PROBABILITY THEORY IN SCIENCE AND ENGINEERING

BY

PROFESSOR E. T. JAYNES

STANFORD UNIVERSITY

COLLOQUIUM LECTURES IN PURE AND APPLIED SCIENCE

NO. 4

FEBRUARY, 1958

FIELD RESEARCH LABORATORY

SOCONY MOBIL OIL COMPANY, INC.

DALLAS, TEXAS

Printed in U.S.A.

TABLE OF CONTENTS

LECTURE 1

It's a real pleasure to be here. I've heard jokes about Texas for so many years that I'm very glad of the chance to come here and see for myself just what the situation is, and I see that Texas is exactly like California, especially the weather.

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, so if you want a paraphrase, that's it.

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 specific 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. Professor Kac has emphasized several times, not only in his lectures here but in other places, that the question of how one introduces probability methods into physics at all is a matter that is very important, very largely unsolved, and even more largely neglected. In most of the problems in physics and engineering where you might think of using probability theory, your decision as to whether use of probability theory 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 application, we'll have to make up our minds about that. My main purpose in these talks is to show that, with a little different approach 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 mathematics a little bit.

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 about two-thirds theoretical physicist and one-third electrical engineer could get really worried about the foundations of probability theory. This is something which people like me are not supposed to be interested in. The things that I'm going to talk about here arose from my attempts, over a period of about ten years, to understand what statistical mechanics is all about and how it is related to communication theory. Ten years ago 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 various approaches to statistical mechanics and 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 that Professor Kac has talked about here, but that no two people have done this in the same way.

It was just during this year that Shannon's papers announcing the birth of information theory appeared. I discovered them accidentally in the Princeton library about the time I was thinking about statistical mechanics. I took these papers 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 then applied with great success to problems of engineering which apparently have no relation to thermodynamics. But all of the inequalities, which are usually associated with the second law of thermodynamics, turn out to be statements of the greatest practical usefulness in engineering problems. It seemed to me that there must be something pretty important that we could learn from this situation.

This feeling was shared by a large 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 Information Theory got 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 accumulated hindsight to date, 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.

It perhaps takes a moment of thought to see that the mere fact that a mathematical expression like

$$\sum p_i \log p_i$$

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

$$e^x \quad , \quad \cos\vartheta \quad , \quad J_0(z)$$

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 Einstein's that I like very much.* Science is

* A. Einstein, "The Meaning of Relativity", (Princeton Univ. Press, 1946); pp 56-57.

fully justified in finding some relation between 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 accept this equality as a general principle of physics until he could see inertial mass and gravitational mass as 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.

This is a lesson which could be used with profit in all parts of science. We won't commit any serious error of methodology if we try to follow Einstein's example 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. The 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.

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, 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 introduction to his book. Gibbs simply constructed models in which he assigned certain probabilities for certain situations, and he never made any attempt to say why he chose those particular probabilities. In the introduction to his book 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 5/6 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 than 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 distribution of his canonical ensemble. 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 best 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 statistical mechanics depends on the validity of my arguments. But I know in my bones that this theory is right independently of any such arguments, 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, (I won't quote it exactly) "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 distribution 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 about 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 properties, 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 and included 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.

Probably the best example of all is the tremendous complication which spectroscopy got into just before the discovery of the Schrödinger equation. All the rules of thumb that were developed in predicting what spectrum lines would occur and which ones would not, and 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.

How has the Gibbs model fared? We've had it for over 50 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 equilibrium thermodynamics. 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 in the sense of being able to prove it mathematically. 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. In particular, they have done very little to extend the Gibbs model to more general situations, as real advances always do.

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.

I don't want to go into any 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 distribution, 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.

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 this 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 distribution. 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.

At this point 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.

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 and he looks across the street and sees a jewelry store with a broken window, and there's a gentleman wearing a mask, carrying a bag full of watches and diamond rings, crawling out through the broken window. The policeman doesn't hesitate at all in deciding this gentleman is dishonest. But, of course, this was not a logical deduction from what he saw. 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 against the window and had broken it, 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

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

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.

Now the reasoning the policeman went through in our example was not even of this type. It's 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 learn to ignore such trivial things. This shows that in our reasoning we depend very much on past experience 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 Polya has written three books* on plausible reasoning, pointing out all sorts of very interesting examples, showing that there are fairly definite rules by which we do plausible reasoning. 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 very fast 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. 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

---

* G. Polya, "How to Solve It", (Princeton Univ. Press, 1945; second paperbound edition by Doubleday Anchor Books, 1957); "Mathematics and Plausible Reasoning", Volumes I and II (Princeton Univ. Press, 1954).

idea of inventing a "symbolic logic" for plausible reasoning is extremely old. Leibnitz speculated on such a "Characteristica Universalis" almost 200 years before Boole's "Laws of Thought" appeared. Boole's book came out in 1854, I believe, and gave us a symbolic logic for deductive reasoning.

The theory of probability was originally regarded as a symbolic logic for plausible reasoning, and when Laplace's "Theorie Analytique" came out in 1812, this was widely regarded as the long awaited "calculus of inductive reasoning", fully developed. Throughout almost all of 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 von Mises,* 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, Jeffreys** is equally vigorous in denouncing the view of von Mises, 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. Physics and engineering offer many problems

---

* R. von Mises, "Positivism, A Study in Human Understanding", (G. Braziller, Inc., N. Y., 1956); Chap. 14

** H. Jeffreys, "Theory of Probability", (Oxford, 1939)

where use of probability theory is entirely legitimate on one interpretation, and entirely meaningless 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, which are only "Scheinprobleme", can still depend on which viewpoint you adopt.

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. 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.

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 this endeavor, we will feel that progress has been made if we are able to construct idealized mathematical models which reproduce some of those good features; this is the methodology of Gibbs.

In order to direct attention to constructive things and away from controversial things which we can't answer at present, we will 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 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. This is a definite engineering problem. What's the best way of designing his brain case?

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 it. This means that there will be some kind of association between plausibilities and real numbers. We'll just assume 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 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. These assumptions, you see, are practically forced on us by the requirement that the robot's brain must operate by the carrying out of some definite physical process.

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 we reason. For example, if he gets new information which increases the plausibility of proposition A but does not affect the plausibility of proposition B, this of course will always produce an increase, and not a decrease, in the plausibility that both A and B are true. And it must 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. 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 that he always comes to the same conclusion regardless of the order in which we give him the evidence.

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

1. association of plausibilities with real numbers
2. qualitative correspondence with human common sense
3. consistency

These three requirements, I claim, uniquely determine the rules according to which this robot must reason. There is only one set of mathematical rules which satisfies all those conditions. Next, we will turn to the job of working out these rules.

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 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 being humane, 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.

A Model For Inductive Reasoning

We have first to introduce some more notations of the usual symbolic logic. By the product

$$(AB),$$

we mean the proposition "Both A and B are true". The expression

$$(A+B)$$

means "At least one of the propositions A, B is true". 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. Thus, for example,

$$(A|BC)$$

I'll read this as A given BC. It stands for the plausibility that A is true, given that B and C are both true. Or,

$$(A+B|CD)$$

stands for 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

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

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

We now 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 $(AB|C)$. Now in order for (AB) to be a true proposition, it is certainly necessary that B be true; thus, the plausibility $(B|C)$ must be involved. In addition, if B is true, it is necessary that A should be true; so $(A|BC)$ is 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 $(B|A)$ would not be necessary; 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 that 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 $(A|BC)$:

$$(AB|C) = F[(B|C),(A|BC)] \qquad (1)$$

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 quite plausible given C, B might be quite plausible given C, but A and B together might be impossible. For example, if I'm told that Mr. Jones lives in Dallas, it might be quite plausible that his left eye is blue and it might be quite plausible that his right eye is brown, but it's very implausible that both of those are true. We would have no way of taking such influences into account if we tried to use a formula of this kind. Our robot could not reason the way human beings do, even qualitatively, with that kind of function.

You might try further a relation of this 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 the way we reason.

On the other hand, I don't think you'll be able to produce any situation where the equation

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

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.

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

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

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

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

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:

$$(ABC|D) = F\left[(C|D), (AB|CD)\right]$$

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

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,

$$F\left[F(x,y),z\right] = F\left[x, F(y,z)\right] . \qquad (2)$$

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 (1). And you can show by induction that if the functional Equation (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 Crelles Journal. This was the first issue of Crelle's Journal, and I think this was the first paper that Abel wrote. He considered this functional equation as merely an amusing exercise and found the general solution of it. The solution has been rediscovered probably dozens of times since 1826. In particular, there is a paper by J. Aczel*. There is another article where this functional equation is considered in some detail by R. T. Cox**, who first suggested this approach to probability theory. Let me just quote 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 $f(z)$ such that $f(z) = f(x) + f(y)$. 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 $x \circ y$ must have the following properties:

---

*Jean Aczel, "Sur les Opérations définies pour nombres réels", Bull. Soc. Math. Franc. 76, 59-64 (1948)

**R. T. Cox, Am. Jour. Physics 14, 1-13, (1946)

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 $F(x,y) = f^{-1}[f(x) + f(y)]$, 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 $p(x) = \exp[f(x)]$ is still an arbitrary continuous monotonic function.

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

$$p(AB|C) = p(A|BC)\; p(B|C). \qquad (3)$$

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 the plausibility that both A and B are 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) 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 $p(\chi)$ has to have the property that <u>certainty must always be represented by $p = 1$.</u>

Let's notice what happens if there is a degree of plausibility for which p becomes either zero or infinite. Say, for instance, that $p(A|BC)$ becomes either zero or infinity for some particular case. Then equation (3) says that the plausibility of B becomes irrelevant to the plausibility of AB. This could be so only if A was impossible, given B and C. So we conclude that $p(\chi)$ must have another property. <u>It cannot become zero or infinite for any degree of plausibility except impossibility.</u>

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

$$(AB|C) = (A|C) \quad .$$

Also, 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, our equation (3) reduces to

$$p(A|C) = p(B|C)\, p(A|C) \qquad (4)$$

and again this equation would have to hold no matter what plausibility B might have. Well, there are three possible values of $p(A|C)$ that might satisfy this condition. It could be zero, plus infinity, or minus infinity. But the choice minus infinity we have to exclude because p is a continuous monotonic function which has to get up to plus one for certainty. If it got down to minus infinity for impossibility, it would have to cross zero somewhere between, and we've just seen that p cannot become zero for any plausibility except impossibility if this rule is to represent qualitatively the way we think. The choice minus infinity can be ruled out also by other arguments (for example, see what happens in (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.

Now there are still other conditions of consistency which these rules would have to 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". And because of the fact that these are propositions of the type which must be either true or false, then we can conclude that the product aA is always false, and the 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:

$$(a|B) = S(A|B) \quad . \qquad (5)$$

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

$$(A|B) = S(a|B) \quad . \qquad (6)$$

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

$$S\{S(x)\} = x \quad . \tag{7}$$

Now this alone is not enough to tell us much about this function. So now I'd like to give you another argument. There's another condition which would have to satisfy in order to represent a consistent way of reasoning and for this we already have one rule of calculation worked out.

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

We'll call this Rule 1 from now on. We already know something about this function $p(x)$ and we can use that in the following argument, and just to save writing I will use this notation. We will understand that:

$$[A|B] \equiv p(A|B)$$

Then our Rule 1 can be written in this form:

$$[AB|C] = [B|C]\,[A|BC] \quad . \tag{9}$$

Now this argument that I gave on S of the plausibility numbers of course would apply equally well to the numbers p. In other words, I could put square brackets in equations (5), (6), (7), and everything I said would still hold. There's the arbitrariness of a monotonic function in all this. So let's suppose that S has been defined that way. We can make this step,

$$[AB|C] = S[a|BC][B|C]$$

$$= [B|C]\, S\left\{\frac{[aB|C]}{[B|C]}\right\} \quad .$$

through two applications of Rule 1. In the second equation, I've just written $p(a|BC)$ in a different way, which we get by application of Rule 1. 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 it. Therefore, although it doesn't look like it at all, this final expression must also be symmetric in A and B. In other words,

$$[A|C]\, S\left\{\frac{[bA|C]}{[A|C]}\right\} = [B|C]\, S\left\{\frac{[aB|C]}{[B|C]}\right\} . \qquad (10)$$

Now these two expressions would have to be equal to each other no matter what propositions A, B, and C are. In particular, they would have to be equal when the denial of B is the same as proposition "both A and D" are true; that is, when

$$b = AD ,$$

or

$$B = a + d .$$

But in that particular case, equation (10) simplifies. If B has this meaning, then what is $(bA|C)$? Well b is a 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. Therefore,

$$[bA|C] = [b|C] = S[B|C] .$$

Likewise, $aB = a(a+d) = a + ad = a$

$$[aB|C] = [a|C] = S[A|C] .$$

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

$$x S\left(\frac{S(y)}{x}\right) = y S\left(\frac{S(x)}{y}\right) . \qquad (11)$$

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*, which solves this problem. He shows that the only twice differentiable function which satisfies all of our conditions is

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

This means that our reciprocal relation between the proposition and its denial would then have to take the form

$$[a|B]^m + [A|B]^m = 1$$

or, dropping the square bracket notation,

$$p^m(a|B) + p^m(A|B) = 1 . \qquad (12)$$

---

* R. T. Cox, Am. J. Phys. 14, 1 (1946).

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 possible choice that $p = 0$ is going to represent impossibility. In that case, we'll have to choose m as a positive number, 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 $p_1(\chi)$ satisfies these conditions, then $p_2(\chi) \equiv (p_1(\chi))^m$ 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 $p(\chi)$ was an arbitrary monotonic function. This means that if I choose to write equation (12) in the form

$$p(a|B) + p(A|B) = 1 \tag{13}$$

this is just as general.

Now 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. If you have one function which satisfies that condition, then that function raised to the mth power also satisfies that condition. 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

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

is a function which represents impossibility by zero, and has all the properties that we needed. So regardless of which choice I make by which 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:

$$0 \leq p \leq 1 \quad .$$

(for impossibility) (for certainty)

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

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

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. We could take the equations we have as our fundamental equations, but I prefer to take another, which we deduce from these, as our second fundamental equation. 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. Now this follows from the rules we already have. We just apply (13) and Rule 1 over and over again:

$$\begin{aligned} p(A+B|C) &= 1 - p(ab|C) \\ &= 1 - p(a|bC)\,p(b|C) \end{aligned}$$

$$= 1 - [1 - p(A|bC)]\, p(b|C)$$

$$= p(B|C) + p(Ab|C)$$

$$= p(B|C) + p(b|AC)\, p(A|C)$$

$$= p(B|C) + p(A|C)\, [1 - p(B|AC)].$$

Finally, we get

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

At long last we come out with the above form. And it's this result that I will take as our Rule 2.

LECTURE 2

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

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

Rule 2: $$p(A+B|C) = p(A|C) + p(B|C) - p(AB|C).$$

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 this and it wouldn't have affected anything. So we have the liberty of writing it with A and B interchanged, as shown.

Arbitrariness and Prior Information

We've found so far the most general consistent rules by which our robot can do plausible reasoning, granted that he must associate plausibilities with real numbers in some way so that his brain can operate by some definite physical process. The most general rules can be written in the form of very familiar looking equations, in which p is an arbitrary monotonic function of these numbers which we call plausibility (and of

course we have not said anything at all how we are going to define these numbers). It might appear at this stage as if we have found an infinite number of different possible consistent rules by which our robot can do plausible reasoning: corresponding to every different choice of a monotonic function there'd be a different set of rules. This is not so, however, because no matter which function we chose, the behavior of the robot would be exactly the same. The information we give the robot, it turns out, will determine the numerical values of p , not the numerical values of the plausibilities that we started with. Let's see that in the simplest case. Suppose we have n different propositions, A_1 , A_2 ,.... A_n . These propositions are to be mutually exclusive. Two of them could not be true at the same time, so we could write an equation like this:

$$p(A_i A_j|B) = p(A_i|B)\,\delta_{ij}$$

Now suppose one of these propositions must be true on data B , but data B gives the robot no reason to prefer any one to any other. They're equally likely, to him. In that case, our last term of Rule 2 drops out and the sum of all these functions must certainly be equal to one:

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

If the propositions are all equally likely to the robot, they must all be represented by the same plausibility, and therefore the only possibility is that

$$p(A_i|B) = \frac{1}{n} \quad . \tag{15}$$

You see no matter what function $p(\chi)$ we had chosen, there would be no way of getting around this result, which is called the "Principle of Insufficient Reason" in the literature. The information we gave the robot, the statement that these propositions were mutually exclusive and exhaustive, determines the numerical values of the $p's$. The robot's reasoning process can be carried out entirely in terms of manipulation of the numbers $p(\chi)$, as our equations show, and the robot's final conclusions could be stated equally well in terms of p instead of χ . This means that the quantities χ that we started with have faded completely out of the picture. They correspond to different possible ways that you could design the circuits of his brain; but no matter what function you chose, the observable behavior of the robot would be exactly the same. So rather than saying that p is an arbitrary monotonic function of the plausibility χ , it looks like maybe it's more to the point to say that the plausibility χ is an arbitrary monotonic function of p . As soon as we realize this, we see that there is actually only one consistent set of rules by which this robot can do plausible reasoning.

From now on, instead of writing

$$p(A|B) \quad ,$$

I'm simply going to leave off the $\wp$ and write it

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

You can interpret this two ways. You can say I'm changing my notation, since it's always the function $\wp$ that we're concerned with, I'll simply understand that it's always that function that is meant. Or you can say that I've now adopted the convention that

$$\wp(x) \equiv x$$

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

Rule 1
$$(AB|C) = (A|BC)(B|C) = (B|AC)(A|C) \tag{16}$$

Rule 2
$$(A+B|C) = (A|C) + (B|C) - (AB|C) \tag{17}$$

and from now on we'll call these quantities 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. 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 still harmless. If it's irrelevant, it will cancel out mathematically. Any probabilities which are conditional on X alone we will call a priori probabilities or prior probabilities. If there is any additional evidence in addition to X, which the robot is now reasoning on, we will generally leave off the X. We'll understand that even when we don't write X explicitly, it's always built into all expressions:

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

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

$$(A|X).$$

Because of some strange things that have been thought about a priori probabilities in the past, we have to carefully point out that it would be a big mistake to think of X as being some sort of hidden major premise, that represents 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.

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, $(B|C)$:

$$(A|BC) = (A|C)\frac{(B|AC)}{(B|C)} . \qquad (18)$$

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 essentially 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 way to do this is to calculate the probability of A , based 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 random 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, or making some kind of prediction as to what is likely to happen next, and so on. All of these represent plausible reasoning on the basis of new evidence; the evidence

of the random experiment. Therefore, to our robot it's perfectly obvious that all examples of parameter estimation and hypothesis testing must be special cases 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 processes must come from Bayes' theorem is just as much a necessity of thought as the validity of a syllogism is to us.

Although this conclusion about Bayes' theorem is obvious to our robot, it has not been at all obvious to human statisticians. They largely regard Bayes' theorem as not having any logical basis except in the case where every probability in it can be given a frequency interpretation. In that case, Bayes' theorem can be interpreted as selecting out of an original population of events some sub-population 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.

Since this is perhaps the 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 about the effect of new information. If this procedure is fully consistent and in full qualitative agreement with common sense, then it is

necessarily exactly 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 usually 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, of course, this is a definite issue of fact, not a vague matter of personal opinion. 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.

Maximum Likelihood

The first example in 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.*

*R. A. Fisher, "Statistical Methods For Research Workers" (11th Edition, Hafner, N. Y., 1950).

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:

$$V = S + N$$

Our prior knowledge about the nature of the noise can be described by some probability distribution.

$$W(N)\,dN = (N|X)\,dN\,.$$

I've extended the notation a little bit here. I'm now indicating that the same bracket symbols will be used for probability densities as for probabilities. The distinction is always determined by whether the variables are continuous or discrete, so I don't see any need to invent a new notation for it. Now if we knew that the signal had a certain value S, then the probability of observing the voltage V would, of course, be the probability that the noise would have made up just the difference:

$$(V|SX)\,dV = W(V-S)\,dV. \qquad (19)$$

I'm going to be quite sloppy about putting in differentials of this sort. They would always cancel out of equations anyway. Now let's write this still another way.

$$(V|SX) = L(V,S).$$

If we know the signal then this is the probability that we would observe the given voltage. Now in the 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 function L an absolute maximum for the observed value of V :

$$\frac{\partial L}{\partial S} = 0 \quad .$$

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

$$(S|VX) = (S|X)\frac{(V|SX)}{(V|X)} \quad , \tag{20}$$

$$= (S|X)\ \frac{L(V,S)}{(V|X)} \quad .$$

So if we ask the robot what is the most probable value of the signal, 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 prefer one signal magnitude over another, then the robot's estimate would be exactly equal to 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 obtain by throwing away all the prior information we had about the signal, and basing our estimate only 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 suddenly 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 previous medical history?" And

suppose the doctor said, "Why no, I must not look at your previous 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 a significant amount of prior information, his common sense will always tell him 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 theory 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. Mathematically, it is nothing but Bayes' theorem with uniform prior probability.

Sequential Testing

The second example of statements made about Bayes's theorem

in the literature has been provided by Feller. On page 85 of his book* we have the following quotation. He says, "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 when he invented it, Wald's sequential testing procedure*** has been generally considered the optimum one available, optimum according to several different criteria.

---

*W. Feller, "An Introduction to Probability Theory and Its Applications," (Wiley, 1950).

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

***A. Wald, "Sequential Analysis," (Wiley, 1947).

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, because that would run the production cost up and our competitor would probably run us out of business. 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 that most statisticians agree that Wald's method is the optimum one in this sense of requiring on the average fewer tests than any other. 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 bit later. Just for variety, instead of describing Wald's method first, let's go first into the way the robot would handle this problem. To do this, let's manipulate Bayes' theorem a little bit in a manner suggested by I. J. Good.*

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. J. Good, "Probability and the Weighing of Evidence," (C. Griffin & Co., Ltd., London, 1950).

I'll use E to stand for new evidence.

$$(A|EX) = (A|X)\frac{(E|AX)}{(E|X)} .$$

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

$$(a|EX) = (a|X)\frac{(E|aX)}{(E|X)} ,$$

and we can take the ratio of the two equations:

$$\frac{(A|EX)}{(a|EX)} = \frac{(A|X)(E|AX)}{(a|X)(E|aX)} .$$

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 divided by the probability that it's false, has a technical name. We call it the "odds" on proposition A. So if I write the "odds of A given E and X," as the symbol

$$\mathbf{O}(A|EX) = \frac{(A|EX)}{(a|EX)}$$

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

$$O(A|EX) = O(A|X)\,\frac{(E|AX)}{(E|aX)}. \qquad (21)$$

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 case. And so I'm going to define a new function which I'll call the evidence for A given E :

$$e(A|EX) \equiv 10 \log_{10} O(A|EX).$$

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:

$$e(A|E) = e(A|X) + 10 \log_{10} \frac{(E|A)}{(E|a)}. \qquad (22)$$

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

$$E = E_1 E_2 E_3 \cdots\cdots$$

In that case, we could expand this a little more:

$$e(A|E) = e(A|X) + 10 \log_{10} \frac{(E_1|A)}{(E_1|a)} + 10 \log_{10} \frac{(E_2|E_1 A)}{(E_2|E_1 a)} + \cdots \qquad (23)$$

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:

$$e(A|E) = e(A|X) + 10 \sum_i \log_{10} \frac{(E_i|A)}{(E_i|a)}, \qquad (24)$$

where the sum is over all 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 draw 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

Table 1

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 does 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 1.

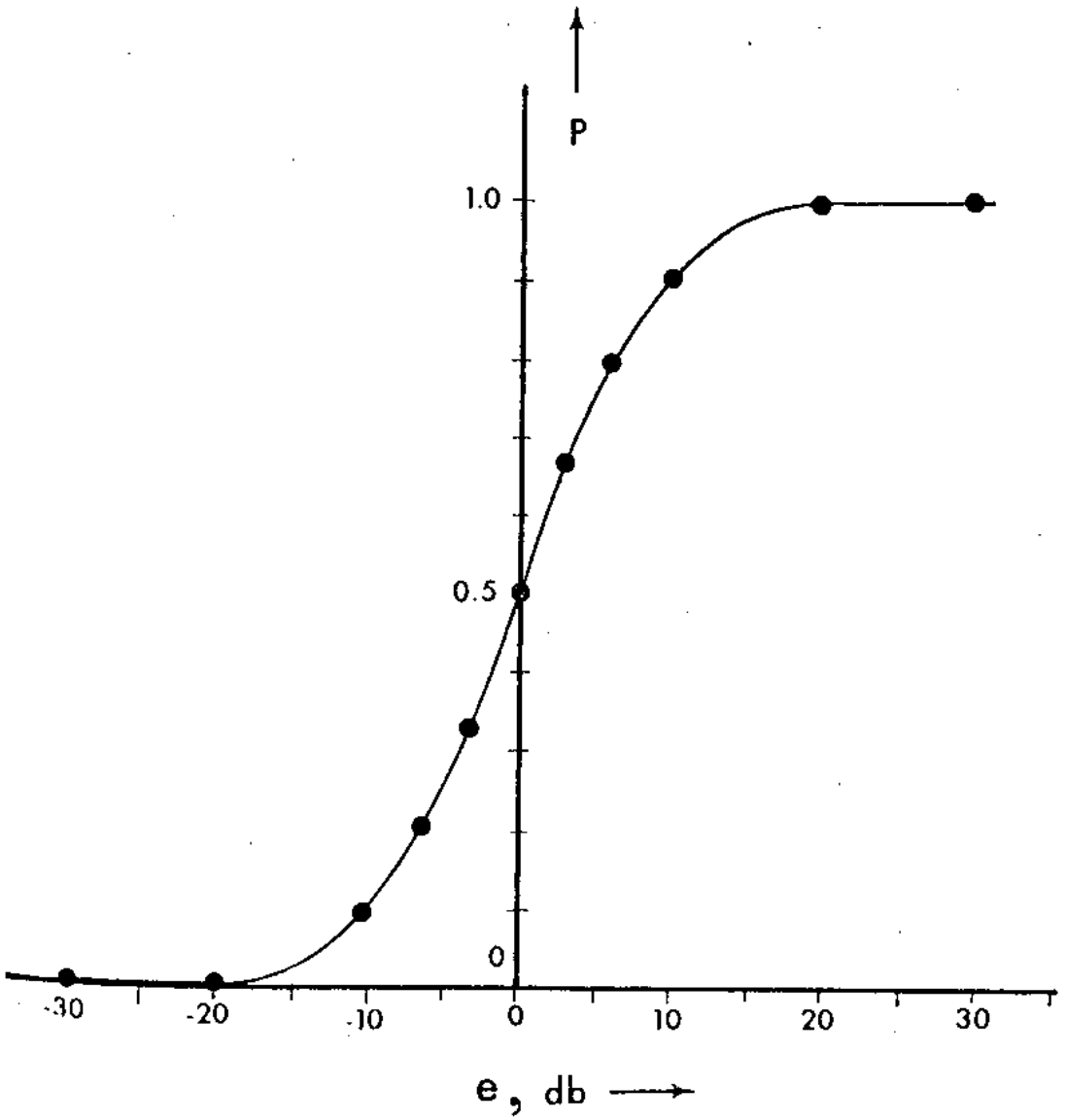


Figure 1.

The graph is symmetric about the center. This is to give just some slight feel for the way these things go.

Now let's take 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, one the average, one in six defective. The eleventh machine is even worse; it makes one in three defective. The output of each machine is going to be collected in a 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.

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, on the basis of 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 insufficient reason,

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

$$= -10 \text{ db.}$$

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, insufficient reason, by which we can convert raw information into numerical values of probability. I mention this here only because of 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

$$10 \log_{10} \frac{(\text{bad} \mid A)}{(\text{bad} \mid a)}$$

where $(\text{bad} \mid A)$ represents the probability of getting a bad diode, given A. Again "Insufficient Reason" tells us what these separate probabilities are. If we have the box in which one in three are bad, the probability will be 1/3, and if we had the box of good ones the probability

would be 1/6. We assume that the number in the box is very large compared to the number tested. So, every bad diode we find gives us

$$10 \log_{10} \frac{1/3}{1/6} = 10 \log_{10} 2 = +3 \text{ db}$$

of evidence for the proposition that we had a bad batch. Now suppose we find a good diode. We'll get evidence for A of

$$10 \log_{10} \frac{(\text{good}|A)}{(\text{good}|a)} = 10 \log_{10} \frac{2/3}{5/6} = -0.96 \text{ db} ,$$

but let's call it - 1 db. 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

$$e(A|E) = -10 + 3N_b - N_g .$$

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

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

or, from Figure 1; the probability of a bad batch is brought up to

$$(A|E) \simeq 0.4 \quad .$$

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

$$N_b \geq 5 + \frac{N}{4}$$

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 that if the probability of proposition A reaches a

certain level then 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 might be very much more serious than making the other kind of error. 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 progress to the right. This is shown in Figure 2.

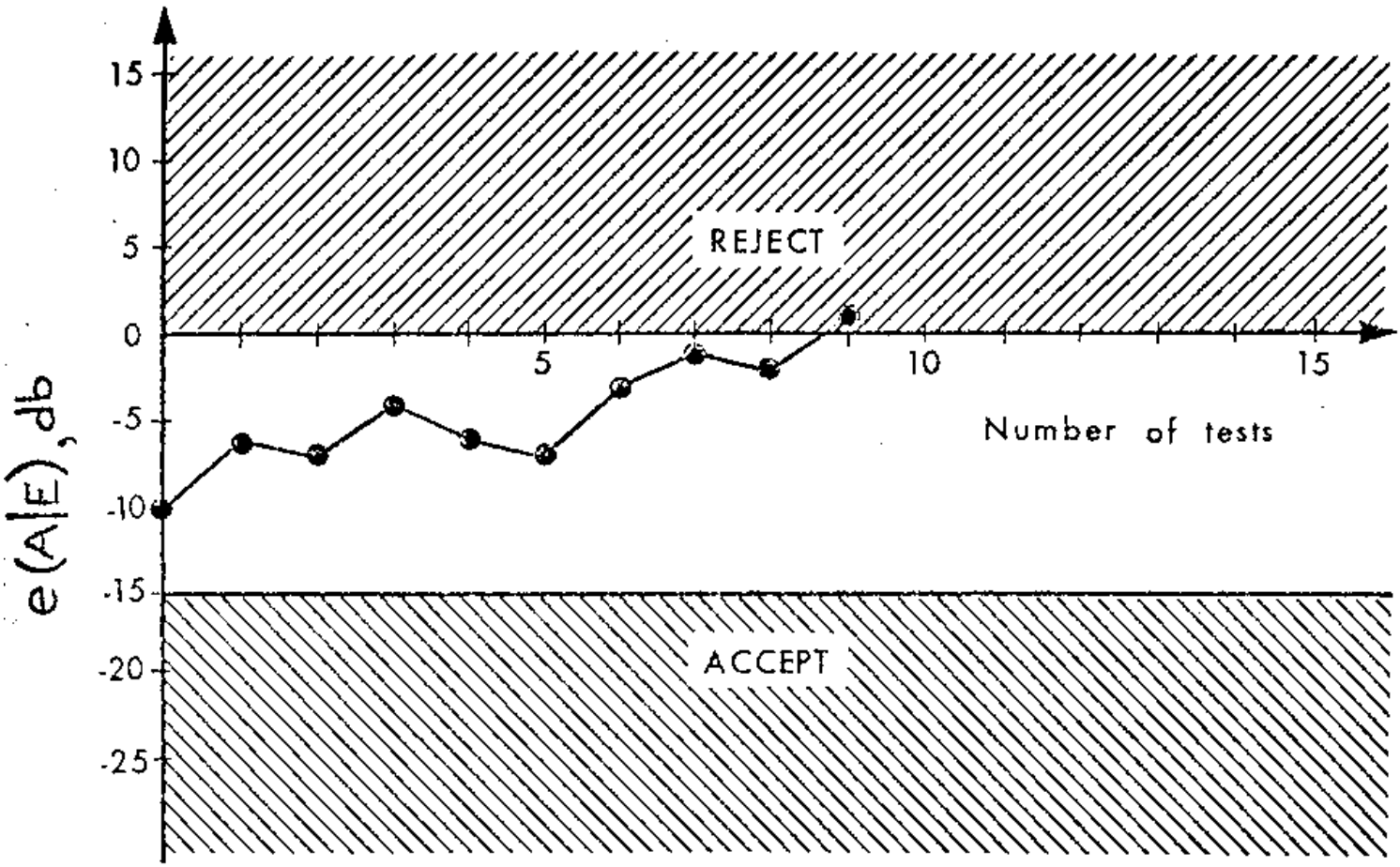


Figure 2.

Now, how does Wald do this? Wald* does not mention Bayes' theorem. But what he actually does is exactly the same, with the one characteristic difference which we find in all these comparisons. He always starts out by throwing away his prior evidence. His graphs always start out at 0 db.

Wald's probability ratio test involves the calculation of exactly the last term of Equation (24). This is the thing which he

* A. Wald, "Sequential Analysis", (Wiley, 1947).

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. Anybody who is familiar with Wald's powers as a mathematician can well imagine how much effort must lie behind that simple little remark in his book. But how does this problem of proof 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 it is the optimum way of proceeding by any sensible criterion of "optimum." Because to our robot, when you have calculated the probability of proposition A on the basis of all the available evidence, then you have got everything about A that is to be had from the evidence. Obviously, you can't do better than this, and there is nothing more to be said.

Does anyone incur any serious error by starting out with zero db? You might think at first that this is bad in the sense that it is inconsistent if we do have prior information and in principle this is right. 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 limits. There is nothing in probability theory which which tells us where to put the limits, but there is something which tells us which prior probability to assign. So we cannot remove all the arbitrariness, but we can remove some of it, by taking into account prior probability. In practice, the statistician would use his common sense to move the limits up and down and thus take account of his prior knowledge, without ever having to admit that there is any such thing as a "prior probability."

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 100 years ago. But so was everything else in science. The simple fact is that the most powerful known method of quality-control testing is nothing but an application of Bayes' theorem with uniform prior probability.

Multiple Hypothesis Testing

Let's suppose something very remarkable happens in this sequential test. Suppose we tested sixty diodes and every one turned out to be bad. According to our equations, that would give us 180 db of evidence for the proposition that we had the bad batch. $e(A|E)$ would end up at + 170 db, which is a probability which differs from 1 by one part in 10^{17}. Now our common sense rejects this conclusion. If you test 60 of them and you find that all 60 are bad, you are not willing to believe that you have a batch in which only 1 of 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 the 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 said we had either 1/3 bad or 1/6 bad. Those are entirely different propositions.

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 way 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. Let's say proposition A means as before that we have a box with 1/3 defective, 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% 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 the 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:

$$\begin{aligned} (A|X) &= \tfrac{1}{11}\ (1 - 10^{-6}) \\ (B|X) &= \tfrac{10}{11}\ (1 - 10^{-6}) \\ (D|X) &= 10^{-6} \end{aligned} \tag{25}$$

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% 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

Let's be explicit and say that 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 logarithm of this probability ratio:

$$e(D|E) = e(D|X) + 10 \log_{10} \frac{(E|DX)}{(E|dX)} \quad . \quad (26)$$

(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?

$$(E|DX) = \left(\frac{99}{100}\right)^m$$

is the probability that the first m are all bad, given that 99% of the machine's output is bad. This is obtained by application of Rule 2 and then Rule 1. We also need the probability that the first m would all be bad given that we <u>had</u> to have one of the first two hypotheses. In this case, signified by proposition d, A and B are exclusive propositions and one of them must be true. The negative term in Rule 2 vanishes for this case, then, and

$$(E|dX) = (EA|dX) + (EB|dX) .$$

Now we can expand these by our Rule 1:

$$(E|dX) = (E|Ad)(A|d) + (E|Bd)(B|d).$$

The probability $(E|Ad)$ may be abbreviated, for the statement that A is true implies that D is false in our problem the way we've set it up. And so this d is really irrelevant in $(E|Ad)$. Likewise the statement that B is true also implies that D must be false, and so

$$(E|dX) = (E|A)(A|d) + (E|B)(B|d).$$

If A is true, 1/3 of them are bad. The probability of experience E would be $(1/3)^m$, if the total number in the box is very large compared to m. And the probability of A given D false is the same as in our first problem, 1/11. If B is true, the probability of E on that basis would be $(1/6)^m$, and the probability that B would be true given D false is again the same as in our first time through this problem. So we have

$$(E|dX) = \left(\frac{1}{3}\right)^m \cdot \frac{1}{11} + \left(\frac{1}{6}\right)^m \cdot \frac{10}{11}.$$

Now if we put all these things together, we come out with this expression for the evidence for proposition D :

$$e(D|E) = -60 + 10 \log_{10}\left[\frac{\left(\frac{99}{100}\right)^m}{\frac{1}{11}\left(\frac{1}{3}\right)^m + \frac{10}{11}\left(\frac{1}{6}\right)^m}\right] . \qquad (27)$$

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:

$$e(D|E) \cong -49.6 + 4.73\, m \quad \text{for} \quad m > 5 . \qquad (28)$$

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

$$e(D|E) \cong -59.6 + 7.73\, m \quad \text{for} \quad m < 3 . \qquad (29)$$

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 each gives us about 7 3/4 db of evidence for the proposition, so it 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 3db 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:

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

$$\cong \begin{cases} -10 + 3m & \text{for } m < 7 \\ 49.6 - 4.73m & \text{for } m > 8 \end{cases} \qquad (30)$$

$$e(B|E) = +10 + 10 \log_{10} \left[\frac{\left(\frac{1}{6}\right)^m}{\left(\frac{1}{3}\right)^m + 11 \times 10^{-6} \left(\frac{99}{100}\right)^m} \right] \tag{31}$$

$$\cong \begin{cases} 10 - 3m & \text{for } m < 10 \\ 59.6 - 7.33\, m & \text{for } m > 11 \end{cases}.$$

These results are summarized in Figure 3. 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 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.

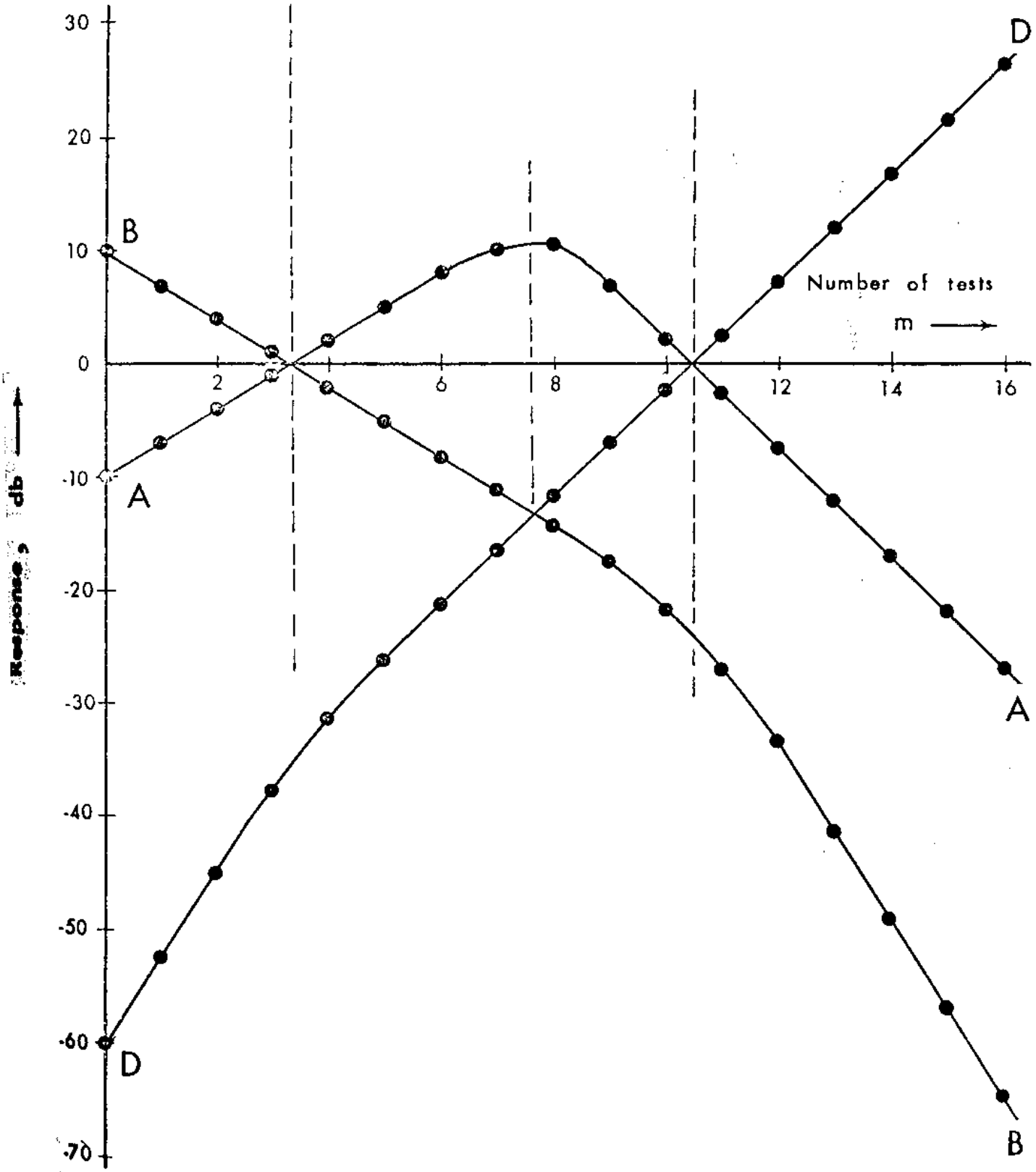

30
20
10
0
-10
-20
-30
-40
-50
-60
-70
2
6
8
12
14
16
B
A
D
D
A
B
Number of tests
m
Response, db

Figure 3.

Now what is going on here? When D has reached the same plausibility as B, that makes 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. 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 several simultaneous repetitions of testing one hypothesis against a single alternative. So, seeing this, you can construct curves of the sort shown in Figure 3 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.

The rule for drawing these curves is exactly the one that electrical engineers use for drawing frequency response curves. They find that it is convenient to plot response in db against the logarithm

of frequency. In certain frequency regions, you may have a flat response, then one of the RC networks, perhaps in your amplifier, starts to cut off. Then you know that the response will drop off, say 6db per octave in a certain frequency range. The place at which this drop starts is determined by equations of the form $\omega RC = 1$. Maybe another network comes in, and after that the response will go down 12 db per octave, and so on. The response curve could be drawn as a number of straight line segments very easily, as shown in Figure 4.

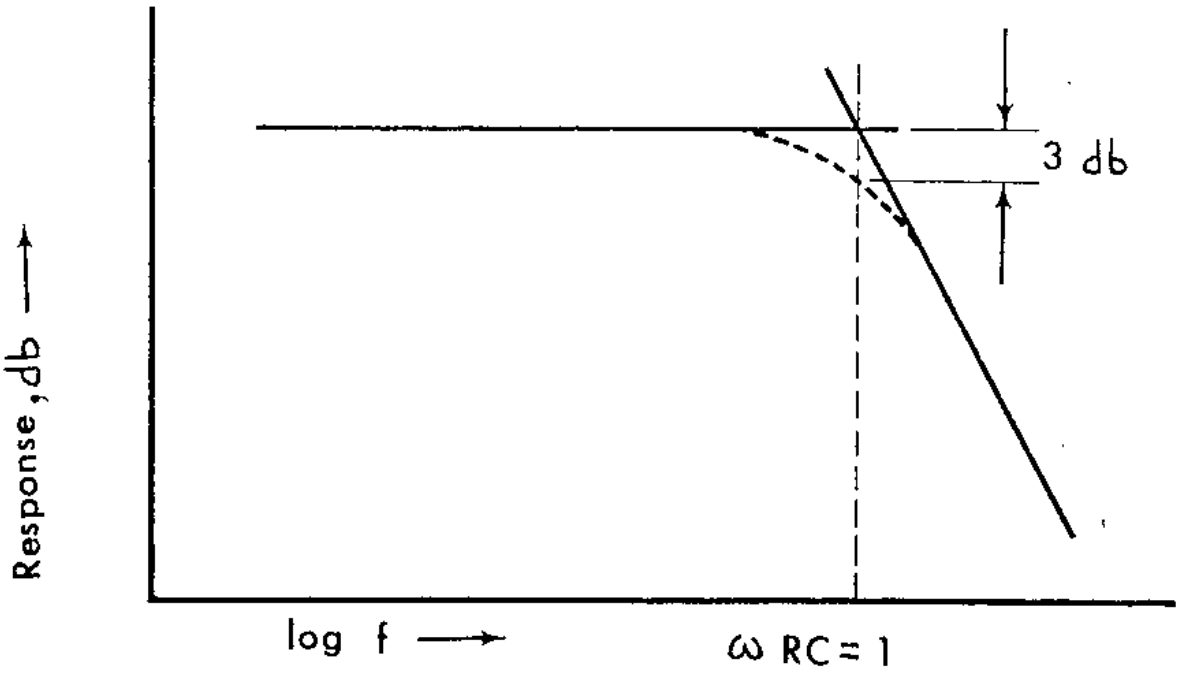


Figure 4.

The rule to round the curve off correctly is that immediately under the intersection of the two straight lines we're down 3 db . This rule is exact both in our case and in the filter case.

Figure 3 shows an interesting thing. Suppose we had decided to stop the test and accept hypothesis A if the evidence reached plus 10db. You see, it would reach plus 10db after about six trials. If we stopped the testing at that point, then of course we would never see the 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 defective." 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. 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 difficutly 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.* 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 value 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

---

* W. A. Shewhart, "Economic Control of Quality of Manufactured Products," (van Nostrand, N. Y., 1931)

of the sample are the only properties of the sample he needs in 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 exactly the same problem as that of detecting a signal in noise, which we will study later on. Also, it is exactly equivalent to the problem of deciding from a set of astronomical observations whether there is some unknown systematic effect, or whether discrepancies should be blamed on errors of observation. Laplace was applying this theory 150 years ago in just that way - 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.

Unfortunately, the general field of quality control is still not very highly developed or unified, and one reason for this is reluctance to use Bayes' theorem the way Laplace did. But considerable progress is being made currently. After fifty years of rejecting Laplace's whole conception of probability theory, statisticians have recently rediscovered some of his methods. They now call it "decision theory," and we will consider it later.

LECTURE 3

Queer Uses for Bayes' Theorem

I. J. Good* has shown how we can use Bayes' theorem backwards to measure our own strengths of belief about propositons. For example, how strongly do you believe in 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 was very worried 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 before you were very worried and seriously considering the possibility of extrasensory perception?

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

* I. J. Good, "Probability and the Weighing of Evidence," (C. Griffin & Co., Ltd., London, 1950)

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 still 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.

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 $\Theta(\text{Guilty} \mid X) = -70$ 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 will be that that is not safe enough. I have my suspicions that actual juries are not even that cautious. 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 worth 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

$$\mathcal{C}(\text{Guilty}\,|\,\text{Motive}) = \mathcal{C}(\text{Guilty}\,|\,X) + 10\log_{10}\frac{(\text{Motive}\,|\,\text{Guilty})}{(\text{Motive}\,|\,\text{Not Guilty})} \qquad (32)$$
$$\approx -70 - 10\,\text{Log}\,(\text{Motive}\,|\,\text{Not Guilty})$$

since $(\text{Motive}\,|\,\text{Guilty}) \approx 1$; i.e., we consider it extremely 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 plausibility $(\text{Motive}\,|\,\text{Not Guilty})$ 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 <u>did</u> 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 know that?

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

$$(\text{Motive} \mid \text{Not Guilty}) = \frac{N_m}{(\text{number of people in New York})} \tag{33}$$

and Equation (32) reduces to

$$e(\text{Guilty} \mid \text{Motive}) = -70 + 10 \log (10^7/N_m) = -10 \log N_m . \tag{34}$$

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 this example 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 tells you.

Interpretation of Random Data From Particle Counters

Now I would like to consider some applications of Bayes' theorem, and comparisons with maximum likelihood, which are less trivial mathematically,

and also correspond pretty closely to problems actually encountered by experimental physicists. Suppose we have a radioactive source (say Co^{60}, for example), which is emitting particles of some sort (say the γ -rays from Co^{60}). There is a counter through which some of these particles pass, and from observing the number of counts registered, we want to infer as much as we can about the number of particles which passed through the counter.

First, we have to define the <u>efficiency</u> of the counter, which I'll denote by "a." By this I mean that each particle passing through the counter has independently the probability "a" of producing a count. The line of reasoning by which we would determine "a" from measurements on the counter requires principles of probability theory which we have not yet discussed (although intuitively, of course, you have no trouble at all in seeing how you would do it). For purposes of this example, we'll just suppose that "a" is a given number.

Now if we <u>knew</u> 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

$$(c|n) = \binom{n}{c} a^c (1-a)^{n-c} \quad . \tag{35}$$

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. 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 through the counter. We're going to assume the number of nuclei N so large 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 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:

$$(n|Np) = \binom{N}{n} p^n (1-p)^{N-n} . \qquad (36)$$

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 when $N \to \infty$, $p \to 0$, $Np = S$ = const., a mathematical argument which is given in every textbook shows that the binomial distribution goes into the simpler Poisson distribution:

$$(n|Np) \cong (n|s) = \frac{e^{-s} S^n}{n!} . \tag{37}$$

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

In all of these talks, I am not going to bother repeating mathematical demonstrations which you can find in any one of a dozen textbooks. I think that most of them are old stuff to you. But if any of these mathematical properties are new to you, let me just say that they are all very elementary, and you can work them out for yourself in less time than it would take to look them up.

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? A handy trick, which almost always works in problems of this sort, is to resolve the proposition C into a set of mutually exclusive alternatives; then apply Rule 2, and then Rule 1. In this case, the propositions Cn for all n form such a set, so we can write

$$(C|S) = \sum_{n=0}^{\infty} (Cn|S) = \sum_{n=0}^{\infty} (C|nS)(n|S) \tag{38}$$

$$= \sum_{n=0}^{\infty} (C|n)(n|S) .$$

Evidently, if we knew the number of particles in the counter, it wouldn't matter any more what s was, so $(c|ns) \equiv (c|n)$. Since we have worked out both $(c|n)$ and $(n|s)$, we just have to substitute them in, and we get

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

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

or,

$$(c|s) = \frac{e^{-sa} (sa)^c}{c!} . \tag{39}$$

This is a Poisson distribution with mean value

$$\bar{c} \equiv \sum_{c=0}^{\infty} c(c|s) = sa . \tag{40}$$

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, of course, that's exactly the guess we would make; but I don't think it's obvious that that would also be the result that you get from calculation. You'd have to go through the calculation to see it; at least, I would.

In practice, the thing which is known is C , and the thing which is unknown would be N . 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 through 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

$$(N|CS) = (N|S)\frac{(C|NS)}{(C|S)} = \frac{(N|S)(C|N)}{(C|S)} . \qquad (41)$$

Again, everything we need for this calculation is on the board, so we just have to substitute:

$$(N|CS) = \frac{\left[\frac{e^{-S}S^N}{N!}\right]\left[\frac{N!}{C!(N-C)!}a^C(1-a)^{N-C}\right]}{\left[\frac{e^{-Sa}(Sa)^C}{C!}\right]} \qquad (42)$$

$$= \frac{e^{-S(1-a)}\left[S(1-a)\right]^{N-C}}{(N-C)!} .$$

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

$$\bar{n} = \sum_n n(n|cs) = c + s(1-a). \tag{43}$$

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 finally make a definite decision, you are going to announce one number. These 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, of course, will depend on what are the consequences of being wrong. If we tell him to take as a criterion that he should minimize the expected square of the error, it is very well known that this leads to taking the mean value $\bar{n}$ of the distribution as his guess. If we ask him to state the one in which he believes most strongly, then he will take the most probable value. But the difference is negligible in this case, because in a Poisson distribution the most probable value 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 pays a visit to our laboratory. We 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, "From a theoretical point of view, the most important general method of estimation so far known is the method of maximum likelihood." (Cramer, p. 498). His likelihood function is, in our notation, $(c|n)$. The value of n which maximizes it is found, within one unit, from

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

or,

$$(n)_{\text{max. likelihood}} = \frac{c}{a} . \qquad (44)$$

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 (40) 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 (44) would be in this case. But what does the robot tell us? Well, he says the best estimate is only

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

More generally, we could write Equation (43) this way:

$$\bar{n} = S + (c - \bar{c}) ; \qquad (45)$$

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 maybe a fluctuation in counting rate by something like the square root of this, $\pm$ 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 $\pm$ 1 in the number of counts. So, you see that an abnormally large number of counts is much easier to blame onto the counter than the particles.

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, 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 way we interpret data in 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 amount 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), 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 rather vague, qualitative fact. We are now going to use Bayes' theorem in four problems where the statistician says categorically that Bayes' theorem is nonsense, and again compare its predictions with maximum likelihood.

Two observers, 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 also notes this fact.) 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 A and B say about the number n_1 of particles? Bayes' theorem for Mr. A reads,

$$(n_1|C_1X_A) = (n_1|X_A)\frac{(C_1|n_1X_A)}{(C_1|X_A)} = \frac{(n_1|X_A)(C_1|n_1)}{(C_1|X_A)} \quad . \qquad (46)$$

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

$$(C_1|X_A) = \sum_{n_1}(C_1|n_1)(n_1|X_A) \quad . \qquad (47)$$

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 arbitrary 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 unform 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 π_1 possesses some intrinsic 'absolute' probability. I deny this. It seems to me that the only meaning of the word 'probability' which makes any sense at all, is simply the best indication of the truth of a proposition, based on whatever evidence we 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 confusion of subject and object. 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

$$(n_1|X_A) = \begin{cases} \frac{1}{N}, & 0 \le n_1 < N \\ 0, & N \le n_1 \end{cases} \qquad (48)$$

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 (46), (47), and we are left with

$$(n_1|c_1 X_A) = \left\{ \begin{array}{ll} \dfrac{(c_1|n_1)}{\sum_{n_1=0}^{N-1}(c_1|n_1)} & , \; 0 \le n_1 < N \\ \\ 0 & , \; N \le n_1 \end{array} \right\} . \quad (49)$$

We have noted, in Equation (44), 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 $n^c(1-a)^n \approx n^c e^{-an}$. Therefore, the sum in (49) converges so rapidly that if N is as large as a few hundred, there is no appreciable difference between

$$\sum_{n=0}^{N-1}(c|n) \quad \text{and} \quad \sum_{n=0}^{\infty}(c|n).$$

So, unless the prior information could justify an upper limit N lower than 200, the value of N turns out not to make any difference. The sum to infinity is easily evaluated, and we get the result

$$(n_1|c_1 X_A) = a(c_1|n_1) = \binom{n_1}{c_1} a^{c_1+1}(1-a)^{n_1-c_1} , \quad (50)$$

So, to Mr. A, the most probable value of n_1 is the same as the maximum-likelihood estimate:

$$\left(n_1\right)_{\substack{\text{most prob.}\\ A}} = \frac{c}{a} = 100 \qquad (51)$$

while the mean value estimate is calculated as follows:

$$\bar{n}_1 - c_1 = \sum_{n_1 = c_1}^{\infty} \frac{n_1!}{c_1!\,(n_1 - c_1 - 1)!}\, a^{c_1+1} (1-a)^{n_1 - c_1}$$

$$= a^{c_1+1}(1-a)(c_1+1) \sum_{n_1 = c_1+1}^{\infty} \binom{n_1}{n_1 - c_1 - 1} (1-a)^{n_1 - c_1 - 1} .$$

The sum is equal to

$$\sum_{m=0}^{\infty} \binom{m + c_1 + 1}{m} (1-a)^m = \sum_{m=0}^{\infty} (-)^m \binom{-c_1 - 2}{m} (1-a)^m$$

$$= \left[1 - (1-a)\right]^{-c_1-2} = \frac{1}{a^{c_1+2}}$$

and, finally, we get

$$\left(\bar{n}_1\right)_A = c_1 + (c_1+1)\frac{1-a}{a} = \frac{c_1 + 1 - a}{a} = 109 . \qquad (52)$$

Now, how about 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 . He will assign a uniform prior probability density for $0 < s < S_o$, and obtain

$$(n_1|X_B) = \int_0^\infty (n_1|s)(s|X_B)\,ds = \frac{1}{S_o}\int_0^{S_o} (n_1|s)\,ds$$

$$= \frac{1}{S_o}\int_0^{S_o} \frac{s^{n_1} e^{-s}}{n_1!}\,ds \quad . \tag{53}$$

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

$$(n_1|X_B) \cong (s|X_B) = \frac{1}{S_o} = \text{const.}, \text{ if } n_1 \ll S_o . \tag{54}$$

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:

$$(n_1|C_1 X_B) = (n_1|C_1 X_A) = a(C_1|n_1) . \tag{55}$$

Jeffreys* has proposed a different way of handling this problem. He suggests that the proper way to express ignorance of a continuous variable known to be positive, is to assign uniform prior probability to its logarithm:

$$(s|X_J) = \frac{1}{s} \quad . \tag{56}$$

Of course, you can't normalize this, but that doesn't stop you from using it. I'll have to admit that I have never been able to follow the argument which Jeffreys advances in support of this rule; 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:

$$(n_1|X_J) = \frac{1}{n_1} \;, \quad (c_1|X_J) = \frac{1}{c_1} \;, \quad (n_1|c_1 X_J) = \frac{c_1}{n_1}(c_1|n_1). \tag{57}$$

This leads to the most probable and mean value estimates:

$$(n_1)_{\substack{\text{most} \\ J}\ \text{prob.}} = \frac{c_1 - 1 + a}{a} = 91 \tag{58}$$

$$(\bar{n}_1)_J = \frac{c}{a} = 100. \tag{59}$$

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

* H. Jeffreys, "Theory of Probability," (Oxford, 1939); Chapter III.

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. 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 prior probabilities 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 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

$$(n_2)_{\substack{\text{most prob.} \\ A}} = \frac{C_2}{a} = 160 \qquad (60)$$

and his mean value estimate is

$$\left(\overline{n}_2\right)_A = \frac{C_2 + 1 - a}{a} = 169 . \qquad (61)$$

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

$$\begin{aligned}(n_1|C_2C_1X_B) &= (n_1|C_1X_B)\frac{(C_2|n_1C_1X_B)}{(C_2|C_1X_B)} \\ &= (n_1|C_1X_B)\ \frac{(C_2|n_1X_B)}{(C_2|C_1X_B)} .\end{aligned} \qquad (62)$$

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

$$\begin{aligned}(C_2|n_1X_B) &= \int_0^\infty (C_2S|n_1X_B)\,ds = \int_0^\infty (C_2|Sn_1)(S|n_1)\,ds \\ &= \int_0^\infty (C_2|S)(S|n_1)\,ds .\end{aligned} \qquad (63)$$

We have already found $(c_2|s)$ in Equation (38), and we need only

$$(s|n_1) = (s|X_B)\frac{(n_1|s)}{(n_1|X_B)} \cong (n_1|s), \text{ if } n_1 \ll s_o \qquad (64)$$

where we have used Equation (54). We have found $(n_1|s)$ in Equation (37), so we have

$$(c_2|n_1X_B) = \int_0^\infty \left[\frac{e^{-sa}(sa)^{c_2}}{c_2!}\right]\left[\frac{e^{-s}s^{n_1}}{n_1!}\right]ds = \binom{n_1+c_2}{c_2}\frac{a^{c_2}}{(1+a)^{n_1+c_2+1}} \qquad (65)$$

Now we just substitute (55) and (65) into (62), carry out an easy summation to get the denominator and the result is

$$(n_1|c_2c_1X_B) = \frac{(2a)^{c_1+c_2+1}}{(c_1+c_2)!(1-a)^{c_1}(1+a)^{c_2+1}}\frac{(n_1+c_2)!}{(n_1-c_1)!}\left(\frac{1-a}{1+a}\right)^{n_1}. \qquad (66)$$

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,

$$(n_1)_{\substack{\text{most prob.}\\ B_2}} = \frac{c_1}{a} + (c_2-c_1)\frac{1-a}{2a} \qquad (67)$$

$$= \frac{c_1+c_2}{2a} + \frac{c_1-c_2}{2}$$

$$= 127.$$

His new mean-value estimate is also readily calculated, and is equal to

$$\left(\bar{n}_1\right)_{B_2} = \frac{c_1 + 1 - a}{a} + \left(c_2 - c_1 - 1\right)\frac{1-a}{2a}$$

$$= \frac{c_1 + c_2 + 1 - a}{2a} + \frac{c_1 - c_2}{2} \qquad (68)$$

$$= 131.5 \quad .$$

You see that both estimates are considerably raised, and the difference between most probable and mean value is only half what it was before. Evidently, this agrees pretty well with common sense; because we see intuitively without any calculation, that to Mr. B knowledge of both c_1 and c_2 enables him to make a better guess about the source strength s. If he could obtain the number of counts in a great many different seconds, c_3 , c_4 , c_5 . . . , he would be able to do better and better, and eventually his estimates of particle numbers would be indistinguishable from those we found from Equation (42), in which the source strength was considered known. This, of course, corresponds exactly to the procedure an experimental physicist would use; he would want to get as much data as possible, use the entire run of data to estimate the source strength, then use this best value for further predictions. I won't go through the details, but you can easily calculate the distributions $\left(s|c_1\right), \left(s|c_1 c_2\right), \left(s|c_1 c_2 c_3\right)\ldots$ and it would turn out that as $m \longrightarrow \infty$, the distribution $\left(s|c_1 \ldots c_m\right)$ 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:

$$(s|c_1 \cdots c_m) \longrightarrow \delta(s - s')$$

where

$$s' \equiv \lim_{m \to \infty}\left[\frac{c_1 + c_2 + \cdots + c_m}{ma}\right] . \tag{69}$$

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

$$(n_2)_{\substack{\text{most prob.} \\ B}} = 133 \tag{70}$$

$$(\bar{n}_2)_B = 137.5 . \tag{71}$$

There is still one feature missing in the comparison of Mr. A and Mr. B in this 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 $\alpha \leq n_1 \leq \beta$. As we will see later in a different problem, the results of doing this would be practically the same as those the statistician would get by an entirely different method, called the method of confidence intervals. But, for present purposes, we will be content to give the standard deviations 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 $(\bar{n}_1 \pm t\sigma)$ satisfies

$$P \geq \left(1 - \frac{1}{t^2}\right) . \tag{72}$$

This tells us nothing when $t \leq 1$, but it tells us more and more as increases beyond unity.

The variances σ^2 of all the distributions we have found are readily calculated. In fact, the calculation of any moment of these distributions is easily performed by making use of the general formula

$$\sum_{m=0}^{\infty} \binom{m+a}{m} m^n x^m = \left(x\frac{d}{dx}\right)^n \frac{1}{(1-x)^{a+1}} , \quad |x|<1, \tag{73}$$

which we have already used in calculation of the mean value of (52). For Mr. A, and Mr. B, and the Jeffreys prior probability distribution, we find the variances

$$\mathrm{Var}\left(n_1 \middle| c_1 X_A\right) = \frac{(c_1+1)(1-a)}{a^2} \tag{74}$$

$$\mathrm{Var}\left(n_1 \middle| c_2 c_1 X_B\right) = \frac{(c_1+c_2+1)(1-a^2)}{4a^2} \tag{75}$$

$$\mathrm{Var}\left(n_1 \middle| c_1 X_J\right) = \frac{c_1(1-a)}{a^2} \tag{76}$$

and the variances for n_2 are found from symmetry.

This has been a rather long discussion, so let's summarize all these results 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) $\pm$ (standard deviation).

		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 $\pm$ s.d.	109 $\pm$ 31	109 $\pm$ 31	169 $\pm$ 48
B	most prob.	100	127	133
	mean $\pm$ s.d.	109 $\pm$ 31	131.5 $\pm$ 26	137.5 $\pm$ 26
J	most prob.	91		
	mean $\pm$ s.d.	100 $\pm$ 30		

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 a substantial decrease in his probable error. However, Mr. B could be helped a good deal more in his estimate of n_1 by acquiring still more data c_3, c_4, ... The standard deviation of the distribution (42) in which the source strength is known exactly, is only $\sqrt{s(1-a)} = 10.8$ for $s = 130$; and Mr. B's standard deviation for his estimate of n_1 would approach this value if we gave him more and more data from other time intervals, such that his estimate of s approached 130.

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 initial information X_β; we would have another example 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.

Well, I said I was going to compare Bayes' theorem with maximum likelihood in this problem. But I have already done that, for Mr. A's most probable values were in all cases just the same as the 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,* "These problems cannot be solved by any theorems of the calculus of probabilities 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; 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 a third 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 and Bayes.

There is a further point which should be made on these estimation problems. For we have seen that the most probable value and the mean value estimates are not the same in general. Which is best? I think this is a matter for common sense to decide. The answer depends on the use to be made of the theory, and on the form of the probability distribution. For example, in Figure 5a we have a distribution for which the most probable value is not only a poorer estimate than the mean value, but is also very unstable against small changes in the problem. But in Figure 5b we have a case where the most probable value is extremely likely to be the correct one, while the mean value is known to be an impossible one. Generally, if the

* A. Wald, "Notes on the Theory of Statistical Estimation and of Testing Hypotheses," (Mimeographed notes, Columbia University, 1941)

distribution has a single peak, the mean value would seem preferable. At at rate, any principle which denies us the choice between them cannot possibly be best in all cases.

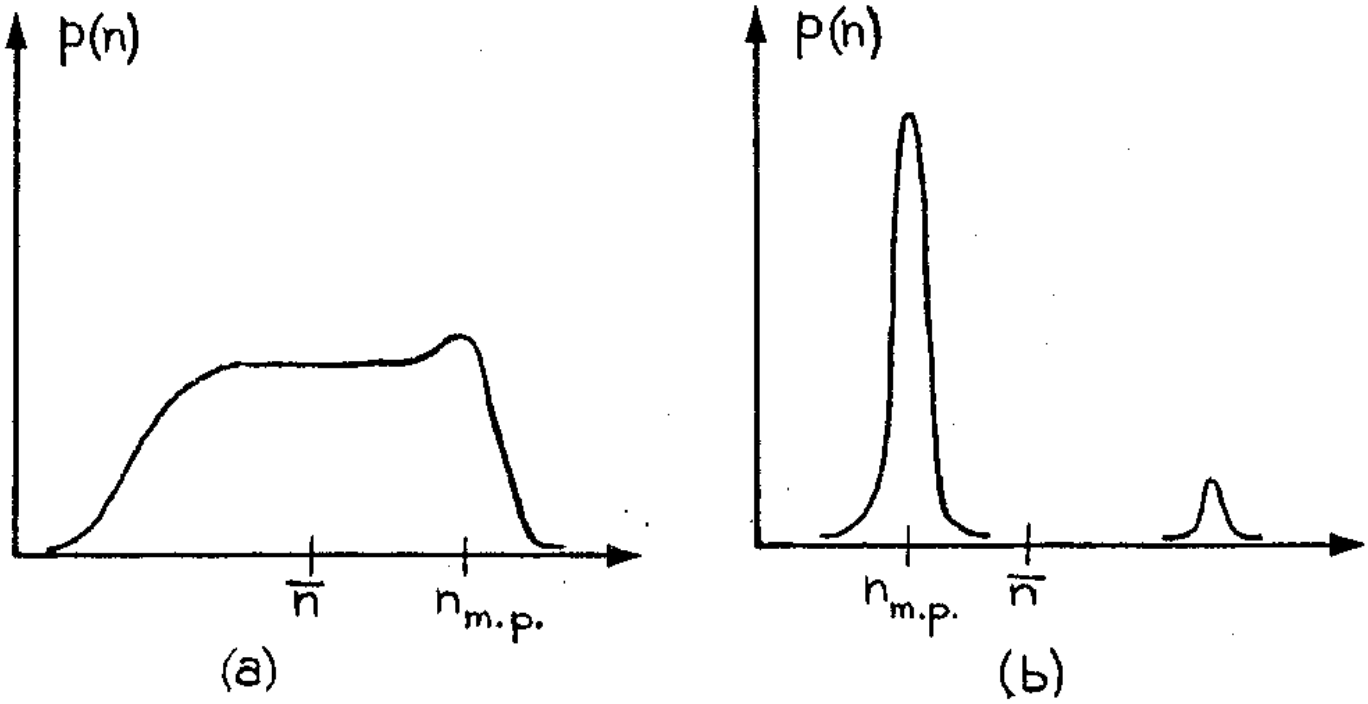


Figure 5.

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 this sort, 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. In any event, whatever desirable properties maximum likelihood might have, asymptotic or otherwise, are also enjoyed by Bayes' theorem with uniform prior probability, because they are

mathematically identical. But Bayes' theorem still has the advantage, 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 5a arise. In the small sample case, the mean-value estimate (i.e., the first moment of the likelihood function) is often far superior to 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 automatically contained in Bayes' theorem as a special case, and whenever it is the appropriate method to use, Bayes' theorem will tell us to use it. If we simply instruct our robot to handle all problems of parameter estimation by Bayes' theorem, he will never do worse than the statistician, he will usually do the same thing, but he will sometimes do better.

LECTURE FOUR

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 couple of simple problems, but he is still not a very efficient reasoning machine, because he has only one means by which he can translate raw information into numerical values of probability; the principle of insufficient reason. In fact, even in order to work through this last example, we had to use principles to set up the problem, which we have not covered yet. He can use insufficient reason if we 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 evidence that gives him some reason for preferring one possibility to some other possibility. What do we do in this case?

Let's imagine a certain class of problems in which the information we give the robot 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 nth 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. This is something which you apparently don't have in Dallas, but we do have it in California, I assure you. The 1,000 cars, are packed 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 pass over an intersection, they go over a machine that weighs each one and totals the result. So we know the total length and the total weight of the 1,000 cars. We can look up statistics from the manufacturers. We know how long the Ford is, how heavy it is; we know 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 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, a robot has no way at all of handling problems of this sort for the time being. 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 some 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 subject to the available information. The most conservative, non-committal distribution is the one which is as "spread out" as possible. It 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 avilable 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. It might 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. But 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 probabiltiy to many cases which were not at all impossible on the evidence we gave him.

Minimum $\sum p_i^2$

Another kind of measure of how spread out a probabiltiy 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

$$\delta\left[\sum_m p_m^2 - \lambda\sum_m m p_m - \mu\sum_m p_m\right] = \tag{77}$$

$$= \sum_m (2p_m - \lambda m - \mu)\delta p_m = 0$$

where λ and μ are Lagrange multipliers.

So p_m will always be a linear function of m :

$$2p_m - \lambda m - \mu = 0 . \tag{78}$$

Now, μ and λ are found from

$$\sum_m p_m = 1 \quad , \quad \sum_m m p_m = \bar{m} ,$$

where $\overline{m}$ is the average value of m .

Let's investigate this and actually 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 $\sum_m p_m$ is

$$p_1 = \frac{4}{3} - \frac{\bar{m}}{2}$$

$$p_2 = \frac{1}{3} \qquad (79)$$

$$p_3 = \frac{\bar{m}}{2} - \frac{2}{3}.$$

In Figure 6 these results are plotted.

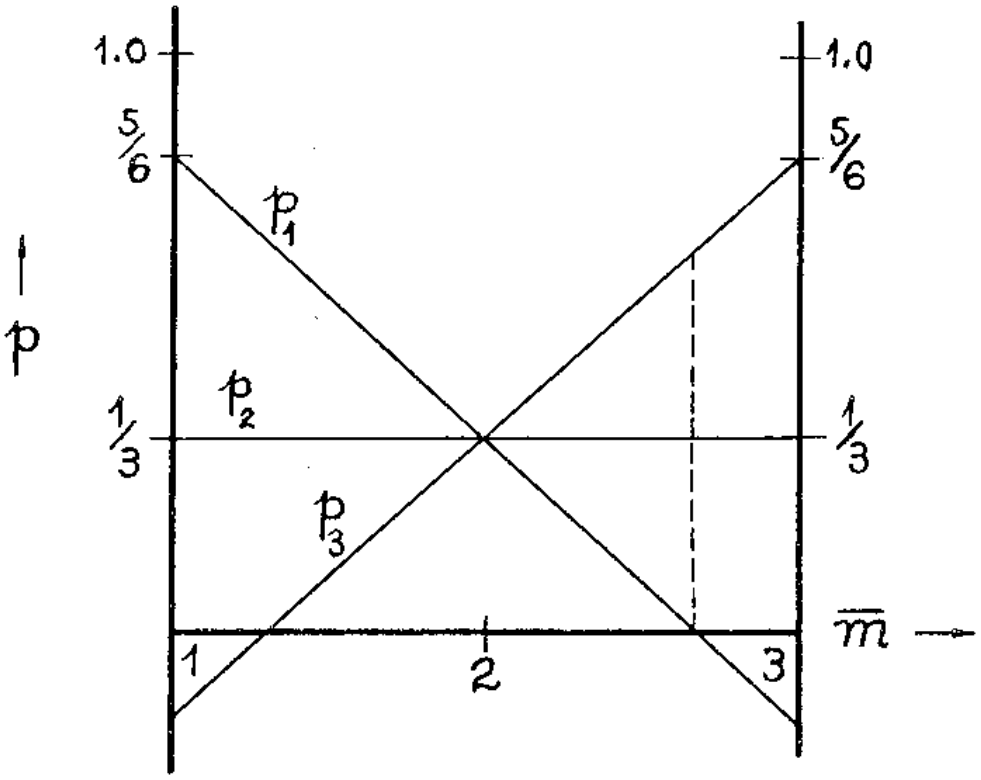


Figure 6.

This shows that p_1 and p_3 become negative and we can't use the solution in the regions where p is negative. In these regions let's say we will replace the negative values by zero and then adjust the other probabilities to account for this action. If we do this the results are shown in Figure 7.

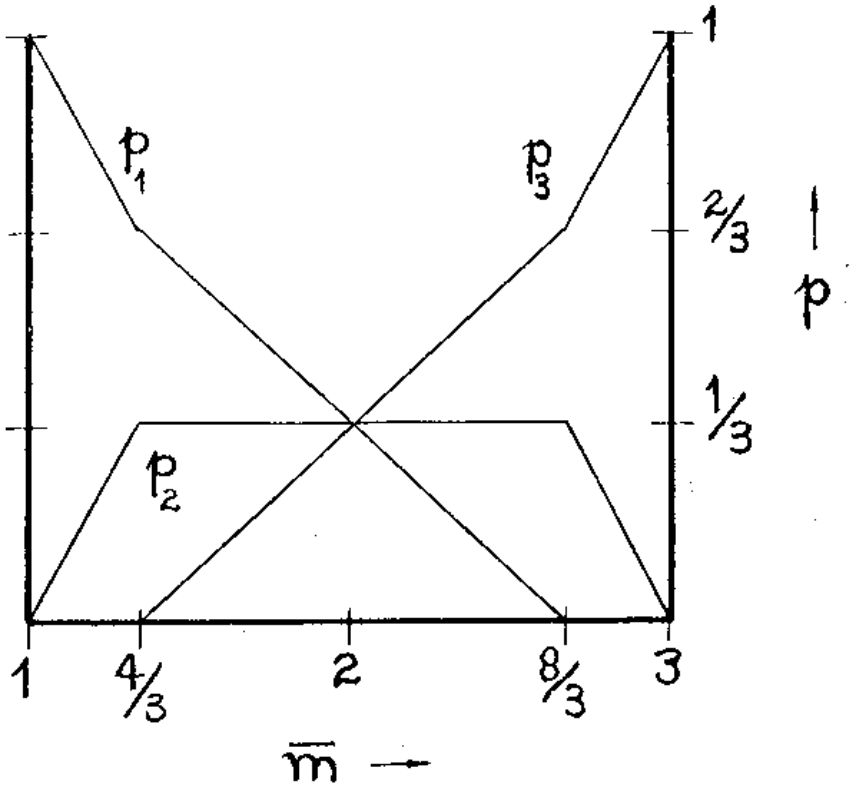


Figure 7.

All right, so that's what this criterion will give to us. Now, is the robot behaving in a reasonable way if we built 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 evidence we gave him which said one was impossible. So he is still jumping to conclusions which are not warranted by the evidence we give him. But the idea behind it still looks like a good one. There should be some consistent measure of the uniformity of a probability distribution which we can maximize, subject to constraints, and which will have the property that it does not permit the robot to draw any conclusions unless those conclusions are really warranted by the evidence he has.

Entropy

Well, at this stage we turn to probably the most important theorem in Shannon's work on information theory. 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 the first two lectures; 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. This condition takes the form that in case the p_i are all equal, the quantity

$$A(n) = H_n\left(\frac{1}{n}, \cdots, \frac{1}{n}\right)$$

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. By this we mean, as before, that if there are several different ways 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 functional equations. 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 $p_2 + p_3 = q$. His new uncertainty $H_3(p_1, p_2, p_3)$ must be the old value, plus the additional uncertainty as to events 2 and 3, weighted according to the probability q that this additional uncertainty will arise:

$$H_3(p_1,p_2,p_3) = H_2(p_1,q) + qH_2\left(\frac{p_2}{q},\frac{p_3}{q}\right) . \qquad (80)$$

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 (80) 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. I don't think that either this demonstration or the ones we gave in the first two lectures are anywhere near in satisfactory form yet. A good mathematician should be able to invent a much simpler and more elegant way of proving these things; but for the time being we'll have to get along on what we have. The best attitude, in my opinion, is that at present the justification for this theory lies partly in these consistency arguments, partly in the fact that so many different lines of reasoning lead to the same conclusions, and partly in the fact that it all works out so well in practice.

At any rate, the next step is perfectly rigorous, because Shannon did it, and I didn't. He proves that the only function H_n which satisfies all the above requirements is

$$H_n(p_1 \cdots p_n) = -\sum_{i=1}^{n} p_i \log p_i . \qquad (81)$$

The only arbitrariness is that we have the option of taking the logarithm to any base we please, since this corresponds only to a multiplicative constant in H_n. This quantity we will, of course, call entropy 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.

It would be a big mistake to try to read too much philosophical significance into this theorem. 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 this theorem, that it was in no way necessary for the theory to follow. By this he meant that the inequalities which H_n satisfies are already quite sufficient to justify its use; it does not really need the further support of the above theorem. 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 both the principle of insufficient reason and the frequency interpretation 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.

First, let's see how it would work out if we ask the robot to assign probabilities in such a way that the entropy is maximized subject to the available information.

We can use our Lagrange multiplier argument again to solve this problem, i.e.,

$$\delta\left[H_n(p_1,\ldots,p_n) - \lambda\sum_m m p_m - \mu\sum_m p_m\right] =$$

$$= \sum_{m=1}^{n}\left[\frac{\partial H_n}{\partial p_m} - \lambda m - \mu\right]\delta p_m = 0.$$

Now,

$$\frac{\partial H_n}{\partial p_m} = -\log p_m - 1 \qquad (82)$$

so our solution is

$$p_m = e^{-\lambda_0 - \lambda m} , \tag{83}$$

where $\lambda_0 \equiv \mu + 1.$

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_0 and λ by forcing this to agree with the fact that the sum of the P's must be one and that the average value must be equal to the average 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

$$Z(\lambda) \equiv \sum_m e^{-\lambda m} \tag{84}$$

which we call the partition function. The equations which fix our Lagrange multipliers are then

$$\lambda_0 = \log Z(\lambda) \tag{85}$$

and

$$\overline{m} = - \frac{\partial}{\partial \lambda} \log Z(\lambda) . \tag{86}$$

Once again, let's put this to the test that we gave our others, in this case, where m can take on only three different values. We find easily that $p_1(\bar{m})$, $p_2(\bar{m})$, $p_3(\bar{m})$ are given in parametric form by

$$p_k = \frac{\exp(2-k)\lambda}{1+2\cosh\lambda} \quad , \quad k = 1, 2, 3 .$$

$$\bar{m} = \frac{e^{2\lambda} + 2e^{\lambda} + 3}{e^{2\lambda} + e^{\lambda} + 1} .$$

In a more complicated problem we would just have to leave it in parametric form, but in this particular case we can eliminate λ mathematically, leading to the explicit solution

$$p_1 = \frac{3 - \bar{m} - p_2}{2}$$

$$p_2 = \frac{1}{3}\left[\sqrt{4 - 3(\bar{m}-2)^2} - 1\right] \qquad (87)$$

$$p_3 = \frac{\bar{m} - 1 - p_2}{2} .$$

These results are plotted in Figure 8, 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.

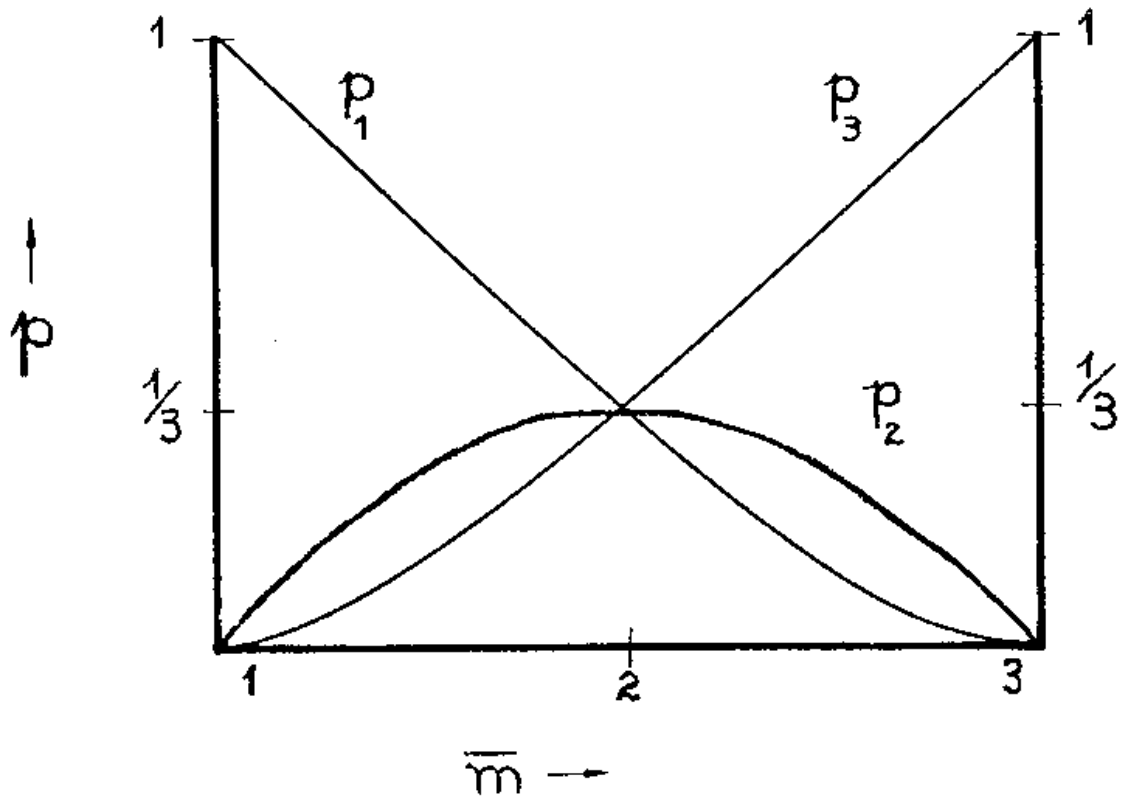


Figure 8.

Let's just notice that we have finally arrived here at a solution which meets the objections we had to the first two examples. 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 place where a probability goes to zero is in the limit where the average 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 the same value and the same slope as the maximum entropy solution, at the end points and at the middle. It represents the best straight-line approximation you could have made to the maximum entropy solution.

Generalization

The maximum-entropy solution can be generalized in many ways. Suppose we have a problem in which there are m different functions, i.e.,

$$f_k(x_i) \qquad \text{where} \quad 1 \leq i \leq n \qquad 1 \leq k \leq m$$

where n is the number of possible values x can assume. The average of $f_k(x_i)$ is known for each of the possible values of k, i.e.,

$$\left\langle f_k(x_i) \right\rangle = \sum_{i=1}^{n} p_i f_k(x_i). \tag{88}$$

If we decide to build this entropy principle into our robot's brain, and we ask him to reason, given this information, he 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.

$$\delta\Big[H(p_1,\dots,p_n) - (\lambda_0 - 1)\sum_i p_i - \lambda_1 \sum_i p_i f_1(x_i) -$$

$$\dots - \lambda_m \sum_i p_i f_m(x_i)\Big] =$$

$$= \sum_i \left[\frac{\partial H}{\partial p_i} - (\lambda_0 - 1) - \lambda_1 f_1(x_i) - \dots - \lambda_m f_m(x_i)\right]\delta p_i = 0$$

and so from (82) our solution is the following:

$$p_i = e^{-\lambda_0 - \lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)} \quad . \tag{89}$$

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.,

$$1 = \sum_i p_i = e^{-\lambda_0} \sum_i e^{-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)} \quad . \tag{90}$$

I will define a partition function as

$$Z(\lambda_1 \cdots \lambda_m) \equiv \sum_{i=1}^{n} e^{-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)} \tag{91}$$

then (90) reduces to

$$\lambda_0 = \log Z(\lambda_1 \cdots \lambda_m) \quad . \tag{92}$$

The average value (88) of f_k is then

$$\left\langle f_k(x_i) \right\rangle = e^{-\lambda_0} \sum_i e^{-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)} f_k(x_i),$$

or,

$$\langle f_k \rangle = -\frac{\partial}{\partial \lambda_k} \log Z . \tag{93}$$

What is the maximum value of the entropy that we get from this probability distribution? I am going to call entropy, S, the way physicists do, instead of H, the way information theory people do:

$$S = -\sum_{i=1}^{m} p_i \log p_i . \tag{94}$$

From (89) we find that

$$S = \lambda_0 + \lambda_1 \langle f_1 \rangle + \cdots + \lambda_m \langle f_m \rangle . \tag{95}$$

Now consider the maximum attainable entropy as a function of the given mean values of f_k;

$$S = S(\langle f_1 \rangle, \cdots, \langle f_m \rangle) ,$$

and let's see what the derivative of S is with respect to some particular $\langle f_k \rangle$;

$$\frac{\partial S}{\partial \langle f_k \rangle} = \frac{\partial \lambda_0}{\partial \langle f_k \rangle} + \frac{\partial \lambda_1}{\partial \langle f_k \rangle} \langle f_1 \rangle + \cdots + \frac{\partial \lambda_m}{\partial \langle f_k \rangle} \cdot \langle f_m \rangle + \lambda_k$$

and, since $\lambda_0 = \log Z$ we can expand this to

$$\frac{\partial S}{\partial \langle f_k \rangle} = \sum_{i=1}^{m} \left[\frac{\partial \log Z}{\partial \lambda_i} \cdot \frac{\partial \lambda_i}{\partial \langle f_k \rangle} + \frac{\partial \lambda_i}{\partial \langle f_k \rangle} \langle f_i \rangle \right] + \lambda_k .$$

But, by (93) the expression in brackets is identically equal to zero; so we have simply

$$\lambda_k = \frac{\partial S}{\partial \langle f_k \rangle} \quad . \tag{96}$$

Now, suppose f_k contains some parameter α in addition to x ; what is the expectation value of the derivative over this distribution?

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = \sum_{i=1}^{n} \frac{\partial f_k(x_i)}{\partial \alpha} \cdot p_i =$$

$$= e^{-\lambda_0} \sum_{i=1}^{n} \frac{\partial f_k(x_i)}{\partial \alpha} e^{-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)} .$$

Now, this sum we have written is proportional to $\partial Z/\partial \alpha$ so that, assuming that α occurs only in $f_k(x_i)$, we can write

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = e^{-\lambda_0} \left[\frac{-1}{\lambda_k} \frac{\partial Z}{\partial \alpha} \right] = -\frac{1}{\lambda_k} \frac{\partial}{\partial \alpha} (\log Z) . \qquad (97)$$

So, there's a large 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.

For example, we tell our robot that the energy of a system has been measured, and there are many different quantum levels E_i. This

measurement of energy, let us say, gives us some sort of average value $\langle E_i \rangle$. If this is the only information we give him, then if he wants to infer any other property of the system on the basis of this knowledge, we will build his brain so that he evaluates the partition function Z ,

$$Z(\lambda) = \sum_i e^{-\lambda E_i} \quad .$$

The probabilities he will assign to the quantum states are

$$p_i = e^{-\lambda_0 - \lambda E_i} \quad .$$

Evidently, this is the Boltzmann distribution, and the robot will soon discover that $\lambda = \frac{1}{kT}$.

More generally, in any problem where we have average values of various physical quantities such as energy, volume, number of molecules of various types, electric or magnetic moments, etc., we ask our robot to do the best job he can of predicting the values of other physical quantities. He will do it by maximizing the entropy, subject to all the constraints we gave him. He will write down all the equations of conventional statistical mechanics such as the distribution of the "grand canonical ensemble," and various generalizations thereof. Our relations (92), (93), (96), and (97) will then acquire all sorts of physical content . They will be the general laws of thermodynamics, giving pressure, stress, chemical potentials, specific heats, and so on. With a little further generalization, which we will look at later, this same process will yield a general statistical mechanics of irreversible processes.

Boltzmann's Approach to Maximum Entropy

Now this result is so important, because of the many applications it opens up, that it's a good idea to see what we're doing in as many different ways as we possibly can. I'd like to give one more line of reasoning which we could have gone through which would lead us to writing down these same equations.

Let's go back and consider the dice problem that we had at the beginning. We have tossed this die on the black table many times and we have the multiple exposure from which we can deduce the total number of spots which were on top during the entire experiment. There were N tosses, and they yielded a total of S spots. We are to do the best job we can of inferring the number of times the six was on top, and the number of times the five was on top, and so on, from nothing but this information.

We could break the situation down into all of the conceivable events that could have happened. Let's say the first toss could have given a 1, 2, 3, 4, 5, or 6, which ever one it was, and the second toss could also have given a 1, 2, ... or 6. And eventually the nth toss could have given any number from one to six. There were, a priori, 6^N conceivable things which could have happened in the course of this experiment. But the information which we have excludes some of those. I will define N_m as the number of times m spots were up. The information we have tells us the following,

$$\sum_{m=1}^{6} N_m = N \ ,$$

the total number of tosses, and

$$\sum_{m=1}^{6} m N_m = S$$

the total number of spots. This last piece of information excludes a large number of possible sequences of events, but still leaves an enormous number which are perfectly possible as far as we know. Evidently, this amount of information not only doesn't show us which sequence of events happened, it doesn't even tell us what the N_m's are. So, if we have to do the best job we can of guessing the N_m's , then we must do some plausible reasoning. How many of the 6^n sequences would have given me a particular set of N_m's ? Well, that is a <u>combinatorial</u> problem which we find solved on page 1 of every textbook on statistical mechanics; namely,

$$W \equiv \frac{N!}{N_1! \cdots N_6!} \quad \text{different ways.}$$

Now, out of all the sets of N's which satisfy the two conditions we know are satisfied, which could have been realized in the greatest number of ways? The mathematical problem is to maximize this combinatorial factor, subject to these constraints representing our information. If we had to guess the N_m's on the basis of no information except this, then it certainly seems that a reasonable way to do it is to see which set of N's could have happened in the greatest possible number of ways. Well, instead of maximizing this combinatorial factor, of course, it is just as good if we maximize any monotonic function of it; and mathematically

because of Mr. Stirling, it is easier to maximize the logarithm of it. Stirling's approximation is

$$\log N! = N\log N - N + \frac{1}{2}\log(2\pi N) + O\left(\frac{1}{N}\right).$$

Now, let us write $\frac{1}{N}\log W$, disregarding terms which tend to zero as $N \longrightarrow \infty$:

$$\frac{1}{N}\log W \cong \log N - 1 - \frac{1}{N}\sum_{m=1}^{6}\left(N_m \log N_m - N_m\right)$$

$$\cong -\sum_{m=1}^{6}\frac{N_m}{N}\log\left(\frac{N_m}{N}\right). \tag{98}$$

So, we can state our problem this way. I want to maximize (98), subject to constraints which I can write as

$$\sum_m \left(\frac{N_m}{N}\right) = 1 \quad , \quad \sum_m m\left(\frac{N_m}{N}\right) = \overline{m} \tag{99}$$

where $\overline{m} \equiv S/N$ is the average number of spots on top. Now, if I were to write N_m/N as p_m, then I have formulated exactly the same mathematical problem that Shannon's theorem led us to, so we'll get the same answer (83). This is a method which has been used in physics since the time of Boltzmann; but we have presented it here in terms of dice, so as to make it clear that the line of reasoning depends not at all on the laws of physics.

In statistical mechanics we are concerned with such things as the distribution of energy among many molecules. Given N molecules, with total energy E, what is the best guess we can make as to the number of molecules in a particular energy level? The set of occupation numbers N_i, which maximizes $\log W$ under these constraints, is the one we call the "most probable" distribution. By that we really mean nothing more than that it is the distribution which can be realized in the greatest number of ways. If we try to go further and assert that it is also the most likely to occur in nature, then we have gone beyond what can be proved by deductive reasoning. We have then done a piece of plausible reasoning - a very excellent piece of plausible reasoning, whose predictions are almost always in agreement with experiment.

I should point out that when we maximize these multinomial coefficients

$$\frac{N!}{N_1! \cdot \cdot \cdot \cdot \cdot N_k!}\ , \tag{100}$$

if N is a large number, the maximum we get is enormously sharp. The set of N_m's which can be realized in the greatest number of ways can be realized in overwhelmingly the greatest number of ways. If N is comparable to Avogadro's number, then for every way in which the N_m's could be significantly different from the Boltzmann distribution, there's something like $10^{10^{20}}$ ways in which they could agree with it.

Notice also that N drops out of the equations in the limit $N \to \infty$, so the result becomes independent of the size of the system. If we were considering finite N and you wanted to take all these other terms in the Stirling approximation into account, you would come out with very slightly different results.

Now, let's try to see exactly how these two approaches are related. Let's consider again the example of the dice tossing. The first approach, via Shannon's theorem, and the second, via the enumeration of all possibilities in a long sequence of tossing, led to the same mathematical problem; but still there was a very fundamental difference between them. In the first solution, maximization of entropy subject to a given mathematical expectation of m, led us to assign probability p_m for the m'th face to come up in a single toss. In the second, our result was the best plausible guess as to the frequency N_m/N with which the m'th face would come up in an indefinitely large number of tosses. The only thing we have shown so far is that the probability at a single trial is numerically equal to an estimate of frequency which is "best" according to a certain criterion. Can we find a single approach which gives both of these results automatically?

Well, what is to prevent us from studying the problem of tosses by direct application of insufficient reason and Bayes' theorem? In N tosses, there are a priori 6^N conceivable outcomes; and with no other information, insufficient reason leads us to assign equal probabilities $(O_i|X) = 6^{-N}$ to each of the, where O_i stands for the i'th

outcome, and i runs from 1 to 6^N. Now we learn that in the N tosses, there were a total of S spots on top. By Bayes' theorem, the probability of any particular outcome is now

$$(O_i|SX) = (O_i|X)\frac{(S|O_iX)}{(S|X)}. \tag{101}$$

But $(S|O_iX)$ is unity if O_i is one of the outcomes with S spots, zero otherwise. Therefore, defining K(N,S) as the total number of outcomes leading to S spots, we have $(O_i|SX) = K^{-1}$ if O_i corresponds to S spots, $(O_i|SX) = 0$ otherwise. The distribution $(O_i|SX)$ contains all the information we have, and so from it we should be able to find <u>both</u> the probability p_m assigned to face m at a single trial (say the first), and the most likely frequency in a large number of trials. The number K(N,S) is to be calculated from

$$K(N,S) = \sum_{(N_m)} \frac{N!}{N_1!\cdots N_6!} \tag{102}$$

where the sum is taken over all N_m compatible with (99). I won't bother to go through this calculation, because you can find it in any statistical mechanics textbook under the heading "Darwin-Fowler Method." Let me just point out the way in which it unifies the two other approaches.

The probability of the m'th face on the first toss is evidently just the fraction of all the K(N,S) possibilities in which the first toss gives the m'th face:

$$p_m = \frac{K(N-1, S-m)}{K(N,S)} . \tag{103}$$

For estimation of the N_m , we might take either the most probable values or the mean values over the distribution $(O_i|SX)$. By now, you will not be surprised if I tell you that in the limit $N \to \infty$, the most probable and mean values of (N_m/N), and the expression (103), all approach our maximum-entropy solution (83). The demonstration is elementary, and requires nothing more than transcription of standard results of the Darwin-Fowler method into our notation.

I have taken the trouble to derive, re-derive, and re-re-derive a simple and well-known mathematical result, because in so doing we have learned something about interpretation of probability theory that is neither simple nor well-known. Virtually all of the modern literature on probability is based on the view that, for some reason, a probability is not respectable unless it is also a frequency. Because of this, the Laplace formulation of probability theory and the principle of insufficient reason are denounced as nonsense. But now we are beginning to see the connection between probability and frequency. In one case (easily generalized) we have seen that the probability which Laplace's theory assigns to an event at a single trial is numerically equal to the best estimate of frequency in a large number of trials. This equality, far from being in conflict with the principles Laplace gave us, is an elementary consequence of those principles.

There are, of course, many different connections between probability and frequency. Some of them we have seen already, in the quality-control testing and particle counter problems, as well as in the maximum-entropy argument. But notice that we never had to stress them in order to solve the problem. We will see this happening all the time; whenever there is any relation between probability and frequency in a problem, this relation will appear automatically, as a consequence of the theory. It never has to be introduced as a separate postulate.

We've seen in one example that you can interpret the principle of maximum entropy as giving us the probability assignment which is "maximally noncommittal" with regard to missing information; or equally well as a slick way of carrying out this process of enumerating all of the possibilities in a very large collection of events, doing a combinatorial analysis, and placing your bets on the situation that can happen in the greatest number of ways. Now, this will work for any problem in which we use the principle of maximum entropy, and in which it makes sense to imagine the experiment repeated a large number of times. You could always make the appropriate generalization of our argument. But once having seen that this process is always going to lead to these same mathematical results, then you don't have to go through this tedious business of enumeration and combinatorial analysis any more. You can simply maximize the entropy and you know that this is the answer you would have got if you had gone through the enumeration.

You might, incidentally, be amused to see the solution of our broken window problem. If we have given that the mathematical expectation of number of pieces is $\langle m \rangle = 10$, then the probability for a window to be broken into m pieces is

$$p_m = \frac{1}{9}\left(\frac{9}{10}\right)^m, \quad 1 \leq m < \infty \quad .$$

That, again, could be interpreted by analyzing a long sequence of window breaking into every possible thing that could have happened, and this result gives also the frequencies according to which the observed fact could have happened in the greatest number of ways.

Now, of what possible use could a result like this maximum entropy solution be for the broken window problem? It certainly seems a rather idle thing to be assigning probabilities of this sort. What could I do with this probability distribution that would be of the least use to anybody? Well, in the case of the broken windows, probably nothing. That was just to illustrate a point. The case of the application of this to the prediction of thermodynamic properties, however, is not at all trivial or useless. But still, the theory has been developed in a way which makes it look very suspicious. We have given a very tiny amount of information. To be sure, we found definite mathematical rules which led us to a definite probability assignment. But, if that is its only justification, what right have we to expect that predictions made from this maximum entropy distribution will have anything at all to do with experimental facts?

Why Does Statistical Mechanics Work?

I'm given, say, the average energy of a system and there are $10^{10^{20}}$ different quantum states in which the system might be. Just knowledge of the average energy is certainly an extremely minute piece of information in a situation like this. And yet, I've got the job of predicting, just from this information, what the pressure might be, what the chemical potentials might be, what the intensity of magnetization might be; things of that sort. Well, of course, we can calculate mean values from our maximum entropy distributions for these things, and it is an experimental fact that they give the right answers. But it certainly doesn't look, from anything we've done so far, that we have any right to expect it to give right answers. There's such an enormous uncertainty that how could we possibly come out with anything reliable from this? Well, now here's a trick. What is it that we are interested in predicting in statistical mechanics? We are not interested in predicting every property of a system.

You go into the lab and perform an experiment. You observe some effect that seems to happen once, but you can't make it repeat. You don't write this up and publish it. By common consent, this is not done. What you saw of course was just as much a legitimate phenomenon as any of the other things. But we have made an arbitrary and unwritten agreement that the subject matter of thermodynamics is restricted to the things which we can reproduce. Now, this is not a trivial step; this is an extremely big restriction. We are restricting our attention to an extremely small fraction of all phenomena which could happen. By phenomena which could happen, I mean phenomena which are compatible with the known laws of physics.

All sorts of behavior are allowed by the Schroedinger equation, which we consider entirely possible. For example, all the air in this room might suddenly more to the back half and leave me up here without anything to breathe. There's nothing in the laws of physics which says that can't happen. But, for some reason, we consider that extremely improbable. And it certainly would not be an experimentally reproducible thing. So, let's look at it in this light; that we're not going to ask this theory to predict everything that a system can do. It's obvious that we couldn't possibly expect it to do that. We're going to ask, is it possible that this theory might predict experimentally reproducible phenomena? If it does that, it will do all that we ever really ask of statistical mechanics.

Well, now, let's make it hard to believe that anything could ever be experimentally reproducible. When we learn how to restore the plausibility, we will understand why this theory works.

So, let's imagine any macroscopic experiment - from a nuclear magnetic resonance measurement, to throwing a baseball. The initial conditions of this experiment are hardly under our control at all, if you look at it from the microscopic standpoint. Again, in typical size systems that we work with in the laboratory, $10^{10^{20}}$ is a pretty good statement of the number of different initial quantum states in which the system might be, as far as we know. This represents the limit of our degree of control over the initial conditions. The number $10^{10^{20}}$ is a very large number; I'll write it another way,

$$10^{10^{20}} = \left(10^{10^{10}}\right)^{10^{10}}.$$

During the experiment we do various things to our system; we push it and pull it; apply RF magnetic fields to it; and various things; and well, you see how the system behaves.

Now, the forces we apply to the system during the course of the experiment, also, if you consider them from the atomic scale, will never be repeated, even approximately, no matter how many times you repeat the experiment, because we never control the conditions of an experiment to atomic precision. So, it's clear that no matter how many times we repeat this experiment, we are never going to repeat the initial quantum state, we're never going to repeat the forces applied, even approximately. Well, how could it ever happen that the result is reproducible then? The experimental fact is that it's easy to reach a degree of control over conditions at which we see reproducible behavior for practically everything on the macroscopic scale. How could it be that anything is ever reproducible?

It seems to me that there is only one answer to this. The fact that a phenomenon can be reproduced experimentally shows that this phenomenon must be characteristic of <u>each</u> of the great majority of all these states, when subjected to <u>each</u> of the great majority of all possible forces we might have applied to it. I don't see any other way in which it could be reproducible. So, now, we begin to see the ray of light, why this theory works, and why it is perfectly reliable for predicting thermodynamic properties. The only thermodynamic properties we know about are the ones which are characteristic of practically all the possible states. Naturally, if we have to predict such a property, then we can do a very good job of

predicting with a very crude amount of initial information. For every way in which the system could behave in a strange manner there are billions and billions of ways in which it could behave in the way we are used to seeing in the laboratory. Our lack of detailed information about the state of the system corresponds exactly to, because it is due to, our lack of detailed control over experimental conditions. The only things which we can predict reliably from the maximum entropy distribution are things which would have been characteristic of practically all the states to which we have assigned any appreciable amount of probability. For any such property, the mean value calculated from the maximum entropy distribution will, of course, be the actual value for practically all possible states. So in this way we can see why it is that the class of phenomena we can predict from maximum entropy arguments and the class of phenomena which are experimentally reproducible, are exactly the same. Let's look at that more carefully.

I would like now to consider briefly one of the arguments that has been given in the past and is still quite commonly seen in textbooks. This is the argument: "The reason why statistical mechanics works is that what we measure experimentally is only a time average over time which is long from an atomic point of view, and that given an ergodic hypothesis, time averages and ensemble averages would be the same." I'd just like to point out something which is not at all new. Even though you might succeed beyond your wildest dreams in proving these ergodic properties (which, of course, no one has succeeded in doing); i.e., even if you could prove

rigorously and universally that ensemble averages were equal to time averages for systems, you still would not have explained why ensemble averages are equal to experimental values. Because the ergodic theorem applies only for time averages over infinite length of time, and you have to consider what is the length of time over which you have to average before you can be sure that the average has approached its limiting value. Let's answer this in two different ways.

First, let's imagine we have a rather smallish crystal which contains 10^{21} nuclei of spin 1/2. I'm going to consider only the states of the nuclear spin system. There are

$$2^N = 2^{10^{21}} > 10^{10^{20}} = \left(10^{10^{10}}\right)^{10^{10}}$$

different possible spin states for the system as a whole. Each spin could be up or down. How rapidly are transitions going to be made between different states of the spin system? To get an idea of the sort of times that are involved, let's say the spins interact with each other, but there are lattice vibrations in the crystal which modulate these terms, and, therefore, each spin sees a varying magnetic field due to the moving spins around it and this can induce transitions where spins flip up or down. Lattice vibrations have frequencies of the order of 10^{12} cycles per second at room temperature. Let's assume that each spin in the crystal has a good chance of making its spin flip once every cycle of the lattice

vibration. In actual fact, the probability of the spin flipping is very much less than that. But, if we had that case, then how many transitions per second would we have to some new quantum state of the spin system? We'd have 10^{21} nuclei, each making 10^{12} spin flips per second. There'd be about 10^{33} transitions per second to new quantum states. How long would it take before the spin system had gone into each of the possible quantum states with reasonable probability? Well, it would take something like the ratio of the two numbers:

$$10^{(10^{20}-33)} \text{ seconds.}$$

Now, the geologists and the astronomers tell us that the age of the universe is something like $6X10^{9}$ years, and this is not far from 10^{17} seconds.

Well, this is a modern version of what the Ehrenfests pointed out in 1911, and Boltzmann a few years before that. The time that it would take to make any reasonably complete sampling of all the microscopic conditions is simply fantastic for any system of the size on which we do experiments. And so, you could be sure the time averages approached the ensemble averages only if the time averages were over these enormous times. If you had to prove anything about time averages over shorter times than that, then you would be forced to make special assumptions about "smoothness" properties for the particular quantities that you are going to predict. But that, of course, is exactly what I did when I said that we are interested in predicting only experimentally reproducible things.

There is a second, even simpler, way of seeing this. For if a system could come close to reaching all microscopic conditions during the time in which measurements are made, the experimental results would always correspond to equilibrium values. We would not even know about irreversible processes. The fact that we can measure the rate of an irreversible process already shows that the time required for exploration of all microscopic conditions must be much longer than the time required to make our measurements.

This shows why, no matter how successful you were in solving the very difficult mathematical problems associated with ergodicity and metric transitivity, it would not have any relevance to statistical mechanics. The problem is not to explain why ensemble averages are equal to time averages; it is to explain the much more restrictive condition that ensemble averages are equal to experimental values. Once you have explained this, then equality with time averages would appear, not as the reason for the success of statistical mechanics, but as a trivial consequence of that success, in the special case of equilibrium conditions. We have seen, in the condition of experimental reproducibility, the reason why ensemble averages will still be equal to experimental values in non-equilibrium conditions, where they are <u>not</u> equal to time averages.

Now this gives us a completely different conception of the role of probability in statistical mechanics. We recognize that a probability distribution over states (and here I use the term "state" in the Gibbs, or Γ-space, sense) does not describe any property of the system, but only a certain state of knowledge about that system. Then, of course, we are not allowed to say that the success of statistical mechanics is due to our having found the "correct" probability distribution; it is quite

meaningless to say that one distribution is correct, another incorrect. But as soon as we see that the only properties we are interested in predicting are the ones which would have been characteristic of practically all the possible states anyway, it follows that we don't really need to use probability at all. We could just choose at random any one of the possible initial states, solve the time-dependent Schrödinger equation for it, and see what predictions we get. There would be an overwhelming probability that we would get the right prediction for any experimentally reproducible phenomenon, whether reversible or irreversible in the thermodynamic sense. An overwhelming probability, but not quite certainty; to try to predict the way the air behaves in this room, there's a very small chance that you might be unlucky and choose the initial state for which the air does go to the back of the room. There will always be a little danger of getting the wrong answer if you do it that way. And, in principle, the only thing which using a probability distribution does for us is that it protects us against that danger. By averaging our predictions over many possible states, we suppress this small minority which would have given a different result. From the standpoint of principle, it is purely incidental that this also simplifies the mathematics by about 10^{20} orders of magnitude. Now, we can get still more out of this.

Suppose we make predictions by the principle of maximum entropy. We then perform the experiment and find that what we predicted was right. Is there anything further we can conclude then? I don't think there is. We made the best guess we could, and what was already strongly indicated by the evidence turned out to be in fact true. Then there is nothing more to be said.

But suppose the predictions turned out to be wrong. Now, can you make more conclusions? Well, of course you can. If statistical mechanics fails to give the right answer, that is a situation which is much more interesting than if it works. Because then we can carry the reasoning a step further.

Let me predict that a certain thing should happen. It is an experimentally reproducible fact that this thing does not happen. In the class of states to which I assigned high probability in my maximum entropy distribution, the overwhelming majority would have given this predicted behavior. But experimentally, we know that the overwhelming majority of all states which are allowed by the experimental conditions, the true "possible states," do not have this property. It follows that my enumeration of possible states was not right; there's something which is keeping the system away from the great majority of all the states which I thought were possible. There must be an enormous number of new possible states that I didn't know about, or there must be new constraints on the states that I did know about. In other words, there's a new law of physics. Perhaps a new "constant of the motion."

As soon as we see that statistical mechanics is not a "physical theory," but only a method of plausible reasoning, we see the reasons for its success, its range of validity, and the significance of its failures, in an entirely different light. Any successes that the theory has make it useful in an engineering sense, as an instrument for prediction. But any failures which we might find would be far more valuable to us, because they would disclose new laws of physics. You can't lose either way!

The transition from classical to quantum statistical mechanics provided some very good examples of this. Classical statistical mechanics made definite predictions that certain things should happen. It was found to be an experimentally reproducible fact that these things did not happen. Therefore, it follows that the enumeration of possible states on which the classical statistical mechanics was based was not correct. We change our enumeration in these respects: we introduce discrete states; we recognize that permutations of identical particles do not produce new physical states; and, we introduce the natural unit of volume of phase space h^3 per particle. When we make these changes in our method of enumerating the possibilities, then it is found that we do get predictions in agreement with experiment.

It doesn't follow from this that our enumeration is now correct. There might be enormously great constraints on the possible quantum states which we haven't any inkling of yet. It might be that instead of $10^{10^{20}}$ states, there are only 1 in $10^{10^{19}}$ of those actually accessible at all to the system, because of new laws of physics we haven't discovered yet. The fact that our present statistical mechanics works doesn't exclude that possibility at all. Though we have no evidence for that possibility until we find a case where it doesn't work.

Whenever we have a situation that is experimentally reproducible; i.e., whenever macroscopic information is sufficient to predict macroscopic phenomena, then it is always possible in principle to build a phenomenological theory which side-steps all microscopic details and describes only

relations between macroscopic things. Thermodynamics, as we know it, is one special case of that. Looked at this way, we see that there's every reason to believe that an irreversible thermodynamics can be developed which will be just as general and have just as many nice formal properties as our present equilibrium thermodynamics. During the past 10 years, quite a bit of progress has been made along this line, but we are still very far from the goal. At present, we have the Onsager reciprocity laws, the principle of minimum entropy production, and the fluctuation-dissipation theorem, which are all related to each other in ways not yet entirely clear. We know very little about the range of validity of these rules. However, the situation here looks so promising that, if I were to pose as a prophet, I would say that the next 20 years will see developments in this field comparable to those in the middle nineteenth century which gave us our present equilibrium thermodynamics.

LECTURE FIVE

THE A_p DISTRIBUTION

Memory Storage for Old Robots

We have given our robot another principle by which he can convert information into numerical values of probabilities, and he is now able to solve lots of problems; but he still operates in a rather inefficient way in one respect. When we given 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.

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 intelligent 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 about 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 what the effect of getting new information would be. 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 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. A few years ago, I even gave a talk at one of the Berkeley Statistical Symposiums, expounding this viewpoint. This is, furthermore, just what Carnap* 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,

---

* R. Carnap, "The Continuum of Inductive Methods," University of Chicago Press, 1952.

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

$$(A|A_pE) \equiv p \tag{104}$$

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 proposition 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 (actually, "nothing" has a very precise meaning which we'll see later), then

$$(A_p|X) = 1, \quad 0 \leqslant p \leqslant 1 . \tag{105}$$

This is a probability density, since p is continuously variable. If X tells us nothing relevant to A, then the distribution of maximum entropy is the uniform one. 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,

$$(A_p|E) = (A_p|X)\frac{(E|A_p)}{(E|X)} = \frac{(E|A_p)}{(E|X)} . \tag{106}$$

Now,

$$(A|E) = \int_0^1 (AA_p|E)\,dp. \tag{107}$$

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 a 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 associate plausibility with 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 meaning you or I might attach to a proposition. So now we can blow up the integrand of (107) by our Rule 1:

$$(A|E) = \int_0^1 (A|A_p E)(A_p|E)\,dp. \tag{108}$$

But from the definition (104) of A_p, the first factor is just p, and so

$$(A|E) = \int_0^1 p\,(A_p|E)\,dp. \tag{109}$$

The probability which our robot assigns to proposition A is just the first moment 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; $E = E_a E_b$, where E_a is relevant to A and, given E_a, E_b is not relevant:

$$(A|E) = (A|E_a E_b) = (A|E_a). \tag{110}$$

By Bayes' theorem, it follows that, given E_a, A must also be irrelevant to E_b, for

Jaynes' A_p – Distribution

Define the proposition

$A(p) \equiv$ "Regardless of anything else you may have been told, the probability of A is p."

Hence, $P(A \mid A(p)\, E) = p$,

where E is any other evidence. By integrating Bayes' theorem,

$$P(A \mid E) = \int_0^1 p\, P(A(p) \mid E)\, dp ;$$

Since p is continuous, $P(A(p) \mid E)$ is a probability density.

Auxiliary condition: If X says "nothing" about A, then

$$P(A(p) \mid X) = 1, \quad 0 \le p \le 1.$$

? By "nothing" we mean "no preference"; so if $E = X$, a X means "nothing", then $P(A \mid X) = 1/2$. Note that $P(A \mid E)$ is the first moment of $P(A(p) \mid E)$. Here, of "nothing" does nothing to change one's state of mind regarding $P(A \mid E)$.

If $P(A(p) \mid X) = 1$, we say that X means there no other prior information except that A must be either true or false; there are no other prior hypotheses — just as with the underlying hypothesis of Bernoulli sequences. Certainly, X must be relevant to A!

$$(E_b|AE_a) = (E_b|E_a)\frac{(A|E_bE_a)}{(A|E_a)} = (E_b|E_a). \qquad (111)$$

Let's call this property "weak irrelevance." Now does this imply that E_b is irrelevant to A_p? Evidently not, for (110) 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, (110) holds independently of what E_a might be; call this "strong irrelevance." Then we have

$$(A|E) = \int_0^1 p(A_p|E_aE_b)\,dp = \int_0^1 p(A_p|E_a)\,dp\,. \qquad (112)$$

If this is to hold for all $(A_p|E_a)$, the integrands must be the same:

$$(A_p|E_aE_b) = (A_p|E_a) \qquad (113)$$

and from Bayes' theorem it follows as in (111) that A_p is irrelevant to E_b:

$$(E_b|A_pE_a) = (E_b|E_a) \qquad (114)$$

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,

$$(A|EF) = \int_0^1 p(A_p|EF)\,dp = \int_0^1 p(A_p|E)\frac{(F|A_pE)}{(F|E)}\,dp. \quad (115)$$

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

$$\frac{(F|A_pE_aE_b)}{(F|E_aE_b)} = \frac{(F|A_pE_a)\left[\frac{(E_b|FA_pE_a)}{(E_b|A_pE_a)}\right]}{(F|E_a)\left[\frac{(E_b|FE_a)}{(E_b|E_a)}\right]} = \frac{(F|A_pE_a)}{(F|E_a)} \quad (116)$$

where we have used (114). 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 (104) 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 (115) reduced to

$$(A|EF) = \frac{1}{(F|E_{aa})}\int_0^1 p(A_p|E)(F|A_p)\,dp. \quad (117)$$

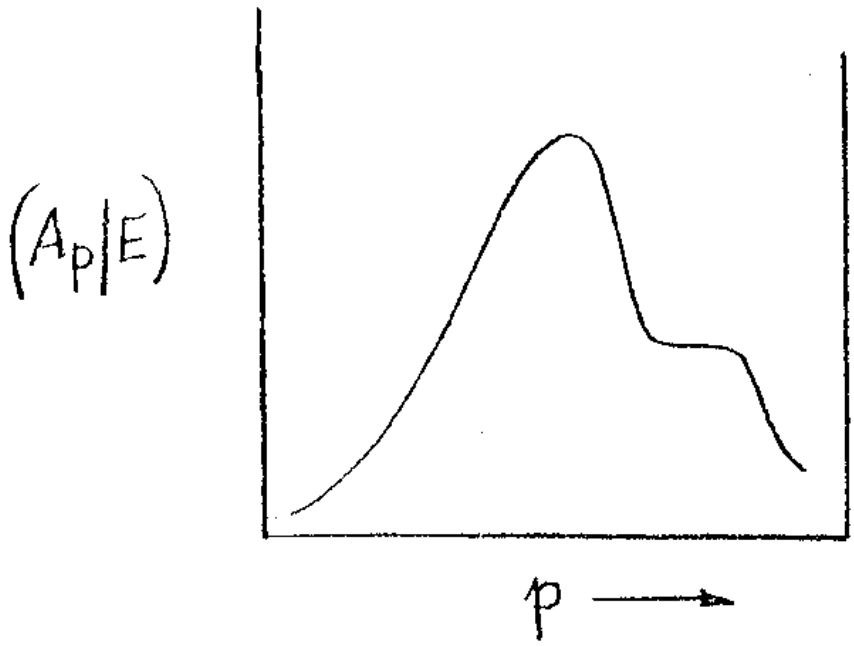
The weak point in this argument is that I haven't proved that it is possible to resolve E into a completely relevant part and completely irrelevant part. However, we'll see in a minute that in many important applications it is possible. So, let's just say that the following results

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 (21); by calculating the odds on A instead of the probability. This is just

$$\frac{(A|EF)}{(a|FE)} = \frac{\int_0^1 p(A_p|E)(F|A_p)\,dp}{\int_0^1 (1-p)(A_p|E)(F|A_p)\,dp} = O(A|EF), \quad (118)$$

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 about 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 proposition about which he is going to have to reason, he can store a function like this:



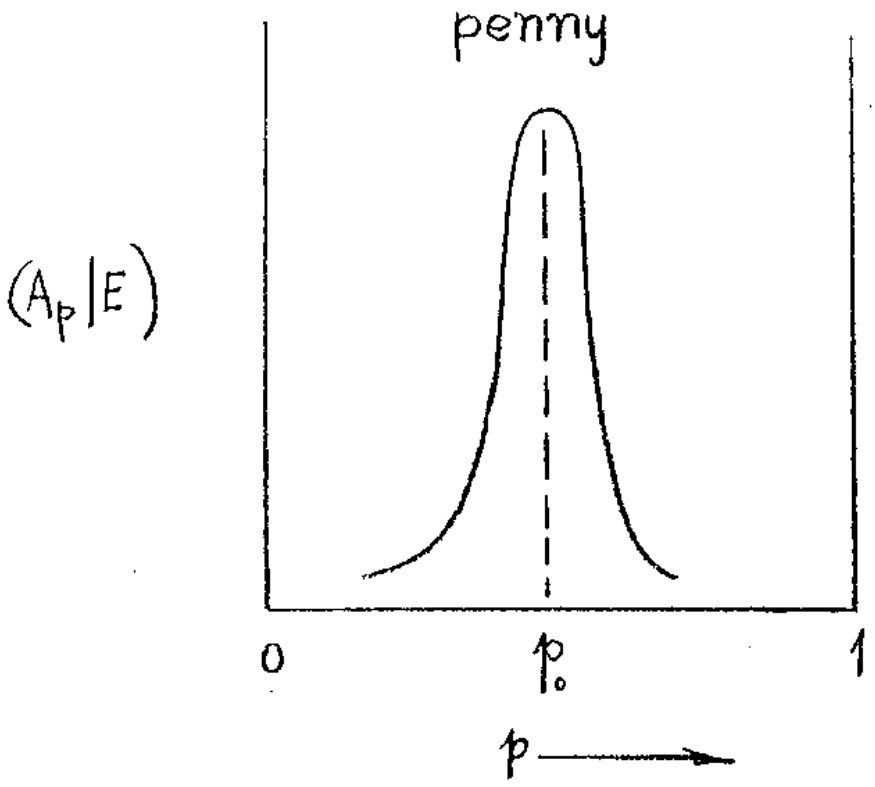
Whenever he receives new information, F, he will be well advised to calculate $(A_p|EF)$, and he then can erase his previous $(A_p|E)$, and for the future store only $(A_p|EF)$. This shows that in a machine which does inductive reasoning, the memory storage problem is very much simpler than it is in a machine which does 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 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

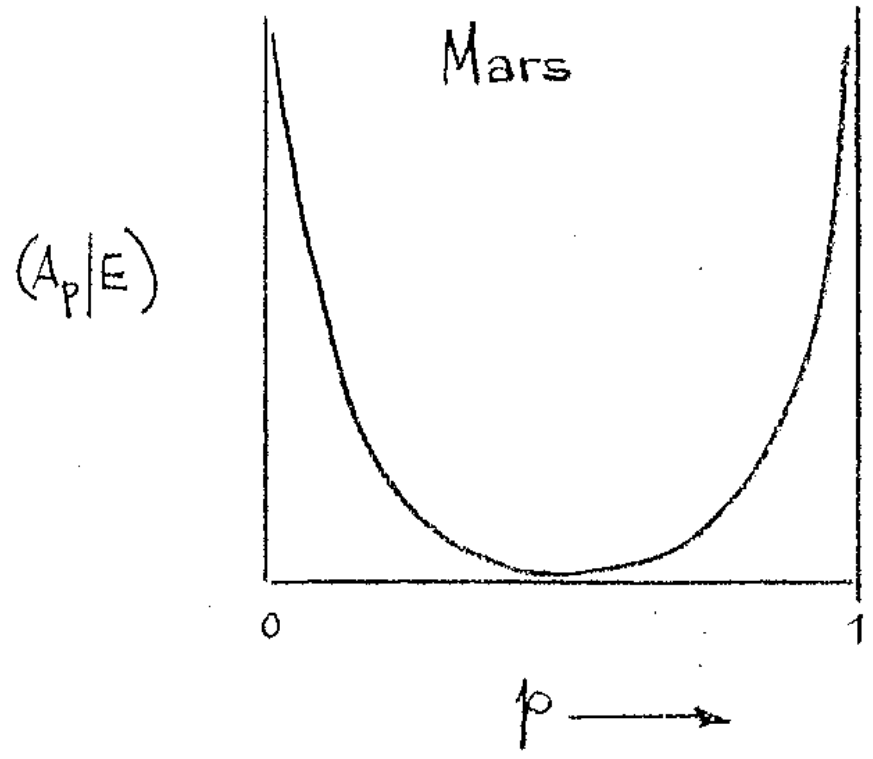
$$\frac{d \sin x}{dx} = \cos x \quad ?$$

Let's look once more at Equation (117). 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 correctly 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 this.



In the case of the question of intelligent life on Mars, my state of knowledge is described by an $(A_p|E)$ distribution something like this, qualitatively.



The first moment is exactly the same 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 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."

Laplace's Law of Succession

Now, let's imagine that a random experiment is being performed. From the results of the random 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:

$X \equiv$ "We admit two prior hypotheses; A true, and A false. 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"

A more precise statement of X is

$$(A_p|X) = 1 \qquad 0 \leq p \leq 1 \ . \tag{119}$$

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 (35), we have the binomial distributions

$$(N_n|A_p) = \binom{N}{n} p^n (1-p)^{N-n} \ , \tag{120}$$

$$(M_m|A_p) = \binom{M}{m} p^m (1-p)^{M-m} \ .$$

I might mention here 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 textbooks. One postulates that an event possesses some intrinsic, "absolute" probability, whose numerical value we can never determine exactly. Nevertheless, no one questions that such an "absolute" probability exists. Cramér*, 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 side 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's 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 we'll see several more examples presently.

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

---

* H. Cramér, "Mathematical Methods of Statistics," (Princeton Press, 1946); p 154.

into mutually exclusive alternatives gives us

$$(N_n|X) = \int_0^1 (N_n A_p|X)\,dp = \int_0^1 (N_n|A_p)(A_p|X)\,dp = \binom{N}{n} \int_0^1 p^n (1-p)^{N-n}\,dp.$$

The integral we have to evaluate is of the form

$$\int_0^1 x^r (1-x)^s\,dx = \frac{r!\,s!}{(r+s+1)!} \tag{121}$$

which is known as an Eulerian integral of the first kind. Thus, we have

$$(N_n|X) = \left\{ \begin{array}{ll} \frac{1}{N+1}, & 0 \leq n \leq N \\ 0, & N < n \end{array} \right\}; \tag{122}$$

i.e., just the uniform distribution of maximum entropy. $(M_m|X)$ is similarly found. Now we can turn (120) around by Bayes' theorem:

$$(A_p|N_n) = (A_p|X)\frac{(N_n|A_p)}{(N_n|X)} = (N+1)(N_n|A_p) \tag{123}$$

and so finally the desired probability is

$$(M_m|N_n) = \int_0^1 (M_m A_p|N_n)\,dp = \int_0^1 (M_m|A_p N_n)(A_p|N_n)\,dp. \tag{124}$$

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 (124), we have again an

Eulerian integral, and our result is

$$(M_m | N_n) = \frac{\binom{n+m}{n}\binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}}. \qquad (125)$$

This is a very old and well-known formula in probability theory. Let's look at it 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

$$(A | N_n) = \frac{n+1}{N+2}. \qquad (126)$$

This is Laplace's law of succession. 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 law 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 law of succession is, next to Bayes' theorem, the second most important rule 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.

Poor old Laplace has been lampooned for generations because he illustrated the 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 law of succession.

I have to confess to you that I am unable to see anything at all absurd about the law 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 law 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 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 law 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.

---

* P. S. Laplace, "A Philosophical Essay on Probabilities," (Dover, 1951); p 19.

Let me give you just two examples, both famous, of the kind of objections to the law of succession which you find in the literature. (1) Suppose the solidification of hydrogen to have been once accomplished. According to the law 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) Consider the law of succession in 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 law says also that there is probability 1/2 that there are no elephants on Mars. The law is self-contradictory!

The trouble with example (1) is obvious in view of our earlier remarks. But let's look a little more closely at example (2). 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 law of succession is going to survive example (2), 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 law of succession was that the robot perceives 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 application of insufficient reason, and this will of course give the same answer $(A|X) = \frac{1}{2}$, that we got from the law 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 law of succession.

If the robot perceives 137 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 a gross misapplication of the law 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."

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 was 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 law 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, 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, $\{A_1, A_2 \ldots, A_K\}$, 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? By a perfectly straightforward generalization of the derivation of (125), we find

$$(m_1 \cdots m_K | n_1 \cdots n_K) = \frac{\binom{n_1+m_1}{n_1} \cdots \binom{n_K+m_K}{n_K}}{\binom{N+M+K-1}{M}} . \qquad (127)$$

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:

$$(A_1 | n_1, N, K) = \frac{n_1 + 1}{N + K} , \qquad (128)$$

You see that in the case $N = n_1 = 0$, this reduces to the answer provided by insufficient reason. In the case that 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 law 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 to help us in deciding between the hypotheses, 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 a minute that this conclusion also follows as a consequence of the theory we're developing here.

The real reason for introducing the law of succession lies in the cases where we do 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 where we have a K-dimensional initial assignment $(A_{p_1\cdots p_K}|X)$, uniform in the case that leads to (128). You see from (128) that as soon as the number of observations is large compared to the number of hypotheses, then the probability assigned to any particular hypothesis depends, for all practical purposes, only on what we have observed, not on how many prior hypotheses there were. If you contemplate this for ten seconds, I think your common sense will tell you that the criterion, $N >> K$, is exactly the right one for this to be so.

Probability and Frequency

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.

The law of succession gives us the solution to problem (a); 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, then 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 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. Even the +1 and +2 in Laplace's formula turn up, in a sense, 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 law of succession without thereby damning his own procedure; he is doing exactly what Laplace's law of succession tells him to do. To define probability in terms of frequency is equivalent to saying that the law of succession is the only rule which can be used for converting observational data into probability assignments.

Now let's consider problem (b); 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

---

* H. Cramér, "Mathematical Methods of Statistics," (Princeton University Press, 1946); pp 509-524. See especially equation (34.2.5) at the 84% confidence level, corresponding to $\lambda = \sqrt{2}$.

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 (125) as $M \longrightarrow \infty$, $m \longrightarrow \infty$, in such a way that $(m/M) \longrightarrow f$. Introducing the proposition

$A_f \equiv$ "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

$$(A_f|N_n) = \frac{(N+1)!}{n!\,(N-n)!} f^n (1-f)^{N-n}, \qquad (129)$$

which is the same as our $(A_p|N_n)$ in (123), with f numerically equal to p. According to (129), the most probable frequency is equal to (n/N), the observed frequency in the past. But we have noted before that in parameter estimation 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

$$\bar{f} = \int_0^1 f(A_f|N_n)\,df = \frac{n+1}{N+2}; \qquad (130)$$

i.e., just the same as the value of $(A|N_n)$ given by Laplace's law 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 argument.

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 (129) 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 (129) is

$$\mathrm{var}\,(A_f|N_n) = \overline{f^2} - \overline{f}^2 = \overline{f}(1-\overline{f})/(N+3),$$

so that the expected error in the estimate (130) decreases like $N^{-\frac{1}{2}}$. More detailed conclusions about the reliability of predictions, which we could make from (129), are for all practical purposes identical with those the statistician would make by the method of confience intervals.

All these results hold also for the generalized law of succession. Taking the limit of (127) as $M \longrightarrow \infty$, $(m_i/M) \longrightarrow f_i$, we find the joint probability distribution for A_i to occur with frequency f_i to be

$$(f_1 \cdots f_k | n_1 \cdots n_k)\, df_1 \cdots df_k = \frac{(N+k-1)!}{n_1! \cdots n_k!} f_1^{n_1} \cdots \qquad (131)$$

$$\cdots f_k^{n_k}\, df_1 \cdots df_k$$

where the df_i are restricted by the condition $\sum_i df_i = 0$. The probability that the frequency f_1 will be in the range df_1 is found by integrating (131) over all values of $f_2 \cdots f_k$ compatible with $f_i \geq 0$, $(f_2 + \cdots + f_k) = 1 - f_1$. This can be carried out by application of Laplace transforms in a well-known way, and the result is

$$(f_1|n_1 \cdots n_k)\,df_1 = \frac{(N+k-1)!}{n_1!\,(N-n_1+k-2)!} f_1^{n_1} (1-f_1)^{N-n_1+k-2}\, df_1 \qquad (132)$$

from which we find the most probable and mean value estimates of f_1 to be

$$(f_1)_{m.p.} = \frac{n_1}{N+k-2} \qquad (133)$$

$$\overline{f_1} = \frac{n_1+1}{N+k}, \quad \text{compare (128)} \qquad (134)$$

Another interesting result is found by taking the limite of (120) as $M \longrightarrow \infty$, $(m/M) \longrightarrow f$. We easily find

$$(A_f|A_p) = \delta(f-p). \qquad (135)$$

Likewise, taking the limit of (123) as $N \longrightarrow \infty$, we find

$$(A_p|A_f) = \delta(p-f), \qquad (136)$$

which follows also from (135) by application of Bayes' theorem. Therefore, if B is any proposition, we have from our standard argument,

$$(B|A_f) = \int_0^1 (BA_p|A_f)\,dp = \int_0^1 (B|A_pA_f)(A_p|A_f)\,dp = \int_0^1 (B|A_p)\delta(p-f)\,dp \,. \quad (137)$$

In the last step we used the property (104) 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 (125) reduces to the binomial distribution $(M_m|A_p)$ as given by (120). The generalized formula (127), in the corresponding limit, goes into the well-known multinomial distribution.

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, Laplace's theory becomes mathematically identical with 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. Examples are the statistical mechanics of irreversible processes, and the theory of radar detection. The axiom (probability) $\equiv$ (frequency), if applied consistently, would prevent us from using probability theory in these problems.

This is, I think, the same thing that Professor Kac pointed out here - that the question of how one can introduce probability methods into physics involves great conceptual difficulties. These difficulties, I suggest, are due only to the attempt to interpret every probability as a frequency. If we admit, with Laplace, that the notion of probability is a respectable concept in its own right, then there is nothing mysterious about a probability distribution in both position and velocity, even though there is no "lack of specification" over which we can average. In the "master equation" approach to kinetic theory, it is meaningless to ask whether or why nature prepares the factorized distributions which lead to the Boltzmann equation. Nature does not prepare distributions, factorized or otherwise; she prepares states. There does not exist any 1:1 correspondence between different probability distributions and different physical situations.

I don't think the present mysteries of kinetic theory are going to be cleared up until workers in the field recognize this, and reformulate the objectives of the theory. For example, the problem which is relevant to physics is not to calculate the "true" contracted distribution functions, or to find what precise mathematical properties of the master probability function lead to the Boltzmann equation. It is to find what physical

predictions are characteristic of "practically all" master probability functions compatible with our macroscopic information. Experimentally reproducible effects can involve only such predictions, and as soon as we learn how to extract them from the master probability function, then it will make very little difference which particular function we use in our calculations.

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

$$(A|EF) = \int_0^1 p(A_p|EF)\,dp = \frac{\int_0^1 p(A_p|F)(A_p|E)\,dp}{\int_0^1 (A_p|F)(A_p|E)\,dp}$$

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 appreciate change in the probability of A; i.e.,

$$(A|EF) \cong (A|E).$$

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, e.g. (123), gives

$$(A_p|N_n) = \frac{(N+1)!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$\cong (\text{constant}) \cdot \exp\left[-\frac{(p-f)^2}{2\sigma^2}\right] \tag{138}$$

where

$$\sigma^2 = \frac{f(1-f)}{N} .$$

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.

$$(A_p|EF) = (\text{constant})(A_p|E)(A_p|F). \tag{139}$$

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 insufficient reason that we started with, and also make contact between this theory and Carnap's methods of inductive reasoning.

Before we can use insufficient reason 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 describe 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 <u>does</u> make a difference. It makes a very important difference, but one that doesn't show up until we start performing this random experiment. It is not in the probability of A, the difference is 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.

Carnap's Inductive Methods

Carnap* 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

* R. Carnap, "The Continuum of Inductive Methods," Univ. of Chicago Press, 1952.

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 re-interpret 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 $(A_p|X) = (\text{const.}) \cdot \left[p(1-p)\right]^r$, with $2r = \lambda - 2$. The result is

$$(A|N_n X) = \frac{2n + \lambda}{2N + 2\lambda} = \frac{(n + r) + 1}{(N + 2r) + 2}, \qquad (140)$$

Greater λ thus corresponds to a more sharply peaked $(A_p|X)$ distribution.

In our coin-tossing example, the gentlemen from the Bureau of Standards reason 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 law 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.)

From the second way I wrote Equation (140), 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 that, with prior knowledge, it is <u>possible</u> for A to be either true or false. It 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.

Our theory also gives "inductive methods" for more general prior distributions for which $(A|X) \neq 1/2$. For any of these we find, in agreement with Carnap, that the probability assigned to A in a single trial is numerically equal to the mean value estimate of the frequency of A in a large number of trials.

NOTE: At this colloquium, Professor Jaynes delivered ten lectures on the subject of probability. Circumstances beyond our control prevented the transcription of the entire lecture series.

SOCONY MOBIL OIL COMPANY, INC.

FIELD RESEARCH LABORATORY

COLLOQUIUM LECTURES IN PURE AND APPLIED SCIENCE

1. "Kinetic Theory Applied to Hydrodynamics", Professor Max Dresden, Northwestern University, June, 1956.

2. "Some Stochastic Problems in Physics and Mathematics", Professor Mark Kac, Cornell University, October, 1956.

3. "Statistical Dynamics", Professor Henry Eyring, The University of Utah, June, 1957.

4. "Probability Theory in Science and Engineering", Professor E. T. Jaynes, Stanford University, February, 1958.