that the condition $g(p) \ge 0$ is not so easily established from (17-32).

The de Finetti theorem is very important to us because it shows that the connections between probability and frequency which we have found in this lecture hold for a fairly wide class of probability functions $p(x_1 \ldots x_N)$, namely the class of all exchangeable sequences. These results, of course, generalize immediately to the case where there are more than two possible outcomes at each trial.

Possibly even more important, however, is the light which the de Finetti theorem sheds on one of the oldest controversies in probability theory--Laplace's first derivation of the rule of succession. The idea of an A_p-distribution is not, needless to say, my own invention. The way I have introduced it here is only my attempt to translate into modern language what I think Laplace was trying to say in that famous passage, "When the probability of a simple event is unknown, we may suppose all possible values of this probability between 0 and 1 as equally likely." This statement, which I interpret as saying that with no prior evidence, $(A_p|X) = \text{const.}$, has been rejected as utter nonsense by virtually everyone who has written on probability theory in this century. And, of course, on any frequency definition

of probability, Laplace's stattment could have no justification at all. But on any theory it is conceptually difficult, since it seems to involve the idea of a "probability of a probability," and the use of an A_p-distribution in calculations has been largely avoided since the time of Laplace.

The de Finetti theorem puts some much more solid ground under these methods. Independently of all conceptual problems, it is a mathematical theorem that whenever you talk about a situation where the probability of a certain sequence of results depends only on the number of successes, not on the particular trials at which they occur, all your probability distributions can be generated from a single function g(p), in just the way we have done here. The use of this generating function is, moreover, a very powerful technique mathematically, as you will quickly discover if you try to repeat some of the above derivations [for example, Equation (16-22)] without using an A_p-distribution. So, it doesn't matter what you or I might think about the A_p-distribution conceptually; its validity as a mathematical tool for dealing with exchangeable sequences is a proven fact, standing beyond the reach of mere philosophical objections.

Lecture 18

APPLICATION OF PROBABILITY THEORY TO PHYSICAL MEASUREMENTS

Suppose we wish to determine the charge e and mass m of the electron. The Millikan oil-drop experiment measures e directly. The deflection of an electron beam in a known electromagnetic field measures the ratio (e/m). The deflection of an electron beam toward a metal plate due to attraction of image charges measures (e^2/m).

From the results of any two of these experiments we can calculate values of e and m. But all the measurements are subject to error, and the values of e, m obtained from different experiments will not agree. How, then, do we process the data so as to make use of all the information available and get the best estimates of e, m? What is the probable error remaining? How much would the situation be improved by including still another experiment of given accuracy? In this lecture I want to show that probability theory gives simple and elegant answers to these questions.

18.1. Reduction of Equations of Condition.

More specifically, suppose we have the results of these experiments:

(1) measures e with ±2% accuracy

(2) measures (e/m) with ±1% accuracy

(3) measures (e^2/m) with ±5% accuracy

Supposing the values of e, m approximately known in advance, $e \approx e_o$, $m \approx m_o$, the measurements are then linear functions of the corrections. Write the

unknown true values of e and m as

$$e = e_o(1 + x_1)$$
$$m = m_o(1 + x_2) \qquad (18\text{-}1)$$

then x_1, x_2 are dimensionless corrections, small compared to unity, and our problem is to find the best estimate of x_1 and x_2. The results of the three measurements are three numbers M_1, M_2, M_3 which we write as

$$M_1 = e_o(1 + y_1)$$
$$M_2 = \frac{e_o}{m_o}(1 + y_2)$$
$$M_3 = \frac{e_o^2}{m_o}(1 + y_3) \qquad (18\text{-}2)$$

where the y_i are also small dimensionless numbers which are defined by (18-2) and are therefore known in terms of the old estimates e_o, m_o and the new measurements M_1, M_2, M_3. On the other hand, the true values of e, e/m, e^2/m are expressible in terms of the x_i:

$$e = e_o(1 + x_1)$$
$$\frac{e}{m} = \frac{e_o(1 + x_1)}{m_o(1 + x_2)} = \frac{e_o}{m_o}(1 + x_1 - x_2 + \ldots)$$
$$\frac{e^2}{m} = \frac{e_o^2(1 + x_1)^2}{m_o(1 + x_2)} = \frac{e_o^2}{m_o}(1 + 2x_1 - x_2 + \ldots) \qquad (18\text{-}3)$$

where higher order terms are considered negligible. Comparing (18-2) and (18-3) we see that if the measurements were exact we would have

$$y_1 = x_1$$
$$y_2 = x_1 - x_2$$
$$y_3 = 2x_1 - x_2$$

But taking into account the errors, the known y_i are related to the unknown x_j by

$$y_1 = a_{11}\, x_1 + a_{12}\, x_2 + \delta_1$$

$$y_2 = a_{21}\, x_1 + a_{22}\, x_2 + \delta_2$$

$$y_3 = a_{31}\, x_1 + a_{32}\, x_2 + \delta_3 \qquad (18\text{-}4)$$

where the coefficients a_{ij} form a (3 x 2) matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \\ 2 & -1 \end{pmatrix} \qquad (18\text{-}5)$$

and the δ_i are the unknown fractional errors of the three measurements. For example, the statement $\delta_2 = -0.01$ means that the second measurement gave a result one per cent too small.

More generally, we have n unknown quantities $\{x_1 \ldots x_n\}$ to be estimated from N imperfect observations $\{y_1 \ldots y_N\}$, with $N \geq n$, and the N "equations of condition,"

$$y_i = \sum_{j=1}^{n} a_{ij}\, x_j + \delta_i \,, \quad i = 1, 2, \ldots, N. \qquad (18\text{-}6)$$

or, in matrix notation,

$$y = Ax + \delta \qquad (18\text{-}7)$$

where A is an (N x n) matrix.

It seems plausible that the best estimate of each x_j will be some linear combination of all the y_i, but if N > n we cannot simply solve equation (18-7) for x, since A is not a square matrix and has no inverse. However, we can get a system of equations solvable for x if we take n linear combinations of the equations of condition; i.e., if we multiply (18-7) on the left by some (n x N) matrix B. Then the product BA exists and is a square (n x n)

matrix. Choose B so that $(BA)^{-1}$ exists. Then the linear combinations are the n rows of

$$By = BAx + B\delta \tag{18-8}$$

which has the unique solution

$$x = (BA)^{-1} B(y - \delta) \tag{18-9}$$

If the probabilities of various fractional errors δ_i are symmetric: $p(\delta_i) = p(-\delta_i)$ so that $\langle\delta_i\rangle = 0$, then corresponding to any given matrix B the "best" estimate of x_j by almost any criterion will be the j'th row of

$$\bar{x} = (BA)^{-1} By \tag{18-10}$$

But by making different choices of B (i.e. taking different linear combinations of the equations of condition) we get different estimates. Which choice of B is best?

In the above I have merely restated, in modern terms, the old problem of "reduction of equations of condition" studied by 18'th century astronomers and described in Laplace's "Essai Philosophique." A popular criterion for solution was the principle of least squares; find that matrix B for which the sum of the squares of the errors in $\bar{x}_j$ is a minimum; or perhaps use a weighted sum. This problem can be solved directly.

18.2. Reformulation as a Decision Problem.

But we really solved this problem in Lecture 13, for we have already shown in full generality that the best estimate of any parameter (or any quantity, if you are squeamish about calling every unknown quantity a "parameter), by the criterion of any loss function, is found by applying Bayes' theorem to find the probability that the parameter lies in various intervals, then making that estimate which minimizes the expected loss taken over the posterior probabilities.

Now in the original formulation of the problem, as given above, it was

only a plausible conjecture that the best estimate of x_j is a linear combination of the y_i as in Equation (18-10). The material in Lecture 13 shows us a much better way of formulating the problem, in which we don't have to depend on conjecture. Instead of trying to take linear combinations without knowing which combinations to take, we should apply Bayes' theorem directly to the equations of condition. Then, if the best estimates are indeed of the form (18-10), Bayes' theorem should not only tell us that fact, it will automatically give us also the best choice of the matrix B.

Let's do this calculation for the case the probabilities assigned to the errors δ_i of the various measurements are independent and gaussian. We expect this to be the most realistic case, since in most physical measurements the total error is the sum of contributions from many small imperfections, and the central limit theorem, to be discussed later, would then lead us to the gaussian form. To anticipate a little, this is subject to one important qualification; that in general the gaussian approximation will be good only for those values of total error δ which can arise in many different ways by combination of the individual elementary errors. For unusually wide deviations the gaussian approximation can be very bad--just how bad we will see later when we study the Cauchy distribution.

The probability that the errors $\{\delta_1 \ldots \delta_N\}$ lie in the intervals $d\delta_1 \ldots d\delta_N$ respectively, is

$$p(\delta_1 \ldots \delta_N)\, d\delta_1 \ldots d\delta_N = (\text{const.})\, \exp\left[-\frac{1}{2} \sum_{i=1}^{N} w_i \delta_i^2 \right] d\delta_1 \ldots d\delta_N \qquad (18\text{-}11)$$

where the "weight" w_i is the reciprocal variance of the error of the i'th measurement. For example, the crude statement that the first measurement has ±2 per cent accuracy, now becomes the more precise statement that the first measurement has weight

$$w_1 = \frac{1}{\langle \delta_1^2 \rangle} = \frac{1}{(.02)^2} = 2500 \qquad (18\text{-}12)$$

From (18-6) and (18-11) we have immediately the probability density for obtaining measured values $\{y_1 \ldots y_N\}$ given the true values $\{x_1 \ldots x_n\}$:

$$(y_1 \ldots y_N | x_1 \ldots x_n) = C_1 \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} w_i [y_i - \sum_{j=1}^{n} \alpha_{ij} x_j]^2 \right\} \quad (18\text{-}13)$$

where C_1 is independent of the y_i. According to Bayes' theorem, if we assign uniform prior probabilities to the x_j, then the posterior probability density for the x_j, given the actual measurements y_i, is of the form

$$(x_1 \ldots x_n | y_1 \ldots y_N) = C_2 \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} w_i [y_i - \sum_{j=1}^{n} a_{ij} x_j]^2 \right\} \quad (18\text{-}14)$$

where C_2 is independent of the x_j. Now

$$\sum_{i=1}^{N} w_i (y_i - \sum_{j=1}^{n} a_{ij} x_j)^2$$

$$= \sum_{i=1}^{N} w_i [y_i^2 - 2y_i \sum_{j=1}^{n} a_{ij} x_j + \sum_{j,k=1}^{n} a_{ij} a_{ik} x_j x_k]$$

$$= \sum_{j,k=1}^{n} K_{jk} x_i x_k - 2 \sum_{j=1}^{n} L_j x_j + \sum_{i=1}^{N} w_i y_i^2 \quad (18\text{-}15)$$

where

$$K_{jk} \equiv \sum_{i=1}^{N} w_i a_{ij} a_{ik} \quad (18\text{-}16)$$

$$L_j \equiv \sum_{i=1}^{N} w_i y_i a_{ij} \quad (18\text{-}17)$$

or, defining a diagonal "weight" matrix $W_{ij} = w_i \delta_{ij}$, we have a matrix K and a vector L:

$$K = \tilde{A} W A \quad (18\text{-}18)$$

$$L = \tilde{A} W y \quad (18\text{-}19)$$

where $\tilde{A}$ is the transposed matrix. We want to write (18-14) in the form

$$(x_1 \ldots x_n | y_1 \ldots y_N) = C_3 \exp[-\frac{1}{2} \sum_{j,k=1}^{n} K_{jk} (x_j - \bar{x}_j)(x_k - \bar{x}_k)] \quad (18\text{-}20)$$

whereupon the $\bar{x}_j$ will be the mean value estimates desired. Comparing (18-15) and (18-20) we see that

$$\sum_{k=1}^{n} K_{jk} \bar{x}_k = L_j \quad (18\text{-}21)$$

or,

$$\bar{x}_k = \sum_{j=1}^{n} (K^{-1})_{kj} L_j \tag{18-22}$$

and this is the solution for best estimates of the $\bar{x}_k$ by the mean-square error criterion. From (18-18) and (18-19), we can write the result as

$$\bar{x} = (\tilde{A} W A)^{-1} \tilde{A} W y \tag{18-23}$$

and, comparing with (18-10), we see that in the gaussian case with uniform prior probabilities, the best estimates are indeed of the form (18-10), and the best choice of the matrix B is

$$B = \tilde{A} W \quad , \tag{18-24}$$

a result given by Laplace (1819).

Let us apply this solution to our determination of e and m. Here the measurements of e, (e/m), (e^2/m) were of 2%, 1%, 5% accuracy respectively, and so

$$w_2 = \frac{1}{(.01)^2} = 10{,}000$$

$$w_3 = \frac{1}{(.05)^2} = 400 \tag{18-25}$$

and we found $w_1 = 2500$ before. Thus we have

$$B = \tilde{A} W = \begin{pmatrix} 1 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix} \begin{pmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ 0 & 0 & w_3 \end{pmatrix} = \begin{pmatrix} w_1 & w_2 & 2w_3 \\ 0 & -w_2 & -w_3 \end{pmatrix} \tag{18-26}$$

$$K = \tilde{A} W A = \begin{pmatrix} (w_1+w_2+4w_3) & -(w_2+2w_3) \\ -(w_2+2w_3) & (w_2+w_3) \end{pmatrix} \tag{18-27}$$

$$K^{-1} = (\tilde{A} W A)^{-1} = \frac{1}{\Delta} \begin{pmatrix} (w_2+w_3) & (w_2+2w_3) \\ (w_2+2w_3) & (w_1+w_2+4w_3) \end{pmatrix} \tag{18-28}$$

where

$$\Delta = \det(K) = w_1w_2 + w_2w_3 + w_3w_1 \tag{18-29}$$

Thus the final result is

$$(\tilde{A}\ W\ A)^{-1}\ \tilde{A}\ W = \frac{1}{\Delta}\begin{pmatrix} w_1(w_2+w_3) & -w_2w_3 & w_2w_3 \\ w_1(w_2+2w_3) & -w_2(w_1+2w_3) & w_3(w_2-w_1) \end{pmatrix} \tag{18-30}$$

and the best estimates of x_1, x_2 are

$$\bar{x}_1 = \frac{w_1(w_2+w_3)y_1 + w_2w_3(y_3-y_2)}{w_1w_2 + w_2w_3 + w_3w_1} \tag{18-31}$$

$$\bar{x}_2 = \frac{w_1w_2(y_1-y_2) + w_2w_3(y_3-2y_2) + w_3w_1(2y_1-y_3)}{w_1w_2 + w_2w_3 + w_3w_1} \tag{18-32}$$

Inserting the numerical values of w_1, w_2, w_3, we have

$$\bar{x}_1 = \frac{13}{15}\,y_1 + \frac{2}{15}\,(y_2-y_3) \tag{18-33}$$

$$\bar{x}_2 = \frac{5}{6}\,(y_1-y_2) + \frac{2}{15}\,(y_3-2y_2) + \frac{1}{30}\,(2y_1-y_3) \tag{18-34}$$

which exhibits the best estimates as weighted averages of the estimates taken from all possible pairs of experiments. Thus, y_1 is the estimate of x_1 obtained in the first experiment, which measures e directly. The second and third experiments combined give an estimate of e given by $(e^2/m)(e/m)^{-1}$. Since

$$\frac{\frac{e_o^2}{m_o}(1+y_3)}{\frac{e_o}{m_o}(1+y_2)} \simeq e_o(1+y_3-y_2)$$

(y_3-y_2) is the estimate of x_1 given by experiments 2 and 3. Equation (18-33) says that these two independent estimates of x_1 should be combined with weights 13/15, 2/15. Likewise, Equation (18-34) gives $\bar{x}_2$ as a weighted average of three different (although not independent) estimates of x_2.

But how accurate are these estimates $\bar{x}_j$? From (18-20) we find the well-known formula for the second central moments of $(x_1 \ldots x_n | y_1 \ldots y_N)$:

$$\langle \Delta x_j \Delta x_k \rangle \equiv \langle (x_j - \bar{x}_j)(x_k - \bar{x}_k) \rangle = \langle x_j x_k \rangle - \langle x_j \rangle \langle x_k \rangle = (K^{-1})_{jk} \qquad (18\text{-}35)$$

Thus from the inverse matrix

$$K^{-1} = (\tilde{A}\, W\, A)^{-1} \qquad (18\text{-}36)$$

already found in our calculation of $\bar{x}_j$, we can also read off the probable errors, or more conveniently, the standard deviations. From (18-27) we can state the results in the form (mean) ± (standard deviation) as

$$x_j = \bar{x}_j \pm \sqrt{(K^{-1})_{ij}} \qquad (18\text{-}37)$$

or,

$$x_1 = \bar{x}_1 \pm \left[\frac{w_2 + w_3}{w_1 w_2 + w_2 w_3 + w_3 w_1} \right]^{1/2} \qquad (18\text{-}38)$$

$$x_2 = \bar{x}_2 \pm \left[\frac{w_1 + w_2 + 4w_3}{w_1 w_2 + w_2 w_3 + w_3 w_1} \right]^{1/2} \qquad (18\text{-}39)$$

with numerical values

$$x_1 = \bar{x}_1 \pm 0.0186$$
$$x_2 = \bar{x}_2 \pm 0.0216 \qquad (18\ 40)$$

so that from the three measurements we obtain e with ±1.86 per cent accuracy, m with ±2.16 per cent accuracy.

How much did the rather poor measurement of (e^2/m), with only ±5 per cent accuracy, help us? To answer this, note that in the absence of this experiment we would have arrived at conclusions given by (18-27), (18-31) and (18-32) in the limit $w_3 \to 0$. The results (also easily verified directly from the statement of the problem) are

$$\bar{x}_1 = y_1$$
$$\bar{x}_2 = y_1 - y_2 \qquad (18\text{-}41)$$

$$K^{-1} = \frac{1}{w_1 w_2}\begin{pmatrix} w_2 & w_2 \\ w_2 & (w_1+w_2) \end{pmatrix} \tag{18-42}$$

or, the (mean) ± (standard deviation) values are

$$x_1 = y_1 \pm \frac{1}{w_1} = y_1 \pm 0.020$$

$$x_2 = y_1 - y_2 \pm \left[\frac{w_1+w_2}{w_1 w_2}\right]^{1/2} = y_1 - y_2 \pm 0.024 \tag{18-43}$$

As might have been anticipated by common sense, a low-accuracy measurement can add very little to the results of accurate measurements, and if the (e^2/m) measurement had been much worse than ±5 per cent it would hardly be worth-while to include it in our calculations. But suppose that an improved technique gives us an (e^2/m) measurement of ±2 per cent accuracy. How much would this help? The answer is given by our previous formulas with $w_1 = w_3 = 2500$, $w_2 = 10{,}000$. We find now that the mean-value estimates give much higher weight to the estimates using the (e^2/m)) measurement:

$$\bar{x}_1 = 0.556\, y_1 + 0.444(y_3 - y_2)$$

$$\bar{x}_2 = 0.444(y_1 - y_2) + 0.444(y_3 - 2y_2) + 0.112(2y_1 - y_3) \tag{18-44}$$

which is to be compared with (18-33), (18-34). The standard deviations are given by

$$x_1 = \bar{x}_1 \pm 0.0149$$

$$x_2 = \bar{x}_2 \pm 0.020 \tag{18-45}$$

The accuracy of e is improved roughly twice as much as that of m, since the improved measurement involves e^2, but only the first power of m.

18.3. Discussion: A Paradox.

We can learn many more things from studying this problem. For example,

I want to point out something which you will find astonishing at first. If you study Equation (18-32), which gives the best estimate of m from the three measurements, you will see that y_3, the result of the (e^2/m) measurement, enters into the formula in a different way than y_1 and y_2. It appears once with a positive coefficient, and once with a negative one. If $w_1 = w_2$, these coefficients are equal and (18-32) collapses to

$$\overline{x}_2 = y_1 - y_2 \tag{18-46}$$

Now, realize the full implications of this: it says that the only reason we make use of the (e^2/m) measurement in estimating m is that the (e) measurement and the (e/m) measurement have different accuracy. No matter how accurately we know (e^2/m), if the (e) and (e/m) measurements happen to have the same accuracy, however poor, then we should ignore the good measurement and base our estimate of m only on the (e) and (e/m) measurements!

I think that your common sense will instantly revolt against this conclusion, and you will say that there must be an error in Equation(18-32). So, let's take a minute off while you check the derivation.

This is a perfect example of the kind of result which probability theory gives us almost without effort, but which our unaided common sense might not notice in years of thinking about the problem. I won't deprive you of the pleasure of resolving this "paradox" for yourself, and explaining to your friends how it can happen that consistent inductive reasoning may demand that you throw away your best measurement.

You recall that, back at the end of Lecture 9, I complained about the fact that orthodox statisticians sometimes throw away relevant data in order to fit a problem to their model of "independent random errors." Am I now guilty of advocating the same thing? No doubt, it looks very much that way! Yet I plead innocence--the numerical value of (e^2/m) is in fact irrelevant to inference about m, if we already have measurements of e and e/m of equal

accuracy. Try drawing diagrams--or just try and figure out how you would use (e^2/m) in this situation--and I think you'll see why this is so.

As another example, it is important that we understand the way our conclusions depend on our choice of loss functions and probability distributions for the errors δ_i. If we use instead of the Gaussian distribution (18-11) one with wider tails, such as the Cauchy distribution $p(\delta) \propto (1 + \frac{1}{2}w\delta^2)^{-1}$, the posterior distribution $(x_1x_2|y_1y_2y_3)$ may have more than one peak in the (x_1,x_2)-plane. Then a quadratic loss function, or more generally any concave loss function (i.e. doubling the error more than doubles the loss) will lead you to make estimates of x_1 and x_2 which lie between the peaks, and are known to be very unlikely. With a convex loss function a different "paradox" appears, in that the basic equation (13-16) for constructing the best estimator may have more than one solution, with nothing to tell us which one to use.

The appearance of these situations is the robot's way of telling us this: our state of knowledge about x_1 and x_2 is too complicated to be described adequately simply by giving estimates and probable errors. The only honest way of describing what we know is to give the actual distribution $(x_1x_2|y_1y_2y_3)$. This is one of the limitations of decision theory which we have to understand in order to use it properly.

Lecture 19

PHYSICS OF "RANDOM" EXPERIMENTS

As we have already noted several times in these lectures, the idea that probability assignments must be based ultimately on observed frequencies in random experiments is fundamental to almost all recent expositions of probability theory; which would seem to make it a branch of experimental science. At the end of Lecture 9 we saw some of the difficulties that this view leads us to, in that in some real physical experiments the distinction between random and nonrandom quantities is so obscure and artificial that you have to resort to black magic in order to force this distinction into the problem. But in that discussion we didn't really get into the serious physics of the situation. In this lecture, I want to take time off from development of probability theory, and have a little interlude of more physical considerations that show the fundamental difficulty with the notion of "random" experiments--even the ones, such as coin tossing, which at first glance seem most appropriately regarded as "random."

We have also noted that there have always been dissenters from the orthodox view who have maintained, with Laplace, that probability theory is properly regarded as the "calculas of inductive reasoning," and is not fundamentally related to random experiments at all. According to this second view, consideration of random experiments is only one particular <u>application</u> of probability theory (and not even the most important one); for probability theory accounts equally well for general inductive inferences where no random

experiment is involved. But we haven't yet noted that there is an interesting correlation; those who have advocated the second view have tended to be physicists rather than mathematicians. So, it will be of interest to examine the historical background of this question with particular emphasis on the physics of the situation.

With the rise of the "Neo-Bayesian" school of thought, this question has flared up again in the recent literature of statistics. Several participants have recognized that the issue is not merely one of philosophy or mathematics; in some way not yet made entirely clear, it also involves physics. The mathematician tends to think of a random experiment as an abstraction--really nothing more than a sequence of numbers. To define the "nature" of of the random experiment he introduces statements--variously termed assumptions, postulates, or axioms--which specify the sample space and assert the existence, and certain other properties, of limiting frequencies. In real life, however, a random experiment is not an abstraction whose properties can be defined at will; it is surely subject to the laws of physics.

As soon as a specific random experiment is described, it is the nature of a physicist to start thinking, not about the abstract sample space thus defined, but about the physical mechanism of the phenomenon being observed. The question whether the usual postulates of probability theory are compatible with the known laws of physics is capable of logical analysis, with results that have a direct bearing on the question, not of the mathematical validity of frequency and non-frequency theories of probability, but of their applicability to real situations. Any such conclusions have, evidently, a relevance to the question of orthodox vs. Bayesian statistical methods.

In a recent discussion of these questions Professor G. E. P. Box (196) has remarked, "I believe, for instance, that it would be very difficult to persuade an intelligent physicist that current statistical practice was sensible,

but that there would be much less difficulty with an approach via likelihood and Bayes' theorem." Let's analyze this statement in the light both of history and of physics.

19.1. Historical Background.

As we know, probability theory started in consideration of gambling devices by Cardano, Pascal, and Fermat; but its development beyond that level, in the 18'th and 19'th centuries, was stimulated by applications in physics and astronomy, and was the work of people--Jacob and Daniel Bernoulli, Laplace, Poisson, Legendre, Gauss--most of whom we would describe today as mathematical physicists.

In the nineteenth century a knowledge of statistical analysis, consisting largely of the work of Laplace, Legendre, and Gauss, was considered an essential part of the training of a scientist. For example, as a young man J. Willard Gibbs spent three years (1866-69) in post-doctoral study at the Universities of Paris, Berlin, and Heidelberg; and the most prominent topic mentioned in the list of lectures he attended was statistical analysis. This study undoubtedly contributed to his discovery, 33 years later, of the basic "canonical ensemble" formalism of statistical mechanics.

A radical change took place early in this century when a new group of workers, not physicists, entered the field. They proceeded to reject virtually everything done by Laplace, and sought to develop statistics anew based on entirely different principles. This extremely aggressive school soon dominated the field so completely that its methods have come to be known as "orthodox" statistics.

Simultaneously with this development, the physicists--with Sir Harold Jeffreys as almost the sole exception--quietly retired from the field, and statistical analysis disappeared from the physics curriculum. This disappear-

ance has been so complete that, if today someone were to take a poll of physicists, I think he would find that not one in a hundred could identify such names as Fisher, Neyman, Wald; or such terms as maximum likelihood, confidence interval, analysis of variance.

This course of events--the leading role of physicists in development of the original Bayesian methods, and their later withdrawal from orthodox statistics--was no accident. As further evidence that there is some kind of basic conflict between orthodox statistical doctrine and physics, we may note that two of the most eloquent proponents of non-frequency definitions in this century--Poincaré and Jeffreys--have been mathematical physicists of the very highest competence, as was Laplace. Professor Box's statement thus has a clear basis in historical fact.

But what is the nature of this conflict? What is there in the physicist's knowledge that has led him to reject the very thing that the orthodox statistician regards as conferring "objectivity" on his methods? To see where the difficulty lies, we examine a few simple random experiments from the physicist's viewpoint. The facts I want to point out are so elementary that you can't believe they are really unknown to modern writers on probability theory. The continual appearance of new statistical textbooks which ignore them merely illustrates what we physics teachers have always known; you can teach a student the laws of physics, but you cannot teach him the art of recognizing the relevance of this knowledge, much less the habit of applying it, in his everyday problems.

19.2. How to Cheat at Coin and Die Tossing.

Cramér (1946) takes it as an axiom that "Any random variable has a unique probability distribution." From the later context, it is clear that what he really means is that it has a unique frequency distribution. If one assumes

that the number obtained by tossing a die is a random variable, this leads to the conclusion that the frequency with which a certain face comes up is a physical property of the die; just as much so as its mass, moment of inertia, or chemical composition. Thus, Cramér (loc. cit., p. 154) states, "The numbers p_r should, in fact, be regarded as physical constants of the particular die that we are using, and the question as to their numerical values cannot be answered by the axioms of probability theory, any more than the size and the weight of the die are determined by the geometrical and mechanical axioms. However, experience shows that in a well-made die the frequency of any event r in a long series of throws usually approaches 1/6, and accordingly we shall often assume that all the p_r are equal to 1/6 . . . ."

To a physicist, such an attitude seems to show utter contempt for the known laws of mechanics. The results of tossing a die many times do not tell us any definite number characteristic of the die. Thyy tell us something about the way the die was tossed. If you toss "loaded" dice in different ways, you can easily alter the relative frequencies of the faces. With more difficulty, and over a smaller range, you can even do this if the die is perfectly "honest."

Although the principles will be just the same, it will be simpler to discuss a random experiment with only two possible outcomes per trial. Consider, therefore, a "biased" coin, about which I. J. Good has remarked (Savage, 1962): "Most of us probably think about a biased coin as if it had a physical probability. Now whether it is defined in terms of frequency or just falls out of another type of theory, I think we do argue that way. I suspect that even the most extreme subjectivist such as de Finetti would have to agree that he did sometimes think that way, though he would perhaps avoid doing it in print." It is, of course, just the famous theorem of de Finetti that we studied in Lecture 17, which shows us how to carry out a probability

analysis of the biased coin without thinking in the manner suggested (it does not follow, however, that this analysis is applicable to a real biased coin). In any event, it is quite easy to show how a physicist would analyze the problem. Let us suppose that the center of gravity of this coin lies on its axis, but displaced a distance x from its geometrical center. If we agree that the result of tossing this coin is a "random variable," then according to the axiom stated by Cramér and hinted at by Good, there must exist a definite functional relationship between the frequency of heads and x:

$$p_H = f(x)$$

But this assertion goes far beyond the mathematician's traditional range of freedom to invent arbitrary axioms, and encroaches on the domain of physics; for the laws of mechanics are quite competent to tell us whether such a functional relationship does or does not exist.

The easiest game to analyze turns out to be just the one most often played to decide such practical matters as the starting side in a football game. Your opponent first calls "heads" or "tails" at will. You then toss the coin into the air, catch it in your hand, and without looking at it, show it first to your opponent, who wins if he has called correctly. It is further agreed that a "fair" toss is one in which the coin rises at least nine feet into the air, and thus spends at least 1.5 seconds in free flight.

The laws of mechanics now tell us the following. The ellipsoid of inertia of a thin disc is an oblate spheroid of eccentricity $1/\sqrt{2}$. The displacement x does not affect the symmetry of this ellipsoid, and so according to the Poinsot construction, as found in textbooks on rigid dynamics [such as Routh (19)], the polhodes remain circles concentric with the axis of the coin. In consequence, the character of the tumbling motion of the coin while in flight is exactly the same for a biased as an unbiased coin, except that for the biased one it is the center of gravity, rather than the geo-

metrical center, which describes the parabolic "free particle" trajectory.

An important feature of this tumbling motion is conservation of angular momentum; during its flight the angular momentum of the coin maintains a fixed direction in space (but the angular velocity does not; and so the tumbling may appear chaotic to the eye). Let us denote this direction by the unit vector n; it can be any direction you choose, and it is determined by the particular kind of twist you give the coin at the instant of launching. Whether the coin is biased or not, it will show the same face throughout the motion if viewed from this direction (unless, of course, n is exactly perpendicular to the axis of the coin, in which case it shows no face at all).

Therefore, in order to know which face will be uppermost in your hand, you have only to carry out the following procedure. Denote by k a unit vector passing through the coin along its axis, with its point on the "heads" side. Now toss the coin with a twist so that k and n make an acute angle, then catch it with your palm held flat, in a plane normal to n. On successive tosses, you can let the direction of n, the magnitude of the angular momentum, and the angle between n and k, vary widely; the tumbling motion will then appear entirely different to the eye on different tosses, and it would require almost superhuman powers of observation to discover your strategy.

Thus, anyone familiar with the law of conservation of angular momentum can, after some practice, cheat at the usual coin-toss game and call his shots with 100 per cent accuracy. You can obtain any frequency of heads you want; and the bias of the coin has no influence at all on the results!

Of course, as soon as this result is out, someone will object that the experiment analyzed is too "simple." In other words, those who have postulated a "physical" probability for the biased coin have, without stating so, really had in mind a more complicated experiment in which some kind of "randomness" has more opportunity to make itself felt.

While accepting this criticism, I can't suppress the obvious comment: scanning the literature of probability theory, isn't it curious that so many mathematicians, usually far more careful than physicists to list all the qualifications needed to make a statement correct, should have failed to see the need for any qualifications here? However, to be more constructive, we can just as well analyze a more complicated experiment.

Suppose that now, instead of catching the coin in our hand, we toss it onto a table, and let it spin and bounce in various ways until it comes to rest. Is this experiment sufficiently "random" so that the true "physical probability" will manifest itself? No doubt, the answer will be that it is not sufficiently random if the coin is merely tossed up two inches starting at the table level, but it will become a "fair" experiment if we toss it up higher.

Exactly how high, then, must we toss it before the true "physical probability" can be measured? This is not an easy question to answer, and I certainly won't make any attempt to answer it here. It would appear, however, that anyone who asserts the existence of a "physical" probability for the coin ought to be prepared to answer it; otherwise it is hard to see what content the assertion has (in the sense of operational verifiability).

I don't deny that the bias of the coin will now have some influence on the frequency of heads; I claim only that the amount of that influence depends very much on how you toss the coin so that, again in this experiment, there is no definite number $p_H = f(x)$ describing a physical property of the coin. Indeed, even the direction of this influence can be reversed by different methods of tossing, as follows.

However high we toss the coin, we still have the law of conservation of angular momentum; and so we can toss it by Method A: to ensure that heads will be uppermost when the coin first strikes the table, we have only to hold

it heads up, and toss it so that the total angular momentum is directed vertically. Again, we can vary the magnitude of the angular momentum, and the angle between n and k, so that the motion appears quite different to the eye on different tosses, and it would require very close observation to notice that heads remains uppermost throughout the free flight. Although what happens after the coin strikes the table is complicated, the fact that heads is uppermost at first has a strong influence on the result, which is more pronounced for large angular momentum.

Many people have developed the knack of tossing a coin by Method B: it goes through a phase of standing on edge and spinning rapidly about a vertical axis, before finally falling to one side or the other. If you toss the coin this way, the eccentric position of the center of gravity will have a dominating influence, and render it practically certain that it will fall always showing the same face. Ordinarily, one would suppose that the coin prefers to fall in the position which gives it the lowest center of gravity; i.e., if the center of gravity is displaced toward tails, then the coin should have a tendency to show heads. However, for an interesting mechanical reason, which I leave for you to work out, method B produces the opposite influence, the coin strongly preferring to fall so that its center of gravity is high.

On the other hand, the bias of the coin has a rather small influence in the opposite direction if we toss it by Method C: the coin rotates about a horizontal axis which is perpendicular to the axis of the coin, and so bounces until it can no longer turn over.

In this experiment also, therefore, a person familiar with the laws of mechanics can toss a biased coin so that it will produce predominantly either heads or tails, at will. Furthermore, the effect of method A persists whether the coin is biased or not; and so one can even do this with a perfectly "honest" coin. Finally, although we have been considering only coins, essen-

tially the same mechanical considerations apply to the tossing of any other object, such as a die.

From the fact that we have seen a strong preponderance of heads, we cannot legitimately conclude that the coin is biased; it may be biased, or it may have been tossed in a way that systematically favors heads. Likewise, from the fact that we have seen equal numbers of heads and tails, we cannot legitimately conclude that the coin is "honest." It may be honest, or it may have been tossed in a way that nullifies the effect of its bias.

19.3. Experimental Evidence.

Since the conclusions just stated are in direct contradiction to what is postulated, almost universally, in expositions of probability theory, it is worth noting that anyone can easily verify them for himself, in a few minutes of experimentation in his kitchen. An excellent "biased coin" is provided by the metal lid of a small pickle jar, of the type which is not knurled on the outside, and has the edge rolled inward rather than outward, so that the outside surface is accurately round and smooth, and so symmetrical that on an edge view one cannot tell which is the top side.

Suspecting that many people simply would not believe the things just claimed without experimental proof, I have performed these experiments with a jar lid of diameter d = 2 5/8", height h = 3/8". Assuming a uniform thickness for the metal, the center of gravity should be displaced from the geometrical center by a distance $x = dh/(2d+8h) = 0.120$ inches; and this was confirmed by hanging the lid by its edge and measuring the angle at which it comes to rest. Ordinarily, one expects this bias to make the lid prefer to fall bottom side up; and so this side will be called "heads." The lid was tossed up about 6 feet, and fell onto a smooth linoleum floor. I allowed myself ten practice tosses by each of the three methods described, and then

recorded the results of a number of tosses by: method A deliberately favoring heads, method A deliberately favoring tails, method B, and method C, as given in Table 19.1.

Method	No. of Tosses	No. of Heads
A(H)	100	99
A(T)	50	0
B	100	0
C	100	54

Table 19.1. Results of tossing a "biased coin" in four different ways.

In method A the mode of tossing completely dominated the result (the effect of bias would, presumably, have been much greater is the "coin" were tossed onto a surface with a greater coefficient of friction). In method B, the bias completely dominated the result (in about thirty of these tosses it looked for a while as if the result were going to be heads, as one might naively expect; but each time the "coin" eventually righted itself and turned over, as predicted by the laws of rigid dynamics). In method C, there was no significant evidence for any effect of bias.

One can, of course, always claim that tossing the coin in any of the four specific ways described is "cheating," and that there exists a "fair" way of tossing it, such that the "true" probabilities will emerge from the experiment. But again, the person who asserts this ought to be prepared to define precisely what this fair method is, otherwise the assertion is without content. Presumably, a fair method of tossing ought to be some kind of random mixture of methods A(H), A(T), B, C, and others; but what is a "fair" relative weighting to give them? It is difficult to see how one could define

a "fair" method of tossing except by the condition that it should result in a certain frequency of heads; and so we are involved in a circular argument.

This analysis can be carried much further than we have done here, and I want to go into it some more in a minute; but it is perhaps sufficiently clear already that analysis of coin and die tossing is not a problem of abstract statistics, in which one is free to introduce postulates about "physical" probabilities which ignore the laws of physics. It is a problem of mechanics, highly complicated and irrelevant to probability theory except insofar as it forces us to think a little more carefully about how probability theory must be formulated if it is to be applicable to real situations. Performing a random experiment with a coin does not tell us what the "physical" probability of heads is; it may tell us something about the bias, but it also tells us something about how the coin is being tossed. Indeed, unless we know how it is being tossed, we cannot draw any inferences about its bias from the experiment.

It may not, however, be clear from the above that conclusions of this type hold quite generally for random experiments, and in no way depend on the particular mechanical properties of coins and dies. In order to illustrate this, let's consider an entirely different kind of random experiment.

19.4. Bridge Hands.

In Lectures 5 and 13, we have already quoted Professor Wm. Feller's pronouncements on the use of Bayes' theorem in quality control testing, about Laplace's rule of succession, and about Daniel Bernoulli's conception of the utility function for decision theory. He does not fail us here either; in this interesting textbook (Feller, 1950), he writes: "The number of possible distributions of cards in bridge is almost 10^{30}. Usually, we agree to consider them as equally probable. For a check of this convention more than 10^{30}

experiments would be required ...." Here again, we have the view that bridge hands possess "physical" probabilities, that the uniform probability assignment is a "convention," and that the ultimate criterion for its correctness must be observed frequencies in a random experiment.

The thing which is wrong here is that none of us would be willing to use this criterion in a real-life situation because, if we know that the deck is an "honest" one, our common sense tells us something which carries more weight than 10^{30} random experiments do. We would, in fact, be willing to accept the result of the random experiment only if it agreed with our preconceived notion that all distributions are equally likely.

To many of you this last statement may seem like pure blasphemy--it stands in violent contradiction to what we have all been taught. Yet in order to see why it is true, we have only to imagine that those 10^{30} experiments had been performed, and the uniform distribution was not forthcoming. We expect, if all distributions of cards have equal frequencies, that any combination of two specified cards will appear together in a given hand, on the average, once in $52 \cdot 51/13 \cdot 12 = 17$ deals. But suppose that the particular combination (Jack of hearts--Seven of clubs) appeared together in each hand three times as often as this. Would we then accept it as an established fact that this particular combination is inherently more likely than others?

We would not. We would say that the cards had not been properly shuffled. But once again we are involved in a circular argument; because there is no way to define a "proper" method of shuffling except by the condition that it should produce all distributions with equal frequency!

In carrying out a probability analysis of bridge hands, are we really concerned with physical probabilities; or with inductive reasoning? In order to help answer this, consider the following scenario: I tell an orthodox statistician that I have dealt at bridge 1000 times, shuffling "fairly" each

time; and that in every case the seven of clubs was in my own hand. What will his reaction be? He will, I think, mentally visualize the number

$$\left(\frac{1}{4}\right)^{1000} \simeq 10^{-602}$$

and conclude instantly that I have not told the truth; and no amount of persuasion on my part will shake that judgment. But what accounts for the strength of his belief? Obviously, it cannot be justified if our assignment of equal probabilities to all distributions of cards is merely a "convention," subject to change in the light of experimental evidence. Even more obviously, he is not making use of any knowledge about the outcome of an experiment involving 10^{30} bridge hands.

What is the extra evidence he has, which his common sense tells him carries more weight than any number of random experiments; but whose help he refuses to acknowledge in expounding probability theory? In order to maintain the claim that probability theory is an experimental science, based fundamentally not on inductive inference but on frequency in a random experiment, it is necessary to suppress some of the information which is available. This suppressed information, however, is just what enables inductive reasoning to approach the certainty of deductive reasoning in this example.

The suppressed evidence is, of course, simply our recognition of the symmetry of the situation. The only difference between a seven and an eight is that there is a different number printed on the face of the card. Our common sense tells us that where a card goes in shuffling depends only on the mechanical forces that are applied to it; and not on which number is printed on its face. If we observe any systematic tendency for one card to appear in the dealer's hand, which persists on indefinite repetitions of the experiment, we can infer from this only that there is some systematic tendency in the procedure of shuffling, which alone determines the outcome of the

experiment.

Once again, therefore, performing the experiment tells you nothing about the "physical" probabilities of different hands. It tells you something about how the cards are being shuffled.

19.5. General Random Experiments.

In the face of the foregoing arguments, one can still take the following position (as a member of the audience did after one of my recent lectures): "You have shown only that coins, dies, and cards represent exceptional cases, where mechanical considerations obviate the usual probability postulates; i.e., they are not really 'random experiments.' But that is of no importance because these devices are used only for illustrative purposes; in the more dignified random experiments which merit the serious attention of the scientist or engineer, there is a physical probability."

To answer this, note that any specific experiment for which the existence of a physical probability is asserted, is subject to physical analysis like the ones just given, which will lead eventually to an understanding of its mechanism. But as soon as this understanding is reached, then this new experiment will also appear as an exceptional case where physical considerations obviate the usual postulates of "physical" probabilities. For, as soon as we have understood the mechanism of any experiment E, then there is logically no room for any postulate that various outcomes possess physical probabilities; for the question: "What are the probabilities of various outcomes O_1, O_2,...?" then reduces immediately to the question: "What are the probabilities of the corresponding initial conditions I_1, I_2, ... that lead to these outcomes?"

We might suppose that the possible initial conditions of experiment E themselves possess physical probabilities. But then we are considering an antecedent random experiment E', which produces conditions I_k as its possible

outcomes. We can analyze the physical mechanism of E' and as soon as this is understood, the question will revert to: "What are the probabilities of the various initial conditions I_k' for experiment E'?" Evidently, we are involved in an infinite regress {E, E', E", ...}; the attempt to introduce a physical probability will be frustrated at every level where our knowledge of physical law permits us to analyze the mechanism involved. The notion of "physical probability" must retreat continually from one level to the next, as knowledge advances.

We are, therefore, in a situation very much like the "warfare between science and theology" of earlier times. For several centuries, theologians insisted on making factual assertions which encroached on the domains of astronomy, physics, biology, and geology--and which they were later forced to retract one by one in the face of advancing knowledge.

Clearly, probability theory ought to be formulated in a way that avoids factual assertions properly belonging to other fields, and which will later need to be retracted (as is now the case for many assertions in the literature concerning coins, dies, and cards). It appears to me that the only formulation which accomplishes this is the original one given by Laplace and expounded by Poincaré and Jeffreys, in which probability theory is regarded as the general "calculus of inductive reasoning," whose validity does not depend on any assumptions about properties of physical experiments. As we saw back in Lecture 3, a very important contribution to the logical foundations of this approach was made recently by R. T. Cox (1946), (1961), who showed that, if we represent degrees of plausibility by real numbers, then the mathematical rules for inductive inference are restricted by elementary conditions of consistency, stated in the form of functional equations whose general solutions are readily found. As already noted, it is no accident that all the aforementioned gentlemen are to be classed as physicists, to whom the things I

have pointed out in this lecture would be obvious from the start.

The Laplace-Poincaré-Jeffreys-Cox formulation of probability theory does not require us to take one reluctant step after another down that infinite regress; it recognizes that anything which continually recedes from the light of detailed analysis can exist only in our imagination. Performing any of the so-called random experiments will not tell us what the "physical" probabilities are, because there is no such thing as a "physical" probability. The experiment tells us, in a very crude and incomplete way, something about how the initial conditions are varying from one repetition to another.

A much more efficient way of obtaining this information would be to study the initial conditions directly. However, in many cases this is beyond our present abilities; as in determining the safety and effectiveness of a new medicine. Here the only fully satisfactory approach would be to analyze the detailed sequence of chemical reactions that follow the taking of this medicine, in persons of every conceivable state of health. Having this analysis one could then predict, for each individual patient, exactly what the effect of the medicine will be.

Such an analysis being entirely out of the question at present, the only feasible way of obtaining the information we want is to perform a "random" experiment. No two patients are in exactly the same state of health; and for a given dose, the unknown variations in this factor constitute the variable initial conditions of the experiment, while the sample space comprises the set of distinguishable reactions to the medicine.

Our use of probability theory in this case is an example of inductive reasoning which amounts to the following: "If the initial conditions of the experiment continue in the future to vary over the same unknown range as they have in the past, then I expect that the relative frequencies of various outcomes will, in the future, approximate those which I have observed in the

past. In the absence of positive evidence giving a reason why there should be some change in the future, and indicating in which direction this change should go, I can only suppose that things will continue in more or less the same way. As I observe the relative frequencies to remain stable over longer and longer times, I become more and more confident about this conclusion. But still, I am doing only inductive reasoning--there is no deductive proof that frequencies in the future will not be entirely different than those in the past.

Lecture 20

INTRODUCTION TO COMMUNICATION THEORY

At this point we have all the basic machinery of our theory developed, and have seen its application in some of the "classical" problems. We said back in the first talk that what started all this was the attempt to see statistical mechanics and communication theory as examples of the same line of reasoning. A generalized form of statistical mechanics appeared as soon as we supplemented Laplace's theory of inductive reasoning by the notion of entropy, and we ought now to be in a position to treat communication theory in a similar way.

One difference is that in statistical mechanics the prior information has nothing to do with frequencies (it consists of measured values of quantities such as pressure); while in communication theory the prior information is obtained in a different way, which makes the probability-frequency paradoxes much more acute. For this reason I thought it best to take up communication theory only after we had seen some of the general connections between probability and frequency, via the A_p distribution and the de Finetti theorem.

First the difficult matter of giving credit where credit is due. All major advances in understanding have their precursors, whose full significance is never recognized at the time. Relativity theory had them in the work of Mach, Fitzgerald, and Lorentz, to mention only the most obvious examples. Communication theory had many precursors, in the work of Gibbs, Nyquist, Hartley, Szilard, von Neumann, and Wiener. But there is no denying that the

work of Shannon (1948) represents the arrival of the main signal, just as did Einstein's of 1905. Here for the first time, ideas which had long been, so to speak, "in the air" in a vague form, are grasped and put into sharp focus.

Shannon's papers were so full of important new concepts and results that they exercised not only a stimulating effect, but also a paralyzing effect. During the first few years after their appearance, it was common to hear the opinion expressed, rather sadly, that Shannon had anticipated and solved all problems of the field, and left nothing else for others to do. Today, I think, no one entertains any such ideas, and the field has seen considerably more development.

The psst-Shannon developments, with few exceptions, can be classed into efforts in two entirely different directions. On the one hand we have the expansionists, who try to apply Shannon's ideas to other fields, as I have been doing. Others range from the entropy calculator (who works out the entropy of a television signal, the French language, a chromosome, or almost anything else you can imagine; and often finds that nobody knows what to do with the result), to the universalist (who assures us that communication theory will revolutionize all intellectual activity; but seldom offers a specific example of anything that has been changed by it).

We should not be critical of these efforts because, as J. R. Pierce has said, it is very hard to tell at present which ones make sense, which are pure nonsense, and which are the beginning of something that will in time make sense. My own efforts have received all three classifications from various quarters. I have a very strong hope, and a moderately strong belief, that the ideas introduced by Shannon will eventually be indispensable to the linguist, the geneticist, the television engineer, the neurologist, etc. But I share with many others a feeling of disappointment that twenty years

of effort along these lines has led to so little in the way of really useful advances in these fields. We have today an abundance of vague philosophy, and of abstract mathematics, but a rather embarrassing shortage of examples where specific practical problems have been solved by using communication theory.

The moral of this is, I think, that more than half the battle is in learning how to ask the right question. People who want to apply communication theory to new fields must learn that the first, and hardest, step is to state precisely what is the problem we want solved. Once we succeed in doing this, real progress comes easily. I will give some examples pertaining to statistical mechanics and decision theory in these lectures.

In almost diametric opposition to the above efforts, as far as aim is concerned, stand the mathematicians, who view communication theory simply as a branch of pure mathematics. Characteristic of this school is a belief that, before introducing a continuous probability distribution, you have to talk about set theory, Borel fields, measure theory, the Lebesgue-Stieltjes integral, and the Radon-Nikodym theorem. The important thing is to make the theorems rigorous by the criteria of rigor currently popular, even if in so doing we limit the scope of the practical theory, and/or make it unintelligible to the average scientist or engineer. The recently published books on information theory by A. Khinchin (1957) and A. Feinstein (1958) can serve as typical examples of the style prevalent in this literature.

Here again, no valid criticism of these efforts is possible. Of course, we want our principles to be subjected to the closest scrutiny one can bring to bear on them. If important applications exist, the need for this is so much the greater; fortunately, mathematicians have found the subject interesting enough to take on a not very easy task. However, the present talks are not addressed to mathematicians, but to scientists and engineers who are

interested in applications; and so I am going to dwell on this side of the story only to the extent of pointing out that the particular theorems which the mathematicians have chosen to rigorize are not always the ones relevant to real situations.

Now, in order to explain this rather cryptic remark, let's turn to some of the specific things in Shannon's papers.

20.1. The Noiseless Channel.

We deal with the transmission of information from some sender to some receiver. I will speak of them in anthropomorphic terms, such as "the man at the receiving end," although either or both might actually be machines, as in telemetry or remote control systems. Transmission takes place via some channel, which might be a telephone or telegraph circuit, a microwave link, a frequency band assigned by the FCC, the German language, the postman, the neighborhood gossip, or a chromosome. If, after having received a message, the receiver can always determine with certainty which message was intended by the sender, we say that the channel is noiseless.

It was recognized very early in the game, particularly by Nyquist and Hartley, that the capability of a channel is not described by any property of the specific messages it sends, but rather by what it could have sent. The usefulness of a channel depends on its ability to transmit any one of a large class of messages, which the sender can choose at will.

In a noiseless channel, the obvious measure of this ability is simply the maximum number, $W(t)$, of distinguishable (at the destination) messages which the channel is capable of transmitting in time t. In all cases of interest to us, this number eventually goes into an exponential increase for sufficiently large t: $W(t) \sim e^{Ct}$, so the measure of channel performance which is independent of any particular time interval is the coefficient C

of this increase. We define the channel capacity as

$$C \equiv \lim_{t\to\infty} \left[\frac{1}{t} \log W(t)\right] \tag{20-1}$$

The units in which C is measured will depend on which base we choose for our logarithms. Usually one takes the base 2, in which case C is given in "bits per second," one bit being the amount of information contained in a single binary (yes-no) decision. For easy interpretation of numerical values the bit is by far the best unit to use; but in formal operations it is easier to use the base e of natural logarithms, and I will do that in this discussion. Our channel capacities are therefore measured in natural units, or "nits per second." To convert, we note that 1 bit = $(\log_e 2)$ = 0.69315 nits, or 1 nit = 1.4427 bits.

The capacity of a noiseless channel is a definite number, characteristic of the channel, which contains no subjective features. Thus, if a noiseless channel can transmit n symbols per second, chosen in any order from an alphabet of a letters, we have $W(t) = a^{nt}$, or C = n log a nits/second. Any constraint on the possible sequences of letters can only lower this number. For example, if the alphabet is $A_1, A_2, \ldots, A_a$, and it is required that in a long message of N = nt symbols the letter A_i must occur with relative frequency f_i, then the number of possible messages in time t is only

$$W(t) = \frac{N!}{(Nf_1)! \ \ldots \ (Nf_a)!} \tag{20-2}$$

and from Stirling's approximation, we find, as in Eq. (10-17),

$$C = -n \sum_i f_i \log f_i \quad \text{nits/second.} \tag{20-3}$$

This attains its maximum value, equal to the previous C = n log a, in the case of equal frequencies, $f_i = a^{-1}$. Thus we have the interesting result that a constraint requiring all letters to occur with equal frequencies does not decrease channel capacity at all. It does, of course, decrease the number

W(t) by an enormous factor; but the decrease in log W is what counts, and this grows less rapidly than t, so it makes no difference in the limit.

Suppose now that symbol A_i has transmission time t_i, but there is no other constraint on the allowable sequences of letters. What is the channel capacity? Well, consider first the class of messages in which letter A_i occurs n_i times, i = 1, 2, ..., a. The number of such messages is

$$W(n_1 \ldots n_a) = \frac{N!}{n_1! \ldots n_a!} \tag{20-4}$$

where

$$N = \sum_{i=1}^{a} n_i \quad . \tag{20-5}$$

The total number of different messages that can be transmitted in time t is then

$$W(t) = \sum_{n_i} W(n_1 \ldots n_a) \tag{20-6}$$

where we sum over all choices of $(n_1 \ldots n_a)$ compatible with $n_i \geq 0$ and

$$\sum_{i=1}^{a} n_i t_i \leq t \; . \tag{20-7}$$

The number K(t) of terms in the sum (20-6) satisfies $K(t) \leq (Bt)^a$ for some $B < \infty$. This is seen most easily by imagining the n_i as coordinates in an a-dimensional space and noting the geometrical interpretation of (20-7).

Exact evaluation of (20-6) would be quite an unpleasant job. But it's only the limiting value that we care about right now, and we can get out of the hard work by the following trick. Note that W(t) cannot be less than the greatest term $W_m = W_{max}(n_1 \ldots n_a)$ in (20-6) nor greater than $W_m K(t)$:

$$\log W_m \leq \log W(t) \leq \log W_m + a \log (Bt) \tag{20-8}$$

and so we have

$$C \equiv \lim_{t \to \infty} \frac{1}{t} \log W(t) = \lim_{t \to \infty} \frac{1}{t} \log W_m \tag{20-9}$$

i.e., to find the channel capacity, it is sufficient to maximize log $W(n_1 \ldots n_a)$ subject to the constraint (20-7). This rather surprising fact can be understood as follows. The logarithm of W(t) is given, rather crudely, by

$$\log W(t) = \log W_{max} + \log [\text{number of reasonably large terms in (20-6)}]$$

Even though the number of large terms tends to infinity as t^a, this is not rapid enough to make any difference in comparison with the exponential increase of W_{max}. This same mathematical fact is the reason why, in statistical mechanics, the Darwin-Fowler method and the method of the most probable distribution lead to the same results in the limit of large systems.

We can solve the problem of maximizing log $W(n_1 \ldots n_a)$ by the same Lagrange multiplier argument used in Lecture 10, Section (10.6). The problem is not quite the same, however, because now N is also to be varied in finding the maximum.

Using the Stirling approximation, which is valid for large n_i, we have as before

$$\log W(n_1 \ldots n_a) \simeq N \log N - \sum_{i=1}^{a} n_i \log n_i \qquad (20\text{-}10)$$

The variational problem with λ a Lagrangian multiplier, is

$$\delta[\log W + \lambda \sum n_i t_i] = 0 \qquad (20\text{-}11)$$

but since $\delta N = \sum \delta n_i$, we have

$$\delta \log W = \delta N \log N - \delta N - \sum_i (\delta n_i \log n_i - \delta n_i)$$

$$= - \sum \delta n_i \log\left(\frac{n_i}{N}\right) \qquad (20\text{-}12)$$

Therefore (20-11) reduces to

$$\sum_{i=1}^{a} \left[\log\left(\frac{n_i}{N}\right) + \lambda t_i\right] \delta n_i = 0$$

with the solution

$$n_i = N\, e^{-\lambda t_i} \tag{20-13}$$

To fix the value of λ we require

$$N = \sum n_i = N \sum e^{-\lambda t_i} \tag{20-14}$$

With this choice of n_i, we find

$$\frac{1}{t} \log W_m = -\frac{1}{t} \sum n_i \log\left(\frac{n_i}{N}\right) = \frac{1}{t} \sum n_i (\lambda t_i) \tag{20-15}$$

In the limit, $t^{-1} \sum n_i t_i \to 1$, and we find

$$C = \lim_{t\to\infty} \frac{1}{t} \log W(t) = \lambda \ . \tag{20-16}$$

So, our final result can be stated very simply:

To calculate the capacity of a noiseless channel in which symbol A_i has transmission time t_i and which has no other constraints on the possible messages, define a partition function

$$Z(\lambda) = \sum_i e^{-\lambda t_i} \tag{20-17}$$

Then the channel capacity C is the real root of

$$Z(\lambda) = 1. \tag{20-18}$$

You see already a very strong resemblance to the reasoning and the formalism of statistical mechanics, in spite of the fact that we have not yet said anything about probability. From (20-14) we see that $W(n_1 \ldots n_a)$ is maximized when the relative frequency of symbol A_i is given by the canonical distribution

$$f_i = \frac{n_i}{N} = e^{-\lambda t_i} = e^{-Ct_i} \tag{20-19}$$

Should we conclude from this that the channel is being "used most efficiently"

when we have encoded our messages so that (20-19) holds? No, that wouldn't be quite the right way of putting it. Because, of course, in time t the channel will actually transmit one message and only one; and this remains true regardless of what relative frequencies we use. Equation (20-19) tells us only that the overwhelming majority of all possible messages that the channel could have transmitted in time t are ones where the relative frequencies are canonical.

On the other hand we have a generalization of the remark following (20-3); if we impose an additional constraint requiring that the relative frequencies are given by (20-19), which might be regarded as defining a new channel, the channel capacity would not be decreased. But any constraint requiring that all possible messages have letter frequencies different from (20-19) will decrease channel capacity.

There are many other ways of interpreting these equations. For example, in our above arguments we supposed that the total time of transmission is fixed and we wanted to maximize the number W of possible messages among which the sender can choose. In a practical communications system, the situation is usually the other way around; we know in advance the extent of choice which we demand in the messages which might be sent over the channel, so that W is fixed. We then ask for the condition that the total transmission time of the message be minimized subject to a fixed W.

It is well known that variational problems can be transformed into several different forms, the same mathematical result giving the solution to many different problems. A circle has maximum area for a given perimeter; and also it has minimum perimeter for a given area. In statistical mechanics, the canonical distribution can be characterized as the one with maximum entropy for a given expectation of energy; or equally well as the one with minimum expectation of energy for a given entropy. Similarly, the channel capacity

found from (20-18) gives the maximum attainable W for a given transmission time, while its reciprocal is equal to the minimum attainable transmission time for a fixed W.

As another extension of the meaning of these equations, note that we don't have to interpret the quantity t_i as a time; it can stand equally well for the "cost," as measured by any criterion, of transmitting the i'th symbol. For example, it might be that the total length of time the channel is in operation is of no importance, because the apparatus has to sit there in readiness whether it is being used or not. The real economic criterion might be the total amount of choice W_b of different messages which the apparatus is capable of transmitting before breaking down, for a given installation cost. The lifetime of the apparatus might be limited by the total number of times a certain relay has to open and close. In this case, we could define t_i as the number of times this relay must operate in the course of transmitting the i'th symbol. The channel capacity given by Equation (20-18) would then be measured, not in nits per second, but in "nits per relay operation," and its reciprocal is equal to the minimum attainable number of relay operations per nit of transmitted information.

A more complicated type of noiseless channel, also considered by Shannon, is one where the channel has a memory; it may be in any one of a set of "states," $\{S_1 \ldots S_k\}$ and the possible future symbols, or their transmission times, depend on the present state. For example, suppose that if the channel is in state S_i, it can transmit symbol A_n, which leaves the channel in state S_j, the corresponding transmission time being t_{inj}. Surprisingly, the calculation of channel capacity in this case is quite easy.

Let $W_i(t)$ be the total number of different messages the channel can transmit in time t, starting from state S_i. Breaking down $W_i(t)$ into several terms according to the first symbol transmitted, we have

$$W_i(t) = \sum_{jn} W_j(t - t_{inj}) \qquad (20\text{-}20)$$

where the sum is over all possible sequences $S_i \rightarrow A_n \rightarrow S_j$. This is a linear difference equation with constant coefficients, so its assymptotic solution must be an exponential function:

$$W_i(t) \approx B_i \exp(Ct) \qquad (20\text{-}21)$$

and from the definition (20-1) it is clear that, for finite k, the coefficient C is the channel capacity. Substituting (20-21) into (20-20), we obtain

$$B_i = \sum_{j=1}^{k} Z_{ij}(C)\, B_j \qquad (20\text{-}22)$$

where

$$Z_{ij}(\lambda) = \sum_{n} \exp(-\lambda t_{inj}) \qquad (20\text{-}23)$$

is the "partition matrix." If the sequence $S_i \rightarrow A_n \rightarrow S_j$ is impossible, we set $t_{inj} = \infty$. By this device we can understand the sum in (20-23) as extending over all symbols in the alphabet.

Equation (20-22) says that the matrix Z_{ij} has an eigenvalue equal to unity. Thus, the channel capacity is the greatest real root of $D(\lambda) = 0$, where

$$D(\lambda) \equiv \det[Z_{ij}(\lambda) - \delta_{ij}] \; . \qquad (20\text{-}24)$$

In the case of a single state, k = 1, this reduces to the previous rule, Equation (20-18).

The problems solved above are, of course, only especially simple ones. By inventing channels with more complicated types of constraints on the allowable sequences (i.e. with a long memory), you can generate mathematical problems as involved as you please. But it would still be just the mathematics--as long as the channel is noiseless, there would be no difficulties of principle. In each case you simply have to count up the possibilities and apply the definition (20-1). For some weird channels, you might find that the limit

therein does not exist, in which case we can't speak of a channel capacity, but have to characterize the channel simply by giving the function W(t).

20.2. The Information Source.

When we take the next step and consider the information source feeding our channel, fundamentally new problems arise. There are mathematical problems aplenty, but there are also more basic conceptual problems, which have to be considered before we can state which mathematical problems are the significant ones.

It was Professor Norbert Wiener who first suggested the enormously fruitful idea of representing an information source in probability terms. He applied this to some problems of filter design, which I will take up briefly in a later lecture. This work was an essential step in developing a way of thinking which led to modern communication theory.

It is perhaps difficult nowadays for us to realize what a big step this was. Previously, communication engineers had considered an information source simply as a man with a message to send; for their purposes an information source could be characterized simply by describing that message. But Wiener suggested instead that an information source be characterized by giving the probabilities that it will emit various messages. Already we can see some conceptual difficulties faced by a frequency theory of probability--the man at the sending end presumably knows perfectly well which message he is going to send. What, then, could we possibly mean by speaking of the probability that he will send something? There is nothing analogous to "chance" operating here.

By the probability of a message, do we mean the frequency with which he sends that particular message? The question is absurd--a sane man sends a given message at most once, and most messages never. Do we mean the frequency

with which the message M occurs in some imaginary "ensemble" of communication acts? Well, it's all right to state it this way if you want to, but it doesn't answer the question. It merely leads us to re-state the question as: what do we mean by the ensemble? How is it to be set up? Calling it by a different name doesn't help us.

Right at this point we have to state clearly what is the specific problem we want solved. A probability distribution is a means of describing a state of knowledge. But whose state of knowledge do we want to talk about? Evidently, not the man at the sending end. Is it the man at the receiving end? Well, that might be relevant to the problem I have in mind. But basically, since I am talking to scientists and engineers, I want to consider communication theory, not as describing the "general philosophy" of communication between sender and receiver, but as something of practical value to an engineer whose job is to design the technical equipment in the communication system. In other words, the state of knowledge we want to describe is that of the communication engineer when he designs the equipment.

This consideration is something you will not find in the previous literature based on the viewpoint which sees no distinction between probability and frequency; on this view, the notion of a probability for a person with a certain state of knowledge simply doesn't exist. Nevertheless, from any viewpoint, the problem of choosing some probability distribution to represent the information source does exist. It cannot be evaded, and the whole content of the theory depends on how we do this.

I have already emphasized several times that in probability theory we never solve an actual problem of practice. We solve only some abstract mathematical model of the real problem. Setting up this model requires not only mathematical ability, but also practical judgment. If our model does not correspond well to the actual situation, our theorems, however rigorous,

may be more misleading than helpful.

This is so with a vengeance in communication theory because, as I will show in this lecture, not only the quantitative details, but even the qualitative nature of the theorems that can be proved, depend on which probability model we use to represent an information source.

The purpose of this probability model is to describe the communication engineer's prior knowledge about the messages to be sent. In principle, this prior knowledge could be of any sort; but in "traditional" communication theory the only kind of prior knowledge considered consists of frequencies of letters, or combinations of letters, which have been observed in past samples of similar messages. A typical practical problem is to design equipment which will transmit English text at a given rate, while using the smallest possible channel capacity. The engineer will then, according to the usual viewpoint, need accurate data giving the correct frequencies of English text. Let's think about that a little more.

Suppose we try to characterize the English language, for purposes of communication theory, by specifying the relative frequencies of various letters, or combinations of letters. Now we all know that there is a great deal of truth in statements such as "the letter E occurs more frequently than the letter Z." Long before the days of communication theory, many people made obvious common-sense use of this knowledge. One of the earliest examples is the design of the Morse telegraphic code, in which the most frequently used letters are represented by the shortest codes--the exact prototype of what Shannon formalized and made precise a century later.

The design of our standard typewriter keyboard makes considerable use of knowledge of letter frequencies. This knowledge was used in a much more direct and drastic way by Ottmar Mergenthaler, whose immortal phrase

ETAOIN SHRDLU

was a common sight in the newspapers not so many years ago. But already we are getting into trouble, because there does not seem to be complete agreement even as to the relative order of the twelve most common letters in English, let alone the numerical values of their relative frequencies. For example, according to Pratt (1942) the above phrase should read

ETANOR ISHDLF

while Tribus (1961) gives it as

ETOANI RSHDLC

As we go into the less frequently used letters, the situation becomes still more chaotic.

Of course, we readily see the reason for these differences. People who have obtained different values for the relative frequencies of letters in English have consulted different samples of English text. It is obvious enough that the last volume of an encyclopaedia might have a higher relative frequency for the letter Z than the first volume. There is no reason to expect that letter frequencies would be the same in, say, a textbook on organic chemistry, a treatise on the history of Egypt, and a modern American novel. The writing of educated people would reveal systematic differences in word frequencies from the writings of people who had never gone beyond grade school. Even within a much narrower field, we would expect to find significant differences in letter and word frequencies in the writings of James Michener and Ernest Hemingway. The letter frequencies in the transcript of the tape recording of this lecture will probably be noticeably different from those I would produce if I sat down and wrote out the lecture verbatim.

The fact that statistical properties of a language vary with the author and circumstances of writing is so clear that it has become a useful research tool. A recent doctoral thesis in classics submitted to Columbia University by James T. McDonough (1961) contains a computer-run statistical analysis of

Homer's Iliad. Classicists have long debated whether all parts of the Iliad were written by the same man, and indeed whether Homer is an actual historical person. The analysis showed stylistic patterns consistent throughout the work. For example, 40.4 per cent of the 15,693 lines end on a word with one short syllable followed by two long ones, and a word of this structure never once appears in the middle of a line. Such consistency in a thing which is not a characteristic property of the Greek language, seems very strong evidence that the Iliad was written by a single person in a relatively short period of time, and it was not, as had been supposed by many nineteenth century classicists, the result of an evolutionary process over several centuries.

Of course, the evolutionary theory is not demolished by this evidence alone. If the Iliad was sung, we must suppose that the music had the very monotonous rhythmic pattern of primitive music, which persisted to a large extent as late as Bach and Haydn. Characteristic word patterns may have been forced on the composers, by the nature of the music. Archaeologists tell us that the siege of Troy, described in the Iliad, is not a myth but an historical fact, occurring about 1200 B. C., some four centuries before Homer. The decipherment of Minoan Linear B script by Michael Ventris in 1952 established that Greek existed already as a spoken language in the Aegean area several centuries before the siege of Troy; but the introduction of the Phoenician alphabet, which made possible a written Greek language in the modern sense, occurred only about the time of Homer. You see that the question is very complex and far from settled; but I find it fascinating that a statistical analysis of word and syllable frequencies, representing evidence which has been there in the Iliad for some twenty-eight centuries for anyone who had the wit to extract it, is now recognized as having a definite bearing on the problem. Undoubtedly, this is only the beginning of this type of analysis.

Well, to get back to communication theory, the point I am making is simply this: it is utterly wrong to say that there exists one and only one "true" set of letter or word frequencies for English text. If we use a mathematical model which presupposes the existence of such uniquely defined frequencies, we might easily end up proving things which, while perfectly valid as mathematical theorems, are worse than useless to an engineer who is faced with the job of actually designing a communication system to transmit English text efficiently.

But suppose our engineer does have extensive frequency data, and no other prior knowledge. How is he to make use of this in describing the information source? Many of the standard results of communication theory can, from the viewpoint I am advocating, be seen as simple examples of maximum-entropy inference; i.e. as examples of the same kind of reasoning as in statistical mechanics. To understand this was my original goal, discussed in Lecture 1.

20.3. Optimum Encoding: Letter Frequencies Known.

Suppose our alphabet consists of a different symbols $A_1, A_2, \ldots, A_a$, and we denote a general symbol by A_i, A_j, etc. Any message of N symbols then has the form $A_{i_1} A_{i_2} \ldots A_{i_N}$. We denote this message by M, which is a shorthand expression for the set of indices: $M = \{i_1\ i_2 \ldots i_N\}$. The number of conceivable messages is a^N. By $\sum_M$ I mean a sum over all of them. Also, define

$N_j(M) \equiv$ (number of times the letter A_j appears in the message M)

$N_{ij}(M) \equiv$ (number of times the digram $A_i A_j$ appears in M),

and so on.

Consider first an engineer E_1, who has a set of numbers $(f_1 \ldots f_a)$ giving the relative frequencies of the letters A_i, as observed in past samples of

messages, but has no other prior knowledge. What communication system represents rational design on the basis of this much information, and what channel capacity does E_1 require in order to transmit messages at a given rate of n symbols per second? To answer this, we need the probabilities p(M) which E_1 assigns to the various conceivable messages. Now Mr. E_1 has no deductive proof that the letter frequencies in future messages will be equal to the f_i observed in the past. On the other hand, his state of knowledge affords no grounds for supposing that the frequency of A_i will be greater than f_i rather than less, or vice versa. So he is going to suppose that frequencies in the future will be more or less the same as in the past, but he is not going to be too dogmatic about it. He can do this by requiring of the distribution p(M) only that it yield <u>expected</u> frequencies equal to the past ones. In other words,

$$\langle N_i \rangle = \sum_M N_i(M)\ p(M) = Nf_i, \qquad i = 1, 2, \ldots, a \qquad (20\text{-}25)$$

Of course, p(M) is not uniquely determined by these constraints, and so E_1 must at this point make a free choice of some distribution.

Let me emphasize again that it makes no sense to say that there exists any "physical" or "objective" distribution p(M) for this problem. This becomes especially clear if we suppose that only a single message is ever going to be sent over the communication system; thus there is no conceivable procedure by which p(M) could be measured as a frequency. But this would in no way affect the problem of engineering design which we are considering.

In choosing a distribution p(M), it would be perfectly possible for E_1 to assume some message structure involving more than single letters. For example, he might suppose that the digram A_1A_2 is more likely than A_3A_4. But from the standpoint of E_1 this could not be justified, for <u>as far as he knows</u>, a design based on any such assumption is as likely to hurt as to help. From E_1's standpoint, rational conservative design consists just in carefully

avoiding any such assumptions. This means, in short, that E_1 should choose the distribution p(M) by maximum entropy consistent with (20-25).

All the formalism of the maximum-entropy inference developed in Lecture 10 now becomes available to E_1. His distribution p(M) will have the form

$$\log p(M) + \lambda_0 + \lambda_1 N_1(M) + \lambda_2 N_2(M) + \ldots + \lambda_a N_a(M) = 0 \qquad (20\text{-}26)$$

and in order to evaluate the Lagrangian multipliers λ_i, he will use the partition function

$$Z(\lambda_1 \ldots \lambda_a) = \sum_M \exp[-\lambda_1 N_1(M) - \ldots - \lambda_a N_a(M)] = z^N \qquad (20\text{-}27)$$

where

$$z = e^{-\lambda_1} + \ldots + e^{-\lambda_a} \; . \qquad (20\text{-}28)$$

From (20-25) and the general relation

$$\langle N_i \rangle = - \frac{\partial}{\partial \lambda_i} \log Z(\lambda_1 \ldots \lambda_a) \qquad (20\text{-}29)$$

we find

$$\lambda_i = - \log(z f_i) \; , \qquad 1 \leq i \leq a \qquad (20\text{-}30)$$

and, substituting back into (20-26), we find the distribution which describes E_1's state of knowledge is just the multinomial distribution:

$$p(M) = f_1^{N_1} f_2^{N_2} \ldots f_a^{N_a} \qquad (20\text{-}26a)$$

which is a special case of an exchangeable sequence; the probability of any particular message depends only on how many times the letters A_1, A_2, ... appear, not on their order. The number of different messages possible for specified N_i is just the multinomial coefficient

$$\frac{N!}{N_1! \ldots N_a!} \; .$$

The entropy per symbol of the distribution (20-26a) is

$$H_1 = \frac{S}{N} = -\frac{1}{N}\sum_{M} p(M)\ \log p(M) = \frac{\log Z}{N} + \sum_{i=1}^{a} \lambda_i f_i$$

$$= -\sum_{i=1}^{a} f_i \log f_i \qquad (20\text{-}31)$$

Having found the assignment p(M), he can encode into binary digits in the most efficient way by a method found independently by R. M. Fano and C. E. Shannon (1948, Sec. 9). Arrange the messages in order of decreasing probability, and by a cut separate them into two classes so the total probability of all messages to the left of the cut is as nearly as possible equal to the probability of messages to the right. If a given message falls in the left class, the first binary digit in its code is 0; if in the right, 1. By a similar division of these classes into subclasses with as nearly as possible a total probability of 1/4, we determine the second binary digit, etc. I leave it for you to prove that (1) the expected number of binary digits required to transmit the message is numerically equal to H_1, when expressed in bits, and (2) in order to transmit at a rate of n of the original message symbols per second, E_1 requires a channel capacity $C \geq nH_1$, a result first given by Shannon.

The preceding mathematical steps are so well-known that they might be called trivial. However, the rationale which we have given them differs essentially from that of conventional treatments, and in that difference lies the main point of this section. Conventionally, one would use the frequency definition of probability, and say that E_1's probability assignment p(M) is the one resulting from the assumption that there are no intersymbol influences. Such a manner of speaking carries a connotation that the assumption might or might not be correct, and that its correctness must be demonstrated if the resulting design is to be justified; i.e. that the resulting

encoding rules might not be satisfactory if there are in fact intersymbol influences.

On the other hand, I contend that the probability assignment (20-26) is not an assumption at all, but the exact opposite. Eq. (20-26) represents, in a certain naive sense which I want to come back to later, the complete absence of any assumption on the part of E_1, beyond specification of expected single-letter frequencies, and it is uniquely determined by this property. The design based on (20-26) is the safest one possible on his state of knowledge. By that I mean the following. If, in fact, strong intersymbol influences do exist unknown to E_1, his encoding system will still be able to handle the messages perfectly well. If he had been given this additional information about intersymbol influences, he could have used it to arrive at an encoding system which would be still more efficient (i.e. would require a smaller channel capacity), as long as messages with only the specified type of correlation were transmitted. But if the type of intersymbol influence in the messages were suddenly to change, this new encoding system would likely become worse than the original one.

20.4. Better Encoding From Knowledge of Digram Frequencies.

Here is a rather long mathematical derivation which has, however, useful applications outside the particular problem at hand. Consider a second engineer, E_2. He has a set of numbers f_{ij}, $1 \leq i \leq a$, $1 \leq j \leq a$, which represent the expected relative frequencies of the digrams A_iA_j. E_2 will assign message probabilities $p(M)$ so as to agree with his state of knowledge,

$$\langle N_{ij} \rangle = \sum_M N_{ij}(M)\ p(M) = (N-1)\ f_{ij} \qquad (20\text{-}32)$$

and in order to avoid any further assumptions which are as likely to hurt as to help as far as he knows, he will determine the distribution $p(M)$ which has maximum entropy subject to this constraint. The problem is solved if he

can evaluate the partition function

$$Z(\lambda_{ij}) = \sum_{M} \exp\left[- \sum_{i,j=1}^{a} \lambda_{ij} N_{ij}(M) \right] \quad . \tag{20-33}$$

This can be done by solving the combinatorial problem of the number of different messages with given N_{ij}, or by observing that (20-33) can be written in the form of a matrix product:

$$Z = \sum_{i,j=1}^{a} \left(Q^{N-1} \right)_{ij} \tag{20-34}$$

where the matrix Q is defined by

$$Q_{ij} = e^{-\lambda_{ij}} \tag{20-35}$$

The result can be simplified formally if we suppose that the message $A_{i_1} \ldots A_{i_N}$ is always terminated by repetition of the first symbol A_{i_1}, so that it becomes $A_{i_1} \ldots A_{i_N} A_{i_1}$. The digram $A_{i_N} A_{i_1}$ is added to the message and an extra factor $\exp(-\lambda_{ij})$ appears in (20-33). The modified partition function then becomes a trace:

$$Z' = \mathrm{Tr}(Q^N) = \sum_{k=1}^{a} q_k^N \tag{20-36}$$

where the q_k are the roots of $\left| Q_{ij} - q\delta_{ij} \right| = 0$. This simplification would be termed "use of periodic boundary conditions" by the physicist. Clearly, the modification leads to no difference in the limit of long messages; as $N \to \infty$,

$$\lim \frac{1}{N} \log Z = \lim \frac{1}{N} \log Z' = \log q_{max} \tag{20-37}$$

where q_{max} is the greatest eigenvalue of Q.

The probability of a particular message is now a special case of (10-28):

$$p(M) = \frac{1}{Z} \exp[- \sum \lambda_{ij} N_{ij}(M)] \tag{20-38}$$

which yields the entropy as a special case of (10-34):

$$S = - \sum_{M} p(M) \log p(M)$$

$$= \log Z + \sum_{ij} \lambda_{ij} \langle N_{ij} \rangle \tag{20-39}$$

In view of (20-32) and (20-37), Mr. E_2's entropy per symbol reduces, in the limit $N \to \infty$, to

$$H_2 = \frac{S}{N} = \log q_{max} + \sum_{ij} \lambda_{ij} f_{ij} \tag{20-40}$$

or, since $\sum_{ij} f_{ij} = 1$, we can write (20-40) as

$$H_2 = \sum_{ij} f_{ij} (\log q_{max} + \lambda_{ij})$$

$$= \sum_{ij} f_{ij} \log \left(\frac{q_{max}}{Q_{ij}} \right) \tag{20-41}$$

Thus, to calculate the entropy we do not need q_{max} as a function of the λ_{ij} (which would be impractical for a > 3), but we need find only the ratio q_{max}/Q_{ij} as a function of the f_{ij}.

To do this, we first introduce the characteristic polynomial of the matrix Q:

$$D(q) \equiv \det(Q_{ij} - q\delta_{ij}) \tag{20-42}$$

and note, for later purposes, some well-known properties of determinants (Bocher, 1907, pp. 31-33). The first is

$$D(q)\delta_{ik} = \sum_{j=1}^{a} M_{ij} (Q_{kj} - q\delta_{kj})$$

$$= \sum_{j} M_{ij} Q_{kj} - qM_{ik} \tag{20-43a}$$

and similarly,

$$D(q)\delta_{ik} = \sum_{j} M_{ji} Q_{jk} - qM_{ki} \tag{20-43b}$$

in which M_{ij} is the cofactor of $(Q_{ij} - q\delta_{ij})$ in the determinant D(q); i.e. $(-)^{i+j} M_{ij}$ is the determinant of the matrix formed by striking out the i'th row and j'th column of the matrix (Q - ql). If q is any eigenvalue of Q,

the expressions (20-43) vanish for all choices of i and k.

The second identity applies only when q is an eigenvalue of Q. In this case, all minors of the matrix M are known to vanish. In particular, the second order minors are

$$M_{ik}M_{jl} - M_{il}M_{jk} = 0 \quad , \text{ if } D(q) = 0. \tag{20-44a}$$

This implies that the ratios (M_{ik}/M_{jk}) and (M_{ki}/M_{kj}) are independent of k; i.e. that M_{ij} must have the form

$$M_{ij} = a_i b_j \quad , \quad \text{if } D(q) = 0 \ . \tag{20-44b}$$

Substitution into (20-43a) and (20-43b) then shows that the quantities b_j form a <u>right eigenvector</u> of Q, while a_i is a <u>left eigenvector</u>:

$$\sum_j Q_{kj}b_j = qb_k \quad , \quad \text{if } D(q) = 0 \tag{20-43c}$$

$$\sum_i a_i Q_{ik} = qa_k \quad , \quad \text{if } D(q) = 0 \tag{20-43d}$$

Suppose now that any eigenvalue q of Q is expressed as an explicit function $q(\lambda_{11},\lambda_{12},\ldots,\lambda_{aa})$ of the Lagrangian multipliers λ_{ij}. Then, varying a particular λ_{kl} while keeping the other λ_{ij} fixed, q will vary so as to keep D(q) identically zero. By the rule for differentiating the determinant (20-42), this gives

$$\frac{dD}{d\lambda_{kl}} = \frac{\partial D}{\partial \lambda_{kl}} + \frac{\partial D}{\partial q}\frac{\partial q}{\partial \lambda_{kl}}$$

$$= - M_{kl}Q_{kl} - \frac{\partial q}{\partial \lambda_{kl}} \mathrm{Tr}(M) = 0 \tag{20-45}$$

where

$$\mathrm{Tr}(M) \equiv \sum_{i=1}^{a} M_{ii} \tag{20-46}$$

is the trace, or diagonal sum, of the matrix M.

Using this relation, the condition (20-32) fixing the Lagrangian multipliers λ_{ij} in terms of the prescribed digram frequencies f_{ij}, becomes

$$f_{ij} = -\frac{\partial}{\partial \lambda_{ij}} \log q_{max} = \frac{M_{ij} Q_{ij}}{q_{max} Tr(M)} \tag{20-47}$$

The single-letter frequencies are proportional to the diagonal elements of M:

$$f_i = \sum_{j=1}^{a} f_{ij} = \frac{M_{ii}}{Tr(M)} \tag{20-48}$$

where we have used the fact that (20-43a) vanishes for $q = q_{max}$, $i = k$. Thus, from (20-47) and (20-48), the ratio needed in computing the entropy per symbol is

$$\frac{Q_{ij}}{q_{max}} = \frac{f_{ij}}{f_i} \frac{M_{ii}}{M_{ij}} = \frac{f_{ij}}{f_i} \frac{b_i}{b_j} \tag{20-49}$$

where we have used (20-44b). Substituting this into (20-41), we find that the terms involving b_i and b_j cancel out, and E_2's entropy per symbol is just

$$\begin{aligned} H_2 &= -\sum_{ij} f_{ij} \log \left(\frac{f_{ij}}{f_i} \right) \\ &= -\sum_{ij} f_{ij} \log f_{ij} + \sum_i f_i \log f_i \end{aligned} \tag{20-50}$$

This is never greater than E_1's H_1, for from (20-31), (20-50),

$$\begin{aligned} H_2 - H_1 &= \sum_{ij} f_{ij} \log \left(\frac{f_i f_j}{f_{ij}} \right) \\ &\leq \sum_{ij} f_{ij} \left[\frac{f_i f_j}{f_{ij}} - 1 \right] = 0 \end{aligned}$$

where we used the fact that $\log x \leq x - 1$ with equality if and only if $x = 1$. Therefore,

$$H_2 \leq H_1 \tag{20-51}$$

with equality if and only if $f_{ij} = f_i f_j$, in which case E_2's extra information was only what E_1 would have inferred. To see this, note that in the message

$M = \{i_1 \ldots i_N\}$, the number of times the digram $A_i A_j$ occurs is

$$N_{ij}(M) = \delta(i,i_1)\delta(j,i_2) + \delta(i,i_2)\delta(j,i_3) + \ldots + \delta(i,i_{N-1})\delta(j,i_N) \qquad (20\text{-}52)$$

and so, if we ask E_1 to estimate the frequency of digram $A_i A_j$, by the criterion of minimizing the expected square of the error, he will make the estimate

$$\langle f_{ij} \rangle = \frac{\langle N_{ij} \rangle}{N-1} = \frac{1}{N-1} \sum_M p(M)\, N_{ij}(M) = f_i f_j \qquad (20\text{-}53)$$

using for $p(M)$ the distribution (20-26a) of E_1. In fact, the distributions $p(M)$ found by E_1 and E_2 are identical if $f_{ij} = f_i f_j$, for then we have from (20-47), (20-48), and (20-44b),

$$Q_{ij} = e^{-\lambda_{ij}} = q_{max}\sqrt{f_i f_j} \quad . \qquad (20\text{-}54)$$

Using (20-37), (20-52), and (20-54), we find that E_2's distribution (20-38) reduces to (20-26a). This is a rather nontrivial example of what we noted in Lecture 10, Eq. (10-76).

20.5. Relation to a Stochastic Model.

The quantities introduced above acquire a deeper meaning in terms of the following problem. Suppose that part of the message has been received, what can Mr. E_2 then say about the remainder of the message? This is answered by recalling our Rule 1:

$$(AB|X) = (A|BX)(B|X)$$

or, the conditional probability of A, given B, is

$$(A|BX) = \frac{(AB|X)}{(B|X)} \qquad (20\text{-}55)$$

a relation which in conventional theory, which does not use X, is taken as the definition of a conditional probability (i.e., as a ratio of two "absolute" probabilities). In our case, let X stand for the general statement of the

problem leading to the solution (20-38), and let

$$B \equiv \text{"The first } (m-1) \text{ symbols are } \{i_1 i_2 \cdots i_{m-1}\}."$$

$$A \equiv \text{"The remainder of the message is } \{i_m \cdots i_N\}."$$

Then $(AB|X)$ is the same as $p(M)$ in (20-38). Using (20-52), this reduces to

$$(AB|X) = (i_1 \cdots i_N|X) = Z^{-1} Q_{i_1 i_2} Q_{i_2 i_3} \cdots Q_{i_{N-1} i_N} \tag{20-56}$$

and in

$$(B|X) = \sum_{i_m=1}^{a} \cdots \sum_{i_N=1}^{a} (i_1 \cdots i_N|X) \tag{20-57}$$

the sum generates a power of the matrix Q, just as in the partition function (20-34). Writing, for brevity, $i_{m-1} = i$, $i_m = j$, $i_N = k$, and

$$R \equiv \frac{1}{Z} Q_{i_1 i_2} \cdots Q_{i_{m-2} i_{m-1}} , \tag{20-58}$$

we have

$$(B|X) = R \sum_{k=1}^{a} (Q^{N-m+1})_{ik} = R \sum_{j,k=1}^{a} Q_{ij} (Q^{N-m})_{jk} \tag{20-59}$$

and so

$$(A|BX) = \frac{Q_{ij} Q_{i_m i_{m+1}} \cdots Q_{i_{N-1} i_N}}{\sum_{k=1}^{a} (Q^{N-m+1})_{ik}} \tag{20-60}$$

since all the Q's contained in R cancel out, we see that the probabilities for the remainder $\{i_m \cdots i_N\}$ of the message depend only on the immediately ppeceding symbol A_i, and not on any other details of B. This property defines a Markov Chain. There is a huge literature dealing with them; it is perhaps the most thoroughly worked out branch of probability theory. The basic tool, from which essentially all else follows, is the matrix p_{ij} of "elementary transistion probabilities." This is the probability $p_{ij} = (A_j|A_i X)$ that the next symbol will be A_j, given that the last one was A_i. Summing (20-60) over $i_{m+1} \cdots i_N$, we find

$$P_{ij}^{(N)} = (A_j | A_i X) = \frac{Q_{ij} - T_j}{\sum_j Q_{ij} T_j} \tag{20-61}$$

where

$$T_j \equiv \sum_{k=1}^{a} (Q^{N-m})_{jk} \tag{20-62}$$

The fact that T_j depends on N and m is an interesting feature. Usually, one considers from the start a chain indefinitely prolonged, and so it is only the limit of (20-61) for $N \to \infty$ that is ever considered. This example shows that prior knowledge of how long the chain is going to be can affect the transition probabilities; however, the limiting case is clearly of greatest interest.

To find this limit we need a little more matrix theory. The equation $D(q) = \det(Q_{ij} - q\delta_{ij}) = 0$ has a roots $(q_1 q_2 \ldots q_a)$, not necessarily all different, or real. Label them so that $|q_1| \geq |q_2| \geq \ldots \geq |q_a|$. There exists a nonsingular matrix A such that $A\,Q\,A^{-1}$ takes the canonical "superdiagonal" form:

$$A\,Q\,A^{-1} = \bar{Q} = \begin{pmatrix} C_1 & 0 & 0 & \ldots \\ 0 & C_2 & 0 & \ldots \\ 0 & 0 & C_3 & \ldots \\ \vdots & \vdots & \vdots & \\ \cdot & \cdot & \cdot & C_m \end{pmatrix} \tag{20-63}$$

where the C_i are sub-matrices which can have either of the forms

$$C_i = \begin{pmatrix} q_i & 1 & 0 & 0 & \ldots & & \\ 0 & q_i & 1 & 0 & \ldots & & \\ 0 & 0 & q_i & 1 & \ldots & & \\ \cdot & \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & \cdot & & q_i & 1 \\ \cdot & \cdot & \cdot & \cdot & & & \\ & & & & & 0 & q_i \end{pmatrix} \quad \text{or,} \quad C_i = \begin{pmatrix} q_i & & & \\ & q_i & & \\ & & \ddots & \\ & & & q_i \end{pmatrix} \tag{20-64}$$

The result of raising Q to the n'th power is

$$Q^n = A \bar{Q}^n A^{-1} \tag{20-65}$$

and as $n \to \infty$, the elements of $\bar{Q}^n$ arising from the greatest eigenvalue $q_{max} = q_1$ become arbitrarily large compared to all others. If q_1 is nondegenerate, so that it appears only in the first row and column of $\bar{Q}$, we have

$$\lim_{N\to\infty} \frac{T_j}{q_1^{N-m}} = A_{j1} \sum_{k=1}^{a} (A^{-1})_{1k} \quad , \tag{20-66}$$

$$\lim_{N\to\infty} \frac{T_j}{\sum_j Q_{ij} T_j} = \frac{A_{j1}}{q_1 A_{i1}} \quad , \tag{20-67}$$

and the limiting transition probabilities are

$$p_{ij}^{(\infty)} = \frac{Q_{ij}}{q_1} \frac{A_{j1}}{A_{i1}} = \frac{Q_{ij}}{q_1} \frac{M_{ij}}{M_{ii}} \tag{20-68}$$

where we have used the fact that the elements A_{j1} (j = 1, 2, ..., a) form an eigenvector of Q with eigenvalue $q_1 = q_{max}$, so that, referring to (20-44b), (20-44c), $A_{j1} = Kb_j$ where K is some constant. Using (20-47), (20-48), we have finally,

$$p_{ij}^{(\infty)} = \frac{f_{ij}}{f_i} \tag{20-69}$$

which is just what would be taken, on the frequency theory, as the <u>definition</u> of the transition probability.

Lecture 24

IRREVERSIBLE STATISTICAL MECHANICS I: HEIMS PERTURBATION THEORY

Back in Lectures 10 and 11, we saw how the principle of maximum entropy leads us to the standard Gibbs formalism of equilibrium statistical mechanics, via arguments very much shorter and simpler than the usual "ergodic" approach of antiquity. The principle is therefore, at the very least, a useful pedagogical device, by which known results may be derived more quickly. But, of course, the real test of any new principle in science is not its ability to re-derive known results, but its ability to give new results, which could not be (or at least, had not been) derived without it. Since we agree with standard formalism in all equilibrium problems, the only place where new results are possible is in the extension to nonequilibrium problems, where previously no general theory existed.

Another respect in which Lecture 11 was left incomplete, appears as soon as we try to apply that formalism to real, nontrivial physical problems; we need more powerful mathematical tools. It is one of the most satisfying things about this approach that both these needs--finding a mathematical technique for complicated equilibrium problems, and setting up a general formalism for nonequilibrium problems--are met by a single mathematical development. I'll give it in this Lecture, and we'll see its applications to equilibrium and nonequilibrium problems in the next two lectures.

24.1 Density Matrix Formulation

So far, I've worked up a formalism in which the enumeration of the possible "states of nature" could take place simply by listing all the stationary quantum states. In other words, quantities that are constants of the motion are the only things that I have allowed myself to specify so far. Evidently, if we are ever going to get to non-equilibrium theory, we have to generalize this to the case where I'm putting in information about things which are not constants of the motion, so something can happen when we let the equations of motion take over. If we started out with the initial canonical probability assignments of Lecture 11 and then solved the Schroedinger equation for the time development, we would find nothing at all happening. It would just sit there. Of course, that is as it ought to be for the equilibrium case; but for the non-equilibrium case, we need a little bit more.

Also, as just noted, even in the equilibrium case, I need to generalize this before I can actually do the calculation for non-trivial physical problems, because in practice I don't have the kind of information assumed above. The theory given so far presupposes an enumeration of the exact energy levels in my system to start with. But in a realistic problem, I can't calculate these. What we know is a Hamiltonian which, in the cases we can actually solve, can eventually be split into a term H_o which is big but simple and another term H_1 which is complicated but small,

$$H = H_o + H_1 \tag{24-1}$$

Then we have to do some kind of perturbation theory in order to find approximate values for the energy levels defined by the entire Hamiltonian. To find them exactly is a problem that we haven't solved.

It will happen in all non-trivial problems that the H_1 simply does not commute with H_o. So we have to learn how to generalize this mathematical machine so we can put in information about quantities which don't commute with each other. I can't enumerate states of nature simply by citing energy levels; in fact, I don't even know the representation in which this would be possible. For this reason, in any representation I can find, the relative phase of these quantum states has to get into the picture even for equilibrium problems. Well, we know the way to do this is to restate this theory in terms of the density matrix; let's turn to that now.

First, let's recall our basic definition of the density matrix. Again, this is perfectly standard material which is in fifty textbooks on quantum theory and statistical mechanics by now. Suppose that I have a state of knowledge about a system; and for the time being, don't worry about how I got this state of knowledge. I just want to describe it. There are various states ψ_1, ψ_2, ..., in which the system might be. I don't know which one it is. All I know is described by assigning some probability w_i to it being in the state ψ_i. Now, if I knew the system was in a definite quantum state ψ_i I could calculate the expectation value of any operator and come out with some formula like this,

$$\langle F \rangle_i = \int \psi_i^* F \psi_i \, d\tau \tag{24-2}$$

where $\int d\tau$ stands for an integration over all particle co-ordinates and, if there are spin indices in the problem, for summation over all those. Now the ψ_i functions that I started with are not necessarily orthogonal functions. They could be any old set of conceivable states of the system. But each of them could be expanded in a complete orthogonal set. Let's say that u_k are a complete orthonormal set of functions in which we can expand any state of this system. For the moment, it doesn't matter what states they are; just

any set that we know is complete. We could expand ψ_i in terms of those, getting some expansion coefficients $a_k^{(i)}$:

$$\psi_i = \sum_k u_k \, a_k^{(i)}$$

and then write

$$\langle F \rangle_i = \int \left(\sum_k u_k^* \, a_k^{(i)*} \right) F \left(\sum_j u_j \, a_j^{(i)} \right) d\tau \quad . \tag{24-4}$$

Now the a_k^* and a_j are constants which can be taken outside,

$$\langle F \rangle_i = \sum_{kj} a_k^{*(i)} \, a_j^{(i)} \int u_k^* \, F \, u_j \, d\tau \tag{24-5}$$

and the integral (or sum)

$$\int u_k^* \, F \, u_j \, d\tau = F_{kj} \tag{24-6}$$

defines the matrix element F_{kj}, in the u_k representation, so that

$$\langle F \rangle_i = \sum_{kj} F_{kj} \, a_k^{(i)*} \, a_j^{(i)} \quad . \tag{24-7}$$

The expectation value of any quantity, if I am given the wave function ψ_i, is a quadratic form in these matrix elements F_{kj}. Now if I'm in this fix where I don't know what the state is, the best expectation value I can give you is not just one of these, but I have to average it also over these w_i which represent my uncertainty as to what the actual state is,

$$\langle F \rangle = \sum_i w_i \langle F \rangle_i = \sum_i w_i \sum_{jk} F_{kj} \, a_k^{(i)*} \, a_j^{(i)} \quad . \tag{24-8}$$

Our expectation values are now double averages. Even if I knew the exact quantum state, there are still statistical things in quantum theory (or, to put it more cautiously, in the current "Copenhagen" interpretation of that theory), which would allow me to give only expectation values in general. I'm not even that well off. I don't even know what the right state is, so I have to average over that ignorance (w_i) also.

When you have a thing like (24-8), the only thing you can possibly do with it is change the order of summations and see what happens. Let me do that;

$$\langle F \rangle = \sum_{jk} F_{kj} \sum_i w_i \, a_k^{(i)*} \, a_j^{(i)} \quad .$$

Now, define a matrix ρ by

$$\sum_i w_i \, a_k^{(i)*} \, a_j^{(i)} = \rho_{jk} \tag{24-9}$$

then

$$\langle F \rangle = \sum_{jk} F_{kj} \, \rho_{jk} \quad . \tag{24-10}$$

The summation over j builds me the matrix product of $F\rho$; and then the summation over k is the sum of the diagonal elements, which we call the trace. Or, I could have written the sum with ρ and F interchanged. In this case I would now say the summation over k builds me the matrix product ρF, and then the summation over j gives the trace, so I could write this equally well as

$$\langle F \rangle = \sum_{jk} F_{kj} \, \rho_{jk} = \mathrm{Tr}(F\rho) = \mathrm{Tr}(\rho F) \quad . \tag{24-11}$$

This matrix ρ is, of course, called the density matrix, and you see that it is a Hermitian matrix,

$$\rho^*_{kj} = \rho_{jk} \quad ,$$

or in matrix notation

$$\rho^\dagger = \rho \quad . \tag{24-12}$$

The neat way to develop our quantum statistics, so the phases are taken into account automatically, is in terms of the density matrix. From now on I will express expectation values of any quantities I want to talk about in the form (24-11). We started out with a problem of how you set up a probability assignment which describes a certain state of knowledge; now we've got the problem of setting up a density matrix which describes a certain state of knowledge. Take a specific case; suppose somebody measures the total magnetic moment of some spin system and they give me a number M. I want to find a density matrix which describes what I know about this spin system when you give me just this number; or rather these three numbers, the three components $\{M_x, M_y, M_z\}$. At the very least I want my density matrix to satisfy

$$\vec{M} = \langle \vec{M}_{op} \rangle = \mathrm{Tr}\left(\rho \vec{M}_{op}\right) \quad . \tag{24-13}$$

In other words, if I give this density matrix to anybody else, and he tries to predict the moments from the density matrix, he should be able to get back the numbers that were given to me, by following the usual rule for prediction in statistical mechanics. If he couldn't do that, then it wouldn't make sense to say that the density matrix "contained" the given information $\{M_x, M_y, M_z\}$. This is all we are doing when we choose ρ to satisfy (24-13).

In general, there are an infinite number of density matrices which would all do this. Again, I am faced with the problem of making a free choice of a density matrix, which is "honest" in the sense that it doesn't assume things that I don't know, and spreads out the probability as evenly as possible over all possibilities allowed by what I do know. Well, we started out with

$$S_I = -\sum_i p_i \log p_i$$

so, suppose we now take

$$S_A = -\sum_i w_i \log w_i \qquad (24\text{-}14)$$

and we might choose the density matrix which makes S_A a maximum. If we took that as our measure of amount of uncertainty, we would be in a little bit of trouble. A sort of Gibbs paradox would show up. I said that these initial states ψ_i that we started out with are not necessarily orthogonal to each other. In fact, I can have state ψ_1 and I give it a probability w_1; to the state ψ_2 I give probability w_2. Now, let's make a continuous change in the problem such that $\psi_2 \Longrightarrow \psi_1$; my state of knowledge shades continuously into: ψ_1 with a probability (w_1+w_2). But nothing like that happens to S_A. In S_A as $\psi_2 \Longrightarrow \psi_1$ the term $-w_1 \log w_1 - w_2 \log w_2$ would have to be replaced suddenly by

$$-(w_1+w_2)\log(w_1+w_2) \quad .$$

If we took this quantity S_A as the measure of uncertainty about the system, then you would have this phenomenon of sudden discontinuities in my uncertainty when two wave functions suddenly become exactly equal. But my intuitive state of knowledge has no discontinuity when I do that. It goes continuously from one case to another. That's one thing that would be wrong if I tried to use this S_A as my measure of uncertainty.

There's another thing that would be even worse, and perhaps easier to see. For a given density matrix, there's no upper limit to the S_A that I could get. If S_A is going to be the thing that counts, I'll say I've got 26 different states, ψ_a to ψ_z. They all happen to be equal to ψ_1 but I assign probabilities w_a to w_z to them. Now, of course, my summation

$$-\sum_{a=1}^{26} w_a \log w_a$$

over the alphabet--my notation is not quite consistent, but I think you see the point--my summation over all these terms could be a very large number. I can introduce thousands of them. There would be no upper limit to the $-\sum w \log w$ I could get if I used this S_A.

On the other hand, there's one property that is unique. S_A has no upper bound. S_A does have a lower bound. S_A for a given density matrix has an absolute minimum given by

$$S_A \geq -\mathrm{Tr}(\rho \log \rho) \ . \qquad (24\text{-}15)$$

There's one and only only one way, in general, of setting up these states ψ_i and corresponding probabilities w_i so that this lower bound is reached. When I say "in general," I mean if there are no degeneracies in the eigenvalues of ρ. I think that I will not bother to give you the proof of this. The proof is given in this second paper that I had a long time ago [Physical Review 108, 171, (1957)].

Well, now what does $\log \rho$ mean? I have to do that for the next step. ρ is a Hermitian matrix and there's a theorem in matrix theory that says, there is a matrix S such that

$$S\rho S^{-1} = \begin{pmatrix} \rho_1 & & & \\ & \rho_2 & & \\ & & \ddots & \\ & & & \rho_n \end{pmatrix} \qquad (24\text{-}16)$$

I can always find some similarity transformation which would have made this diagonal. Now, in the representation where ρ is diagonal, then by $\log \rho$ I mean the diagonal matrix

$$\log \rho = \begin{pmatrix} \log \rho_1 & & & \\ & \log \rho_2 & & \\ & & \ddots & \\ & & & \log \rho_n \end{pmatrix} \qquad (24\text{-}17)$$

If I choose for my basis u_k the particular set of functions ψ_i for which S_A does reach its absolute minimum value, then the diagonal elements of ρ are just the probabilities w_i assigned to these states. In other words, the choice of possible states ψ_i which makes S_A a minimum for a given ρ, is the one for which the probabilities w_i assigned to these states are the eigenvalues of this matrix ρ.

The reason we had a Gibbs paradox at the beginning here was that I said these different states ψ_i that I'm taking into account are not necessarily orthogonal. If states ψ_1 and ψ_2 are not orghogonal and you tell me the system is in state ψ_1, then, of course, the present Copenhagen interpretation says: the probability that, if I did a measurement, I would actually find it in ψ_2, is not zero. It's the scalar product squared, $|(\psi_1,\psi_2)|^2$; sometimes called the transition probability from one state to another. I'm not writing down the probabilities of mutually exclusive events unless I choose my states ψ_i to be orthogonal, and that's just what I do by making the choice that

minimize S_A. I'm going to say now that the S_I for a density matrix is this unique minimum value of S_A:

$$S_I \equiv -\mathrm{Tr}(\rho \log \rho) \tag{24-18}$$

There are a number of other arguments why you choose (24-18) rather than some other expressions that you could think of, and they are also given in this previously mentioned paper. I want not to show you some examples of equilibrium statistical mechanics using this and I want to develop a general perturbation theory in which, if there's a complicated problem I can break it down into a simple problem plus a small change. I want to learn how to expand this density in powers of some small perturbation and the perturbation theory we get will also be exactly the one we need for our irreversible theory tomorrow. Now, we are back at the same problem that we studied in Lecture 10, but the $\langle F_k \rangle$ are matrices, and the constraints are

$$\langle F \rangle_k = \mathrm{Tr}(\rho F_k), \quad k = 1, 2, \ldots, m \quad . \tag{24-19}$$

This restricts my density matrix. I must find which density matrix will maximize S_I while agreeing with conditions we have imposed on it. Now, the formal solution of this goes through in exactly the same way as we did in Lecture 10. In fact, you recall that my proof back then was based on the fact that when I have an ordinary discrete probability distribution

$$\sum_{i=1}^{n} p_i \log p_i \geq \sum_{i=1}^{n} p_i \log u_i \tag{24-20}$$

the inequality becomes an equality if, and only if, $p_i = u_i$. Now, we have a precisely similar situation here. You can prove that if ρ and σ are any two density matrices, there is an inequality

$$\mathrm{Tr}\{\rho \log \rho\} \geq \mathrm{Tr}\{\rho \log \sigma\} \tag{24-21}$$

I'll leave this as an "exercise for the reader" to prove. The argument goes through precisely the way I did it before. The solution to this problem was given long ago by von Neumann (Göttinger Nachrichten, 1927).

The mathematical properties that I am talking about have been well known for a long time; but the new viewpoint about the <u>significance</u> of those properties is the thing that I'm selling here. These properties provide the justification for choosing certain distributions in preference to certain other ones. The density matrix that maximizes S_I subject to these constraints is again given by

$$\rho = \frac{1}{Z(\lambda_1 \cdots \lambda_m)} \exp\left\{-\lambda_1 F_1 - \cdots \cdots - \lambda_m F_m\right\} \tag{24-22}$$

One would guess, of course, that it generalizes in some such way as this, but I don't think your intuition would tell you whether the proper generalization was exactly this form. All the formal properties that I wrote down this morning follow from this distribution just the same way that we gave before with one exception, which I'll get to after we've developed our mathematics a little bit more.

Of course the number one must have expectation value of one,

$$\langle 1 \rangle = \mathrm{Tr}(\rho 1) = \mathrm{Tr}(\rho) \quad . \tag{24-23}$$

This is one more condition just like the one this morning that p_i had to be equal to one. The normalizing factor which will guarantee that the trace of this thing is one, is evidently

$$Z(\lambda_1 \ldots\ldots \lambda_m) = \mathrm{Tr}\ \exp\left\{-\lambda_1 F_1 - \ldots\ldots - \lambda_m F_m\right\} \tag{24-24}$$

which is the partition function.

Now perhaps I ought to say a word about what is meant by the exponential of a matrix. If I have any function of an ordinary number x that I can expand in a power series,

$$f(x) = \sum_{n=0}^{\infty} a_n\, x^n \quad , \tag{24-25}$$

of course, there is nothing to stop me from defining the same function of a matrix by the same power series,

$$f(M) = \sum_{n=0}^{\infty} a_n\, M^n \quad . \tag{24-26}$$

Then the question arises; does this converge to a definite matrix and if so does the resultin matrix f(M) have any useful properties? There is a theorem: if the original power series converged for x equal to each of the eigenvalues of the matrix M, then the power series is guaranteed to converge to a definite matrix f(M). Now in particular the exponential function,

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{24-27}$$

converges so well it has infinite radius of convergence and, therefore, the exponential of any square matrix with finite elements is guaranteed to exist and to be a well defined matrix.

The choosing of the λ_k is again something which we do in order to make the expectation values agree with the given data. Again it's going to turn out that same formal relations hold when we are talking matrices. Again we have to solve

$$\langle F_k \rangle = -\frac{\partial}{\partial \lambda_k} \log Z \tag{24-28}$$

for the λ_k. But to prove that this is right, we have to give a mathematical argument that is a little more involved than that needed to prove (It turns out that this argument is also fundamental to everything that I want to talk about from now on, so let's take time out for it now.

24.2 Heims Perturbation Theory

I would like to develop what I call the Heims' perturbation theory. This was worked out in about 1959 by my former student, Steve Heims, and we published a very truncated account of it in the appendix to a paper on gyromagnetic effects [Revs. Mod. Physics 34, 143 (1962)]. You see we have always the problem of evaluating exponentials of matrices. First, I would like to work out the well-known perturbation expansion of this. I have a matrix A, and the matrix e^A is something that I can do. That is simple. But the thing I really want to evaluate is

$$e^{(A+\text{something else})}$$

or

$$e^{A+\varepsilon B} = e^A\left[1 + \sum_{n=1}^{\infty} \varepsilon^n S_n\right] \quad . \tag{24-29}$$

And I will indicate that this something else is small by putting ε in it and expanding in powers of ε. You see this is the typical situation we would have if we tried to evaluate a density matrix

$$\rho = \frac{1}{Z} \exp\left\{-\lambda_1 F_1 - \lambda_2 F_2 - \cdots - \lambda_m F_m\right\} \quad . \tag{24-30}$$

Some of these operators might be simple so I could evaluate their exponentials; then some others might be complicated and not commute with the others, and they would mess up the whole problem. At that point I would resort to approximations. To put it in general form, let's talk just A and B for a while.

Let me form a quantity

$$e^{-xA}\, e^{x(A+\varepsilon B)}$$

where x is an ordinary number and by xA I mean the matrix in which every element is multiplied by x. If I let ε go to zero, this goes into the unit matrix. But it isn't quite the unit matrix, if ε is not zero. But how does it vary with x? Well, by staring at this power series definition of the exponential function, you can convince yourself very quickly that the same rule for differentiating an exponential function works even if a matrix is in the exponent. I have my choice of writing it either way:

$$\frac{d}{dx}e^{-xA} = -Ae^{-xA} = -e^{-xA}\, A \qquad . \tag{24-31}$$

Therefore,

$$\frac{d}{dx}\left[e^{-xA}e^{x(A+\varepsilon B)}\right] = -e^{-xA}\, Ae^{x(A+\varepsilon B)} + e^{-xA}(A+\varepsilon B)e^{x(A+\varepsilon B)} \tag{24-32}$$

Now two terms cancel, and ε is just a number, so

$$\frac{d}{dx}\left[e^{-xA}e^{x(A+\varepsilon B)}\right] = e^{-xA}\, Be^{x(A+\varepsilon B)} \qquad . \tag{24-33}$$

I can't pull that B outside because in general it doesn't commute with what is either to the left of it or to the right of it. Now that I've differentiated this thing, let me integrate with respect to x and get it back again:

$$\int_0^x \frac{d}{dx_1}\left[e^{-x_1A}\, e^{x_1(A+\varepsilon B)}\right]dx_1 = e^{-xA}e^{x(A+\varepsilon B)} - I$$

$$= \varepsilon\int_0^x e^{-x_1A}\, Be^{x_1(A+\varepsilon B)}\, dx_1 \qquad . \tag{24-34}$$

Now let me clean this up. Multiplying both sides by e^{xA} from the left, we find

$$e^{x(A+\varepsilon B)} = e^{xA}\left[1 + \varepsilon\int_0^x e^{-x_1 A}\, B e^{x(A+\varepsilon B)}\, dx_1\right]. \qquad (24\text{-}35)$$

This is an integral equation which $e^{x(A+\varepsilon B)}$ satisfies. Well now, if you have an integral equation, you grind out perturbation solutions of it simply by iteration--i.e., substituting the equation into itself over and over again. So, let me write this in still easier form,

$$e^{x(A+\varepsilon B)} = e^{xA}\left[1 + \varepsilon\int_0^x e^{-x_1 A}\, B e^{x_1(A+\varepsilon B)}\, dx_1\right] \quad . \qquad (24\text{-}36)$$

The first iteration gives

$$e^{x(A+\varepsilon B)} = e^{xA}\left\{1 + \varepsilon\int_0^x dx_1 e^{-x_1 A}\, B e^{x_1 A}\left[1 + \varepsilon\int_0^{x_1} dx_2 e^{-x_2 A}\, B e^{x_2(A+\varepsilon B)}\right]\right\}$$

$$= e^{xA}\left\{1 + \varepsilon\int_0^x dx_1 e^{-x_1 A}\, B e^{x_1 A} + \varepsilon^2\int_0^x dx_1\int_0^{x_1} dx_2 e^{-x_1 A}\, B e^{(x_1-x_2)A}\, B e^{x_2(A+\varepsilon B)}\right\},$$

and by repeated substitution we get

$$e^{A+\varepsilon B} = e^{A}\left[1 + \varepsilon\int_0^1 e^{-xA}\, B e^{xA}\, dx + \right.$$

$$\varepsilon^2\int_0^1 dx_1\int_0^{x_1} dx_2 e^{-x_1 A}\, B e^{(x_1-x_2)A}\, B e^{x_2 A} +$$

$$\left. B e^{(x_2-x_3)A}\, B e^{x_3 A} + \ldots\right] \varepsilon^3\int_0^1 dx_1\int_0^{x_1} dx_2\int_0^{x_2} dx_3 e^{-x_1 A}\, B e^{(x_1-x_2)A} \quad . \qquad (24\text{-}37)$$

We can keep playing this game as long as we please, and so this generates an infinite series in powers of ε. Or, we can terminate (24-37) at any finite number of terms, replace A by A + εB in the last exponent, and it is an exact equation. The exponential of any matrix is a well-behaved thing, so

we can put in any ε we please--large or small-- and the infinite series is guaranteed to converge to the right thing. Of course, if we have to take more than about two terms of the series, then we'll be wound up in another bad calculation and this whole method will not be too useful.

Let's summarize this: we have found the power series expansion

$$e^{A+\varepsilon B} = e^{A}\left[1 + \sum_{n=1}^{\infty} \varepsilon^{n} S_{n}\right] \tag{24-38}$$

in which

$$S_1 \equiv \int_0^1 e^{-xA} B e^{xA} dx \tag{24-39}$$

$$S_2 \equiv \int_0^1 dx_1 \int_0^{x_1} dx_2 \; e^{-x_1 A} B e^{(x_1 - x_2)} B e^{x_2 A} \tag{24-40}$$

and if we write

$$B(x) \equiv e^{-xA} B e^{xA} \tag{24-41}$$

the general order term is

$$S_n \equiv \int_0^1 dx_1 \int_0^{x_1} dx_2 \cdots \int_0^{x_{n-1}} dx_n \; B(x_1)B(x_2)\cdots B(x_n) \quad . \tag{24-42}$$

Now we have an "unperturbed" density matrix

$$\rho_o = \frac{e^{A}}{Tr\left(e^{A}\right)} \tag{24-43}$$

and a "perturbed" one:

$$\rho = \frac{e^{A+\varepsilon B}}{Tr\left[e^{A+\varepsilon B}\right]} \tag{24-44}$$

In the unperturbed ensemble, any particular operator C has the expectation value

$$\langle C\rangle_o = \mathrm{Tr}(\rho_o C) \tag{24-45}$$

and in the perturbed ensemble, it will be instead,

$$\langle C\rangle = \mathrm{Tr}(\rho C) \quad . \tag{24-46}$$

And what I would really like to get is a power series expansion of <C>. So let's write out the expansion we would like to get; using (24-38),

$$\langle C\rangle = \frac{\mathrm{Tr}\left[e^{A+\varepsilon B}C\right]}{\mathrm{Tr}\left[e^{A+\varepsilon B}\right]} = \frac{\mathrm{Tr}\left(e^{A}C\right) + \sum_{n=1}^{\infty} \varepsilon^n \mathrm{Tr}\left(e^{A}S_n C\right)}{\mathrm{Tr}\left(e^{A}\right) + \sum_{i=1}^{\infty} \varepsilon^n \mathrm{Tr}\left(e^{A}S_n\right)}$$

and divide by $\mathrm{Tr}\left(e^{A}\right)$ to get, from (24-45),

$$\langle C\rangle = \frac{\langle C\rangle_o + \sum_{n=1}^{\infty} \varepsilon^n \langle S_n C\rangle_o}{1 + \sum_{n=1}^{\infty} \varepsilon^n \langle S_n\rangle_o} \tag{24-47}$$

I've got everything reduced to expectation values calculated in the unperturbed distribution, which I assumed was something simple that I could calculate. But still this is in a little messy form. I've got the ratio of two infinite series--I know they are well-behaved series. Both the numerator and denominator series have infinite radius of convergence. But, I would like to write this as a single series over ε and get rid of this denominator. If I can invert the power series for this denominator; that is, find the coefficients a_n in

$$\frac{1}{1 + \sum_{n=1}^{\infty} \varepsilon^n \langle S_n\rangle_o} = 1 - \sum_{n=1}^{\infty} a_n \varepsilon^n \quad ,$$

then we'll have it. This equation is the same as

$$1 = \left(1 + \sum_{n=1}^{\infty} \varepsilon^n \langle S_n \rangle_o\right)\left(1 - \sum_{n=1}^{\infty} \varepsilon^m a_m\right)$$

or

$$1 = 1 + \sum_{n=1}^{\infty} \varepsilon^n \left[\langle S_n \rangle_o - a_n - \sum_{k=1}^{n-1} \langle S_k \rangle_o \, a_{n-k}\right] \quad .$$

Now if a power series in ε is to vanish identically (i.e., for all ε), the coefficient of each term must be zero. So, my problem is: choose the a_n so that

$$\langle S_n \rangle_o = a_n + \sum_{k=1}^{n-1} \langle S_k \rangle_o \, a_{n-k} \quad . \tag{24-48}$$

This is a discrete version of a Volterra integral equation, and is solved as follows. Define a sequence of operators Q_n,

$$Q_1 \equiv S_1 \tag{24-49}$$

$$Q_2 \equiv S_2 - S_1 \langle Q_1 \rangle_o \tag{24-50}$$

$$Q_n \equiv S_n - \sum_{k=1}^{n-1} S_k \langle Q_{n-k} \rangle_o \quad , \quad n > 1 \tag{24-51}$$

Taking the expectation value of (20-51) and comparing with (24-48), you see that the desired solution is just

$$a_n = \langle Q_n \rangle_o \tag{24-52}$$

Now, returning to (24-47) with this result, we have

$$\langle C \rangle = \left[\langle C \rangle_o + \sum_{k=1}^{\infty} \varepsilon^k \langle S_k C \rangle_o\right]\left[1 - \sum_{m=1}^{\infty} \varepsilon^m \langle Q_n \rangle_o\right] \quad . \tag{24-53}$$

In expanding this, note that the double sum can be written as

$$\sum_{k=1}^{\infty}\sum_{m=1}^{\infty}\varepsilon^{k+m}\langle S_k C\rangle_o\langle Q_n\rangle_o = \sum_{n=2}^{\infty}\varepsilon^n\sum_{k=1}^{n-1}\langle S_k C\rangle_o\langle Q_{n-k}\rangle_o \qquad (24\text{-}54)$$

and we might as well add the term with n=1, since it vanishes anyway, having no terms at all. So, we have

$$\langle C\rangle = \langle C\rangle_o + \sum_{n=1}^{\infty}\varepsilon^n\langle S_n C\rangle_o - \sum_{k=1}^{n-1}\langle S_k C\rangle_o\langle Q_{n-k}\rangle_o - \langle Q_n\rangle_o\langle C\rangle_o \qquad (24\text{-}55)$$

and, comparing with (24-51), we get a pleasant surprise; patience and virtue are rewarded at last with what we had no right to expect in such a problem; a neat and simple final result:

$$\langle C\rangle - \langle C\rangle_o = \sum_{n=1}^{\infty}\varepsilon^n\left[\langle Q_n C\rangle_o - \langle Q_n\rangle_o\langle C\rangle_o\right] \quad . \qquad (24\text{-}56)$$

The n'th order contribution to <C> is just the covariance, in the unperturbed ensemble, of Q_n with C. The first-order term in (24-56) has long been known; to the best of my knowledge, Steve Heims was the first person to see that it can be extended to all orders. In several years of living with this formula, and seeing what it can do for us, I have come to regard it as easily the most important general rule of statistical mechanics; almost every "useful" calculation in the field can be seen as a special case of it.

So, this is the general perturbation expansion that we'll use. Every calculation I do from now on will be a special case of the application of Heims' theorem (24-56). Now, the first order correction of course is always the most important one. The first order term has a symmetry property which follows from this cyclic property of the trace, Eq. (24-11); and let me just

show that to you. To first order, since $O_1 = S_1$, I have simply

$$\langle C\rangle - \langle C\rangle_o = \varepsilon\left[\langle S_1 C\rangle_o - \langle S_1\rangle_o\langle C\rangle_o\right] \tag{24-57}$$

but

$$S_1 \equiv \int_o^1 e^{-xA}\, B e^{xA}\, dx$$

so that

$$\langle S_1\rangle_o = \int_o^1 dx \langle e^{-xA}\, B e^{xA}\rangle$$

$$= \frac{\int_o^1 dx\ \mathrm{Tr}\left[e^{(1-x)A}\, B e^{xA}\right]}{\mathrm{Tr}\left(e^{A}\right)} \quad . \tag{24-58}$$

Now, as in (24-11), it is true generally that $\mathrm{Tr}(FG) = \mathrm{Tr}(GF)$; and so

$$\langle S_1\rangle_o = \frac{\int_o^1 dx\ \mathrm{Tr}\left[e^{xA}\, e^{(1-x)A}\, B\right]}{\mathrm{Tr}\left(e^{A}\right)} = \frac{\mathrm{Tr}\left(e^{A} B\right)}{\mathrm{Tr}\left(e^{A}\right)} = \langle B\rangle_o \quad , \tag{24-59}$$

so the first-order correction always reduces to

$$\langle C\rangle - \langle C\rangle_o = \varepsilon\left[\int_o^1 dx \langle e^{-xA}\, B e^{xA} C\rangle_o - \langle B\rangle_o\langle C\rangle_o\right] \quad . \tag{24-60}$$

[At this point, we can verify Eq. (24-28). Make the choices $A = -\lambda_1 F_1 - \ldots - \lambda_m F_m$, $\varepsilon B = -\delta\lambda_k F_k$. Then $Z(\lambda_1 \ldots \lambda_m) = \mathrm{Tr}\left(e^{A}\right)$ and from the definition of a derivative,

$$\frac{\partial \log Z}{\partial\, \lambda_k} = \frac{1}{Z} \lim_{\delta\lambda_k \to o} \frac{Z\left[\lambda_1 \ldots \lambda_k + \delta\lambda_k \ldots \lambda_m\right] - Z\left[\lambda_1 \ldots \lambda_k \ldots \lambda_m\right]}{\delta\lambda_k} \quad . \tag{24-61}$$

In the limit $\delta\lambda_k \to 0$, only the first-order term survives, and so

$$\frac{\partial \log Z}{\partial \lambda_k} = \frac{\mathrm{Tr}\left(e^A S_1\right)}{Z \delta \lambda_k} = \frac{\langle S_1 \rangle_o}{\delta \lambda_k} \quad . \tag{24-63}$$

But, using (24-59), you see that this is just (24-28)].

Now I want to show you a very important symmetry property; if I interchange B and C in the right-hand side of (24-60), I don't change it. The last term I have worked into a form where it is obvious. We still have to play with the first one a little bit. Again, let's write this as the ratio of two traces.

$$\int_o^1 dx \langle e^{-xA} B^{xA} C \rangle_o = \frac{\int_o^1 dx \, \mathrm{Tr}\left[e^{(1-x)A} B e^{xA} C\right]}{\mathrm{Tr}\left(e^A\right)} \tag{24-64}$$

This time I choose to interchange matrices as follows,

$$\int_o^1 dx \, \mathrm{Tr}\, e^{(1-x)A} B e^{xA} C = \int_o^1 dx \, \mathrm{Tr}\left[e^{xA} C e^{(1-x)A} B\right] \quad . \tag{24-65}$$

Now for any $f(x)$, we have

$$\int_o^1 f(x)dx = \int_o^1 f(1-x)dx \tag{24-66}$$

consequently we can write (24-65) as

$$\int_o^1 dx \, \mathrm{Tr}\left[e^{(1-x)A} C e^{xA} B\right] \quad , \tag{24-67}$$

and writing this back as an expectation

$$\int_o^1 dx \langle e^{-xA} B e^{xA} C \rangle_o = \int_o^1 dx \langle e^{-xA} C e^{xA} B \rangle_o \quad . \tag{24-68}$$

After all this, the only thing that has happened is that I've interchanged B and C.

Now this is a very important symmetry property. If I perturb my density matrix by adding in formation about B and I calculate what effect that makes on my prediction of C, it is the same as if I had perturbed my density matrix by putting in information about C and calculated what effect that makes on B. In the next Lecture, I'll show you a whole string of physical reciprocity laws that come out of (24-68).

Again, I'm leaving you on a note where we have an enormous amount of abstract stuff and you haven't seen the physical problem. In the next two Lectures, we'll make up for that.

SUMMARY OF BASIC RULES AND NOTATION
(Continued from inside front cover)

Continuous Distributions: If x is continuously variable, we denote the probability, given A, that it lies in the range (x, x+dx) by

$$(dx|A) = (x|A)\, dx$$

Thus the same bracket symbols (|) are used for probabilities and probability densities. This causes no confusion, since the distinction is determined by whether the quantity is discrete or continuous. Rule 1 and Bayes' theorem then take the same form as above, since dA and/or dB cancel out; and the summations above become integrations.

Prior Probabilities: The initial information available to the robot at the beginning of any problem is denoted by X. $(A|X)$ is then the prior probability of A. Applying Bayes' theorem to take account of new evidence E yields the posterior probability $(A|EX)$. In a posterior probability we sometimes leave off the X for brevity: $(A|E) \equiv (A|EX)$.

Prior probabilities are determined by Rule 4 when applicable; or more generally by the principle of maximum entropy (Lect. 10): choose the $p_i \equiv (A_i|X)$ so as to maximize $H = -\Sigma_i\, p_i \log p_i$ subject to constraints represented by X. In the continuous case this becomes: maximize $H = -\int p(x) \log [p(x)/m(x)]\, dx$, where the measure m(x) is determined by invariance under the group of transformations which convert the problem into an equivalent one, for consistency in sense (b) above (Lect. 12).

Decision Theory: (Lect. 13). Enumerate the possible decisions $D_1 \ldots D_k$ and introduce a function $L(D_i, \theta_j)$ representing the "loss" incurred by making decision D_i if θ_j is the true state of nature. Make that decision D_i which minimizes the expected loss $\langle L \rangle_i = \Sigma_j\, L(D_i, \theta_j)(\theta_j|EX)$ over the posterior distribution of θ_j.

Probability and Frequency: The above rules are shown to apply to general inductive inferences, whether or not any random experiment is involved. Many applications can be carried to completion without ever mentioning frequencies (Lectures 5,6,8,9,11,14,18).

If a problem does involve a random experiment, connections between probability and frequency will appear as mathematical consequences of the theory. Most random experiments are exchangeable sequences (Lect. 17); here the probability of an event is numerically equal to the estimate of frequency which minimizes the expected square of the error. Conversely, if an experiment has been repeated many times, the probability of any event at the next trial approaches its observed frequency (Lect. 16).

Probabilitees derived from maximum entropy subject to constraints are equal to the frequencies which can be realized in the greatest number of ways subject to the same constraints (Lect. 10). Probabilities derived by invariance under a transformation group are equal to the frequencies most likely to be produced in the sense that they require the least "skill" (Lect. 12).