Lecture 13

INTRODUCTION TO DECISION THEORY

"Your act was unwise," I exclaimed "as you see
by the outcome." He solemnly eyed me.
"When choosing the course of my action," said he,
"I had not the outcome to guide me."

----Ambrose Bierce

At this stage we have accumulated quite a few loose ends, which I would now like to clear up. In almost every lecture so far, I had to insert one or more parenthetic remarks to the effect that "there is still an essential point missing here, which will be supplied when we take up decision theory." Actually, we began seeing what it is, as soon as we started applying the theory to our first problem. When we illustrated the use of Bayes' theorem by sequential testing in Lecture 5, we noted that there is nothing in probability theory *per se* which could tell us where to put the threshold levels at which we make our decision: whether to accept the batch, reject it, or make another test. At that time, I said only that the location of this threshold level obviously depends in some way on our judgment as to what are the consequences of making wrong decisions, and what are the costs of making further tests. Qualitatively, this is clear enough; but before we can claim to have a really complete design for our robot, we must re-state this in quantitative terms.

The same situation occurred in Lecture 8 when we studied particle counters, and the robot was faced with the job of estimating the number of particles which had passed through the counter under various conditions. Probability theory told us only the robot's state of knowledge as to the number of particles; it did not tell us what estimate he should in fact make. We noted at that time that taking the mean value over the posterior distribution was the same as making that decision which minimizes the expected square of the error; and in Lecture 11 we followed the same procedure for statistical mechanics. In both of those cases, this seems to be a fairly sensible criterion, and leads to results in good correspondence with common sense. However, why was it the square of the error that we minimized? Why not some other function of the error? The criterion of minimum mean square error has obvious mathematical advantages, because the mean value of a distribution is generally easy to calculate; but in principle it appears to be entirely arbitrary.

You see the common feature of all these problems. In every case, probability theory can give us only a probability distribution which represents the robot's final state of knowledge with all the available data taken into account; but in practice his job is to make a definite decision. He must act as though one hypothesis were true, he must make a definite numerical estimate of some parameter, and so on. The essential thing which is still missing in our design of this robot is the rule by which he converts his final probability assignment into a definite course of action.

13.1. Daniel Bernoulli's Suggestion.

As you might expect from the way this situation appeared in the most elementary applications of probability theory, this problem is by no means new. It was clearly recognized, and a definite solution offered for a certain class of problems, by Daniel Bernoulli in the year 1738. In a cruder form, the

same principle had been seen even earlier, at the time when probability theory was concerned almost exclusively with problems of gambling. The notion which seemed very intuitive to the first workers in probability theory was "expectation of profit." By this we mean, of course, that I consider each possibility, $i = 1, 2, \ldots, n$, assign probabilities p_i to them, and also assign numbers M_i which represent the profit I would obtain if the i'th possibility should in fact turn out to be true. Then the quantity

$$\langle M \rangle = \sum_{i=1}^{n} p_i M_i \qquad (13\text{-}1)$$

is what we call the "expectation of profit." It seemed obvious to the first workers in probability theory that a gambler acting in pure self-interest should always behave in such a way as to maximize his expected profit. This, however, led to some paradoxes (particularly in the famous St. Petersburg problem) which led Bernoulli to recognize that simple expectation of profit is not always a sensible criterion of action.

To give a very simple example, suppose that I assign probability 0.51 to heads in a certain slightly biased coin. Now I am given the choice of two actions: (1) to bet every cent I have at even money, on heads for the next toss of this coin; (2) not to bet at all. According to the criterion of expectation of profit, I should always choose to gamble when faced with this choice. My expectation of profit, if I do not gamble, is zero; but if I do gamble, it is

$$\langle M \rangle = 0.51\, M_o + 0.49\, (-M_o) = 0.02\, M_o > 0 \qquad (13\text{-}2)$$

where M_o is the amount I have now. Nevertheless it seemed obvious to Bernoulli, and I think it does also to you, that very few people would really choose the first alternative in the problem as stated. This means that our common sense, in some cases, rejects the criterion of maximizing expected profit.

Suppose that you are offered the following opportunity. You can bet any

amount you want on the basis that, with probability $(1 - 10^{-6})$, you will lose your money; but with probability 10^{-6}, you will win 1,000,001 times the amount you had wagered. Again, the criterion of maximizing expected profit says that you should bet all the money you have. Our common sense rejects this solution even more forcefully; no sane person would risk all his fortune, which he is practically certain to lose, for an infinitesimal chance of winning a very much larger sum.

Daniel Bernoulli proposed to resolve these paradoxes by recognition that the true value to a person, of receiving a certain amount of money, is not measured simply by the amount received; it depends also upon how much he has already. In other words, Bernoulli said that we should recognize that the mathematical expectation of profit is not the same thing as its "moral expectation." A modern economist is expressing exactly the same idea when he speaks of the "diminishing marginal utility of money."

The original St. Petersburg game consists of the following--we toss an honest coin until it comes up heads for the first time. The game is then terminated. If heads occurs for the first time at the n'th throw, the player receives 2^n dollars. The question is: what is a "fair" entrance fee for him to pay, for the privilege of playing this game? If we use the criterion that a fair game is one where the entrance fee is equal to the expectation of profit, you see what happens. This expectation is

$$\sum_{K=1}^{\infty} (2^{-K})(2^{K}) = \sum_{K=1}^{\infty} 1 \qquad (13\text{-}3)$$

and this is infinite. Nevertheless it is clear again that no sane person would be willing to risk more than a very small part of his fortune for the privilege of playing this game. Let me quote Laplace (1819) at this point:

"Indeed, it is apparent that one franc has much greater value for him who possesses only 100 than for a millionaire. We ought then to distinguish

the absolute value of the hoped-for benefit from its relative value. The latter is regulated by the motives which make it desirable, whereas the first is independent of them. The general principle for appreciating this relative value cannot be given, but here is one proposed by Daniel Bernoulli which will serve in many cases: The relative value of an infinitely small sum is equal to its absolute value divided by the total fortune of the person interested."

In other words, Bernoulli proposed that the "moral value," or what the modern economist would call the "utility" of money should be taken proportional to its logarithm. Laplace, in discussing the St. Petersburg problem and this criterion, reports the following result without giving the calculation: a person whose total fortune is 200 francs ought not reasonably to stake more than 9 francs on the play of this game. I took the trouble of checking this. The fair fee f(200) is found by equating his present utility with his expected utility if he pays the fee and plays the game; a computer gives the root of

$$\log 200 = \sum_{n=1}^{\infty} \frac{1}{2^n} \log(200 - f + 2^n)$$

as f(200) = 8.7204. Likewise, $f(10^3) = 10.98$, $f(10^4) = 14.24$ $f(10^6) = 20.87$.

It seems to me that this kind of numerical result is entirely reasonable. However the logarithmic assignment of utility is not to be taken literally either in the case of extremely small fortunes (as Laplace points out), or in the case of extremely large ones, as the following example of Savage (1954) shows. Suppose your total fortune is 10,000,000 dollars; then if your utility for money is proportional to the logarithm of the amount, the theory says that you should be as willing as not to accept a wager in which, with probability one-half, you'll be left with only 10,000 dollars; and with probability one-half, you will be left with 10,000,000,000 dollars. I think that most of us would consider such a bet to be distinctly disadvantageous to a person with that initial fortune. This shows that our intuitive "utility" for money actually must increase even less rapidly than the logarithm for extremely

large values. There are some who even claim that it is bounded.

The gist of Daniel Bernoulli's suggestion was therefore that, in the gambler's problem of decision making under uncertainty, one should act so as to maximize the expected value, not necessarily of the profit itself, but of some function of the profit which he called the "moral value". In more modern terminology the optimist will call this "maximizing expected utility," while the pessimist will speak instead of "minimizing expected loss", the loss function being taken as the negative of the utility function.

The logarithmic assignment of utility is reasonable for many purposes, as long as it is not pushed to extremes. It is also, incidentally, very closely connected with the notion of entropy, as shown by an argument of Kelly (1956), extended by Bellman and Kalaba (1956). Here, a gambler who receives advance tips on a game which are only partly reliable, acts (i.e., decides on which side and how much to bet) so as to maximize the expected logarithm of his fortune. They show that (1) one can never go broke following this strategy, in contrast to the strategy of maximizing expected profit, where it is easily seen that with probability one this will eventually happen, and (2) the amount one can reasonably expect to win on any one game is clearly proportional to the amount M_o he has to begin with, so after n games, one could hope to have an amount $M = M_o e^{\alpha n}$. With the logarithmic utility function, one acts so as to maximize the expected value of α. The maximum attainable $\langle\alpha\rangle$ turns out to be just $(S_o - S)$, where S is the entropy which describes the gambler's uncertainty as to the truth of his tips, and S_o is the maximum possible entropy, if the tips were completely unreliable. This suggests that, with a little more development of the theory, entropy might have an important place in guiding the strategy of a stock market investor.

Daniel Bernoulli's solution to the problem of decision making has suffered the same fate as did Laplace's solution to the problem of inductive reasoning.

The "objectivist" or "orthodox" school of thought either ignored it or condemned it as metaphysical nonsense until just a few years ago. In one of the best known books on probability theory (Feller, 1950; p. 199), Daniel Bernoulli's solution of the St. Petersburg paradox is rejected without even being described, except to assure the reader that he "tried in vain to solve it by the concept of moral expectation." Well, we will see next just how vain Daniel Bernoulli's efforts were.

13.2. The New Formulation of the Decision Problem.

In the late 1940's a general theory of decision making in the face of uncertainty was developed, largely by Wald (1950) which in its initial stages had no apparent connection with probability theory. I mentioned it briefly in Lecture 5, and now I would like to give you a more specific account of some of the ideas it involved.

We begin by imagining (i.e. enumerating) a set of possible unknown "states of nature", $\{\theta_1, \theta_2, \ldots, \theta_N\}$ whose number might be finite or infinite. The θ_j might also form a continuum. In the quality-control example of Lecture 5, the "state of nature" is the unknown number of defectives in the batch, and the θ_j are discrete. In the particle-counter problem of Lecture 8, the state of nature could be taken as the unknown source strength s, and the θ_j are continuous.

There are certain illusions that tend to grow and propagate here. Let me dispel one right now by noting that, in enumerating the different states of nature, we are not describing any objective (measurable) property of nature --for, one and only one of them is in fact true. The enumeration is only a means of describing our state of ignorance. It is, therefore, meaningless to ask whether one particular enumeration in "correct" without first asking, "what is the information that is being described by the set of θ_j?" Two

observers with different amounts of information may enumerate θ_j differently without either being inconsistent.

The next step in our theory is to make a similar enumeration of the possible decisions $\{D_1, D_2, \ldots, D_k\}$ that might be made. In the quality-control example, there were three possible decisions at each stage:

$$D_1 = \text{accept the batch}$$
$$D_2 = \text{reject it}$$
$$D_3 = \text{make another test} \qquad (13\text{-}4)$$

In the particle counter problem of Mr. B, where we are to estimate the number n_1 of particles passing through the counter in the first second, there are an infinite number of possible decisions:

$$D_i = \text{"}n_1 \text{ is estimated as equal to } i\text{,"} \quad i = 0, 1, 2, \ldots \qquad (13\text{-}5)$$

If we are to estimate the source strength, then there is a continuum of possible decisions.

This theory is clearly of no use unless by "making a decision" we mean "deciding to act as if the decision were correct". It is idle to "decide" that $n_1 = 150$ is the best estimate unless we are then prepared to act on the assumption that $n_1 = 150$. Thus the enumeration of the D_i is a means of describing our knowledge as to what kinds of actions are <u>feasible</u>; it is idle to consider any decision which we know in advance corresponds to an impossible course of action.

There is another reason why a particular decision might be eliminated; even though D_1 is easy to carry out, we might know in advance that it would lead to intolerable consequences. An automobile driver can make a sharp left turn at any time; but his common sense usually tells him not to. Here we see two more points: (1) there is a continuous gradation--the consequences of an action might be serious without being absolutely intolerable, and (2) the consequences of an action (=decision) will in general depend on what is the

true state of nature--a sharp left turn does not always lead to disaster.

This suggests a third concept we need--the loss function $L(D_i,\theta_j)$, which is a set of numbers representing our judgment as to the "loss" incurred by making decision D_i if θ_j should turn out to be the true state of nature. If the D_i and θ_j are both discrete, this becomes a loss matrix L_{ij}.

Quite a bit can be done with just the θ_j, D_i, L_{ij} and there is a rather extensive literature dealing with criteria for making decisions with no more than this. The material we need for our purposes has been summarized in a very readable and entertaining form by Luce and Raiffa (1957), and in the elementary textbook of Chernoff and Moses (1959). The minimax criterion is this: for each D_i find the maximum possible loss $M_i = \max_j(L_{ij})$; then choose that D_i for which M_i is a minimum. The minimax criterion would be a reasonable one if we regard nature as an intelligent adversary who foresees our decision and deliberately chooses the state of nature so as to cause us the maximum frustration. In the theory of some games, this is not a completely unrealistic way of describing the situation, and consequently minimax strategies are of fundamental importance in game theory. But in the decision problems of the scientist or engineer the minimax criterion is that of the long-faced pessimist who concentrates all his attention on the worst possible thing that could happen, and thereby misses out on the favorable opportunities.

Equally unreasonable for us is the opposite extreme of the starry-eyed optimist who uses this "minimin" criterion: for each D_i find the minimum possible loss $m_i \equiv \min_j(L_{ij})$ and choose the D_i that makes m_i a minimum.

Evidently, a reasonable decision criterion for the scientist and engineer is, in some sense, intermediate between minimax and minimin. Many other criteria have been suggested, which go by the names of maximum utility (Wald), α-optimism-pessimism (Hurwicz), minimax regret (Savage), etc. The usual procedure, as described in detail by Luce and Raiffa, has been to analyze any

proposed criterion to see whether it satisfies about a dozen qualitative common-sense conditions such as (1) Transitivity: if D_1 is preferred to D_2, and D_2 preferred to D_3, then D_1 must be preferred to D_3, and (2) Strong Domination: if for all states of nature θ_j we have $L_{ij} < L_{kj}$, then D_i should always be preferred to D_k. This analysis, although straightforward, can become tedious. I will not follow it any further, because the final result is that there is only one class of decision criteria which passes all the tests, and this class is obtained more easily by a different line of reasoning.

A full decision theory, of course, cannot concern itself merely with the θ_j, D_i, L_{ij}. We also, in typical problems, have additional evidence E, which we recognize as relevant to the decision problem, and we have to learn how to incorporate E into the theory. In the quality-control example, E consisted of the results of the previous tests.

At this point, current decision theory takes a long, and I think unnecessary, mathematical detour. One defines a "strategy", which is a set of rules of the form, "If I receive new evidence E_i, then I will make decision D_k." In principle one first enumerates all conceivable strategies (whose number is, however, astronomical even in quite simple problems), and then tries to eliminate the undesirable ones by application of various common-sense conditions. This leads to defining a class of "admissible" strategies, which consists, crudely speaking, of all those any sane person would ever consider adopting; a strategy is admissible if no other exists which is as good or better for all states of nature.

A principal object of the theory is then to characterize the class of admissible strategies in mathematical terms, so that any such strategy can be found by carrying out a definite procedure. The fundamental theorem bearing on this is Wald's Complete Class Theorem which establishes a result already mentioned in Lecture 5. Instead of following this rather difficult argument,

I would like to make a few more remarks about the nature of the problem, and then give a different line of reasoning which leads to the same result by elementary mathematics.

What is it that makes a decision process difficult? Well, if we knew which state of nature was the correct one, there would be no problem at all; if θ_3 is the true state of nature, then the best decision D_i is the one which renders L_{i3} a minimum. In other words, once the loss function has been specified, our uncertainty as to the best decision arises solely from our uncertainty as to the state of nature. Whether the decision minimizing L_{i3} is or is not best depends entirely on this: How strongly do we believe that θ_3 is the true state of nature? How plausible is θ_3?

To a physicist or engineer it seems like a very small step--really only a rephrasing of the question--to ask next, "Conditional on all the available evidence, what is the probability P_3 that θ_3 is the true state of nature?" Not so to the orthodox statistician, who regards the word "probability" as synonomous with "long-run relative frequency in some random experiment". On this definition it is meaningless to speak of the probability of θ_3, because the state of nature is not a "random variable". Thus, if we adhere consistently to the orthodox view of probability, we will have to conclude that probability theory cannot be applied to the decision problem, at least not in this direct way.

It was just this kind of reasoning which led statisticians, in the early part of this century, to relegate problems of parameter estimation and hypothesis testing (which are really decision problems and as such are included in our general formulation) to a new field, Statistical Inference, which was regarded as distinct from probability theory. But let us look in detail at a typical problem of this type, using the loss function criterion, from the orthodox viewpoint. I want to show that a rather simple extension of the usual

orthodox arguments leads us to the same conclusion that Wald's much deeper analysis forced him to (very much against his will): that the original methods proposed by Laplace and Daniel Bernoulli are, in fact, the unique solution of the decision problem.

13.3. Parameter Estimation for Minimum Loss.

One of the situations considered in the discussion of particle counters (Lecture 8) was that of Mr. B, who knew that there was a constant, but unknown, source strength s. By observing the number of counts $\{c_1,\ldots,c_n\}$ in several different seconds, he could make an estimate of the numerical value of s, which presumably became more and more accurate with increasing n. This is a typical example of the general problem of parameter estimation.

More generally, suppose that there is ome unknown parameter α, and we make repeated observations of some quantity, obtaining an observed "sample", $x = \{x_1,\ldots,x_n\}$. We can interpret the symbol x, without subscripts, as standing for a vector in an n-dimensional "sample space". We will suppose that the possible results x_i of individual observations are real numbers. From observation of the sample x, what can we say about the unknown parameter α?

To state the problem more drastically, suppose that we are compelled to choose one specific numerical value as our "best" estimate of α, on the basis of the observed sample x, and any other prior information we might have. This is the decision situation which we all face daily, both in our capacity as scientists and engineers, and in everyday life. The driver approaching a blind intersection cannot know with certainty whether he will have enough time to cross it safely; but still he is compelled to make a decision based on what he can see, and act on it.

Now it is clear that in estimating α, the observed sample x is of no use to us unless α exerts some kind of influence on x. In other words, if we

knew α, but not x, then the probabilities $(x|\alpha) = (x_1 \ldots x_n|\alpha)$ which we would assign to various samples must depend in some way on the value of α. If we consider the different observations as independent, as is almost always done in the orthodox theory of parameter estimation, then the distribution factors:

$$(x|\alpha) = (x_1|\alpha)\ldots(x_n|\alpha) \tag{13-6}$$

However, this very restrictive assumption is not necessary (and in fact doesn't lead to any formal simplification) in discussing the general principles of parameter estimation from the decision theory standpoint.

Let $\beta = \beta(x_1 \ldots x_n)$ be an "estimator", i.e. any function of the sample values, proposed as an estimate of α. Also, let $L(\alpha,\beta)$ be the "loss" incurred by guessing the value β when α is in fact the true value. Then for any given estimator the expected loss <u>for a person who already knows the true value of α</u>, is

$$L_\alpha = \int L(\alpha,\beta)\,(x|\alpha)\; dx \tag{13-7}$$

Call this the α-expected loss. By $\int(\;)\,dx$ we mean the n-fold integration

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\quad)\; dx_1 \ldots dx_n \tag{13-8}$$

There is no need to specify different limits of integration for different problems, since if certain ranges of the x_i are impossible, the factor $(x|\alpha)$ will be zero and remove contributions from those ranges. Also, this notation includes both the continuous and discrete case, since in the latter $(x|\alpha)$ is a sum of delta-functions.

On the view of one who uses the frequency definition of probability, the above phrase, "for a person who already knows the true value of α" is misleading and unwanted. The notion of the probability of sample x <u>for a person with a certain state of knowledge</u> is entirely foreign to him; he regards $(x|\alpha)$ not as a description of a mere state of knowledge about the sample, but as an objective statement of fact, giving the relative frequencies with which dif-

ferent samples are observed "in the long run". Thus the "frequentist" believes that L_α is not merely the "mathematical expectation" of loss in the present situation, but is also, with probability 1, the limit of the average of actual losses which would be incurred by using the estimator β an indefinitely large number of times. Furthermore, the idea of finding the estimator which is "best in the present specific case" is quite foreign to his outlook; because he regards the notion of probability as meaningful only in the sense of limiting frequencies, he is forced to speak instead of finding that estimator "which will prove best in the long run".

On the frequentist view, therefore, it would appear that the best estimator will be the one that minimizes L_α. Is this a variational problem? A change $\delta\beta(x)$ in the estimator produces a change of L_α of

$$\delta L_\alpha = \int \frac{\partial L}{\partial \beta}\,(x|\alpha)\;\delta\beta(x)\;dx. \qquad (13\text{-}9)$$

If we were to require this to vanish for all $\delta\beta(x)$, this would mean

$$\frac{\partial L}{\partial \beta} = 0 \quad \text{for all possible values of } \beta. \qquad (13\text{-}10)$$

Thus the problem as stated has no truly stationary solution except in the trivial case where the loss function is independent of the estimated value β; the best estimator by the criterion of minimum α-expected loss cannot be found by variational methods. Nevertheless, we can get some understanding of the problem by considering (13-7) for some specific choices of loss function. Suppose we take the quadratic loss function $L(\alpha,\beta) = (\alpha - \beta)^2$. Then (13-7) reduces to

$$L_\alpha = \alpha^2 - 2\alpha\langle\beta\rangle + \langle\beta^2\rangle \qquad (13\text{-}11)$$

or,

$$L_\alpha = (\alpha - \langle\beta\rangle)^2 + \text{var}\,(\beta) \qquad (13\text{-}12)$$

where $\text{var}\,(\beta) = \langle\beta^2\rangle - \langle\beta\rangle^2$ is the variance, and the n'th moment

$$<\beta^n> = \int [\beta(x)]^n \ (x|\alpha) \ dx \qquad (13\text{-}13)$$

is the α-expected value of β^n. The α-expected loss is the sum of two positive terms, and a good estimator by the criterion of minimum α-expected loss has two properties:

(1) $<\beta> = \alpha$

(2) var (β) is a minimum. (13-14)

These are just the two conditions which orthodox statistics has considered most important. An estimator with property (1) is called an unbiased estimate [more generally, the function $b(\alpha) = <\beta> - \alpha$ is called the bias of the estimator $\beta(x)$], and one with both properties (1) and (2) was called efficient by R. A. Fisher (although this last condition is ambiguous until we specify the class of functions $\beta(x)$ to be taken into consideration). Nowadays, it is often called an unbiased minimum variance (UMV) estimator.

It has always seemed to me that the above reasoning amounts to looking at the problem backwards. We are describing the situation as it appears to a person who already knows the correct value of α, but does not know which specific sample has been observed. The above equations really refer to only one value of α, but involve many different possible values of x. But this is just the opposite of the state of knowledge which we have when we estimate a parameter; we know x, but not α. Our equations should involve only one sample, namely the one actually observed; but sholld take into account many different possible values of α.

Our job is always to do the best reasoning we can about the single situation that exists here and now, on the basis of the knowledge which we do in fact have; consideration of how things might seem to a person whose state of knowledge is different, or what might happen in some other situation that we are not reasoning about (if some different sample were observed) is

not relevant to our problem. So, we ought to do it the other way around; it is the expected value of $L(\alpha,\beta)$ over the posterior distribution $(\alpha|x)$ of α, conditional on knowledge of the sample, that should logically be minimized. Call this the x-expected loss:

$$L_x(\beta) = \int L(\alpha,\beta)(\alpha|x)\, d\alpha \tag{13-15}$$

where $(\alpha|x)$ is obtained by applying Bayes' theorem. Thus, having observed the sample x, we should calculate $L_x(\beta)$ and take as our estimate that value of β which minimizes $L_x(\beta)$. In the continuous case, subject to some elementary regularity conditions, we would use the estimator $\beta(x_1,\ldots,x_n)$ determined by

$$\frac{\partial L_x(\beta)}{\partial \beta} = 0 \tag{13-16}$$

$$\frac{\partial^2 L_x(\beta)}{\partial \beta^2} > 0 \tag{13-17}$$

These equations make no reference to any sample other than the specific one that has been observed.

But most of the prominent workers in statistics would raise strong objections to this procedure on philosophical grounds [you guessed it--that $(\alpha|x)$ is meaningless because α is not a "random variable"]. So, let's go back and take a closer look at the orthodox formulation of the problem--is there some way we could improve it without conflicting with orthodox principles?

We have already seen a practical difficulty faced in the first formulation; the criterion of minimum α-expected loss does not lead to a variational problem, and therefore even in the simplest case of a quadratic loss function, it gives us no analytical method for constructing the "best" estimator $\beta(x_1\ldots x_n)$. In fact, it is clear from (13-14) that the only really correct solution of the mathematical problem as stated, is $\beta(x_1\ldots x_n) = \alpha$, independent of the observed sample. This shows again that the criterion of minimum α-expected loss essentially describes the reasoning of a person who already

knows the correct value of α. However, the stubborn fact remains that the statistician using this criterion does not know α, and so he cannot use the correct solution of the problem. His estimator must be some function of the sample values only. Once an estimator has been suggested, it can be tested by calculating (13-12). But, except for one special class of sampling distributions $(x_1 \ldots x_n|\alpha)$, which I will consider later, the frequentist has no general principle like (13-16), only his judgment and common sense, to tell him which ones to try out in the first place.

13.4. Should We Use an Unbiased Estimate?*

What is the relative importance of removing bias and minimizing the variance? Well, from (13-12) it would appear that they are of exactly equal importance; there is no advantage in removing the bias $(\langle\beta\rangle - \alpha)$ if in so doing we increase var(β) more than enough to compensate. Yet that is just what the orthodox statistician usually does! Let me give you a specific example of this. Cramér (1946, p. 351) considers the problem of estimating the variance μ_2 of a distribution $(x_1|\mu_2)$:

$$\mu_2 = \langle x_1^2\rangle - \langle x_1\rangle^2 = \langle x_1^2\rangle \tag{13-18}$$

from n independent observations $\{x_1 \ldots x_n\}$. We assume, in (13-18) and in what follows, that $\langle x_1\rangle = 0$ since a trivial change of variables would in any event accomplish this. An elementary calculation shows that the sample variance

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^2 \tag{13-19}$$

has expectation value, over the distribution $(x_1 \ldots x_n|\mu_2) = (x_1|\mu_2)\ldots(x_n|\mu_2)$, of

$$\langle m_2\rangle = \frac{n-1}{n}\mu_2 \tag{13-20}$$

---

*This section is a digression in response to a question from the audience. It may be skipped without losing the main line of argument; however, it does contain an illustration of an important point.

and thus, as an estimator of μ_2 it has a negative bias. So, goes the argument, we should correct this by using the unbiased estimator

$$M_2 = \frac{n}{n-1}\, m_2 \tag{13-21}$$

Now, of course, the only thing that really matters here is the <u>total</u> error of our estimate; the particular way in which you or I separate error into two abstractions labelled "bias" and "variance" is a purely academic matter with no bearing on the actual quality of the estimator. So, let's look at the mean square error criterion. Replacement of m_2 by M_2 removes a term $(\langle m_2\rangle - \mu_2)^2 = \mu_2^2/n^2$ in (13-12); but it also increases the term $\mathrm{var}(m_2)$ by a factor $[n/(n-1)]^2$, so it seems obvious that, at least for large n, this has made things worse instead of better. Let's check this more carefully. Suppose we replace m_2 by the estimator,

$$\gamma_\delta = (1 + \delta)\, m_2 \tag{13-22}$$

What is the best choice of δ? The μ_2-expected loss (13-12) is now

$$\langle(\gamma_\delta - \mu_2)^2\rangle = \mu_2^2 - 2(1+\delta)\mu_2\langle m_2\rangle + (1+\delta)^2\langle m_2^2\rangle$$

$$= [(\langle m_2\rangle - \mu_2)^2 + \mathrm{var}(m_2)] - \langle m_2^2\rangle q^2 + \langle m_2^2\rangle(\delta - q)^2 \tag{13-23}$$

where

$$q \equiv \frac{\langle m_2^2\rangle - \mu_2\langle m_2\rangle}{\langle m_2^2\rangle} \tag{13-24}$$

Evidently, the best estimator in the class γ_δ is the one with $\delta = q$, and the term $-\langle m_2^2\rangle q^2$ in (13-23) represents the decrease in mean-square error obtainable by using γ_q instead of m_2. From Cramér's result (loc. cit., Eq. 27.4.2)

$$\mathrm{var}(m_2) = \langle m_2^2\rangle - \langle m_2\rangle^2 = n^{-3}\,(n-1)[(n-1)\mu_4 - (n-3)\mu_2^2] \tag{13-25}$$

where

$$\mu_4 = \langle(x_1 - \langle x_1\rangle)^4\rangle = \langle x_1^4\rangle$$

is the fourth central moment of $(x_1|\mu_2)$, we find

$$n^3\langle m_2^2\rangle = (n-1)\ [(n^2-n+2)\mu_2^2 + (n-1)\ \mathrm{var}(x^2)] \tag{13-26}$$

$$n^3\langle m_2^2\rangle q = (n-1)\ [(n-2)\mu_2^2 - (n-1)\ \mathrm{var}(x^2)] \tag{13-27}$$

where $\mathrm{var}(x^2) = \mu_4 - \mu_2^2 \geq 0.$ (13-28)

We must understand $n > 1$ in all this, for if $n = 1$, we have $m_2 = 0$; a single observation gives no information at all about the variance of $(x_1|\mu_2)$. But if $n = 2$, we have $q \leq 0$; instead of removing the bias, we should always <u>increase</u> it in order to minimize the mean square error! More generally, if $\mathrm{var}(x^2) = K\mu_2^2$ we have from (13-26), (13-27):

$$q = \frac{(n-2) - (n-1)\ K}{(n^2-n+2) + (n-1)\ K} \tag{13-29}$$

and therefore, if $K < 1$,

$$q \lessgtr 0 \quad \text{if } n \lessgtr \frac{2-K}{1-K} \tag{13-30}$$

while if $K \geq 1$, $q < 0$ for all n.

In the case of a Gaussian distribution,

$$(x_1|\mu_2) = A \exp\left[-\frac{x_1^2}{2\mu_2}\right] \tag{13-31}$$

we have

$$K = \frac{\langle x_1^4\rangle - \langle x_1^2\rangle^2}{\langle x_1^2\rangle^2} = 2 \tag{13-32}$$

We will seldom have $K < 2$, for this would imply that $(x_1|\mu_2)$ cuts off even more rapidly than Gaussian for large x_1. If $K = 2$, (13-29) reduces to

$$q = -\frac{1}{n+1} \tag{13-33}$$

which again says that rather than removing the bias we should approximately double it, in order to minimize the expected square of the error. How much better is the estimator γ_q than M_2? In the Gaussian case the mean square error of the estimator γ_q is

$$\langle(\gamma_q - \mu_2)^2\rangle = \frac{2}{n+1}\mu_2^2 \tag{13-34}$$

For a general choice of δ, it is

$$\langle(\gamma_\delta - \mu_2)^2\rangle = \mu_2^2\left[\frac{2}{n+1} + \frac{n^2-1}{n^2}\left(\delta + \frac{1}{n+1}\right)^2\right] \qquad (13\text{-}35)$$

The unbiased estimator M_2 corresponds to the choice

$$\delta = \frac{1}{n-1} \quad , \qquad (13\text{-}36)$$

and thus to the mean square error

$$\langle(M_2 - \mu_2)^2\rangle = \mu_2^2\left[\frac{2}{n+1} + \frac{2}{n}\right] \qquad (13\text{-}37)$$

which is over twóce the amount incurred by use of γ_q.

Most distributions which arise in practice, if not gaussian, have wider "tails" than gaussian so that $K > 2$. In this case, the difference will be even greater.

Up to this point, it may have seemed that I was quibbling over a very minor thing--changes in the estimator of one or two parts out of n. But now you see that the difference between (13-34) and (13-37) is not at all trivial. For example, with Cramér's unbiased estimator M_2 you will need n = 203 observations in order to get as small a mean-square error as the biased estimator γ_q gives you with only 100 observations.

There is a fantastic example in a recent book on econometrics (Valavanis, 1959; p. 60) where the author attaches such great importance to removing bias that he advocates throwing away practically all the data from the sample, if necessary, to achieve this. One reason for such an undue emphasis on bias is the belief that if we draw N successive samples of n observations each and calculate the estimators $\beta_1 \ldots \beta_N$, the average $\bar{\beta} = N^{-1}\sum\beta_i$ of these estimates will converge in probability to $\langle\beta\rangle$ as $N \to \infty$, and thus an unbiased estimator will, on sufficiently prolonged sampling, give an arbitrarily accurate estimate of α. Such a belief is almost never justified even for the fairly well controlled measurements of the physicist or engineer, not only because of

unknown systematic error, but because successive measurements lack the independence required for these limit theorems to apply. In such uncontrolled situations as economics, the situation is far worse.

But unbiased estimators are, even if we accept these limit theorems, not the only ones which appraach perfect accuracy with indefinitely prolonged sampling. Many other estimators approach the true value of α in this limit, and do it more rapidly. Our γ_q is a specific example. Furthermore, asymptotic behavior of an estimator is not really relevant, because the practical problem is always to do the best we can with a finite sample; therefore the important question is not whether an estimator tends to the true value, but how rapidly it does so.

I have a dark suspicion that a still more important reason for attaching such an undeserved importance to bias is simply that we have been caught in a psycho-semantic trap. It is well known to politicians that our thought processes are influenced to a rather alarming degree by the particular choice of words we use. When we call the quantity $\langle\beta\rangle-\alpha$ the "bias", that makes it sound like something awfully reprehensible, which we must get rid of at all costs. If we had called it instead the "component of error orthogonal to the variance", as suggested by the Pythagorean form of Eq. (13-12), then it would be clear to all that these two contributions to the error are on an exactly equal footing; and that it is folly to decrease one at the expense of increasing the other.

In the book of Chernoff and Moses (1959) these points are clearly recognized, and an even more forceful example is given showing what can be wrong with the criterion of an unbiased estimate. A company is laying a telephone cable across the ocean. They cannot know in advance exactly how much cable will be required, and so they must estimate. If they overestimate, the loss will presumably be proportional to the amount of excess cable to be disposed

of; but if they underestimate and the cable end falls off into the water, the result may be financial disaster. Use of an unbiased estimate here could only be described as foolhardy.

Note, however, that after all this argument, nothing in the above entitles us to conclude that γ_q is the best estimator of μ_2 by the mean-square criterion! For we have considered only the class (13-22) of estimators constructed by multiplying the sample variance (13-19) by some preassigned number; we can say only that γ_q is the best one in that class. The question whether some other function of the sample values, not a multiple of (13-19), might be still better by the mean-square error criterion, remains completely open. This weakness of the orthodox approach to parameter estimation--that it does not tell us how to find the best estimator, but only how to compare different guesses--is due to our having "looked at the problem backwards", in the sense I explained a moment ago. Now I want to show how the trouble can be overcome.

13.5. Reformulation of the Problem.

It is easy to see why the orthodox criterion of minimum α-expected loss is bound to get us into trouble and is unable to furnish any general rule for constructing an estimator. The mathematical problem was: for given $L(\alpha,\beta)$ and $(x|\alpha)$, what function $\beta(x_1 \ldots x_n)$ will minimize

$$L_\alpha = \int L(\alpha,\beta)\,(x|\alpha)\; dx \qquad (13\text{-}38)$$

Although this is not a variational problem, it might have a unique solution; but the solution will still, in general depend on α. Of course, there may be (and in fact are) a few exceptional cases where the α-dependence drops out; but in general the criterion of minimum α-expected loss leads to an impossible situation--even if we could solve the mathematical problem (13-38) and had before us the best estimator $\beta_\alpha(x_1 \ldots x_n)$ for each value of α, we could use the result only if α were already known, in which case we would have no need

to estimate. We were indeed looking at the problem backwards!

This makes it clear that in general we cannot use the criterion (13-38), or in fact any criterion which makes reference to only a single value of α; not for philosophical reasons but because any such criterion is built on self-contradictory premises. The person who advises us to use (13-38) puts himself in exactly the position of the shoe clerk who told a customer, "You will never be able to get those new boots on until you have worn them a while."

This also makes it clear how to correct the trouble. It is of no use to ask what estimator is best for some particular value of α, even though the question might have a definite answer; the only reason for using an estimator is that α is unknown. The estimator must therefore be some compromise that allows for all possibilities within some prescribed range of α; within this range it must do the best job of protecting against loss <u>no matter what the true value of α turns out to be</u>.

Thus it is some weighted average of L_α,

$$\langle L \rangle = \int f(\alpha)\ L_\alpha\ d\alpha \qquad (13\text{-}39)$$

that we should really minimize, where the function $f(\alpha) \geq 0$, which will be given a fuller interpretation later, measures in some way the relative importance of minimizing L_α for various possible values of α.

Merely to recognize this, which amounts to removing a contradiction in the original formulation, already implies the solution. For the mathematical character of the problem is completely changed by adopting (13-39) instead of (13-38). We now have a solvable variational problem with a well-behaved solution. The first variation in $\langle L \rangle$ due to an arbitrary variation $\delta\beta(x_1 \ldots x_n)$ in the estimator is

$$\delta\langle L \rangle = \int f(\alpha)\ d\alpha \int \ldots \int dx_1 \ldots dx_n \frac{\partial L}{\partial \beta}\ (x_1 \ldots x_n | \alpha)\ \delta\beta(x_1 \ldots x_n)$$

which vanishes independently of $\delta\beta$ if

$$\int d\alpha\ f(\alpha)\ \frac{\partial L}{\partial \beta}\ (x_1 \ldots x_n | \alpha) \ = 0 \qquad (13\text{-}40)$$

for all possible samples $\{x_1 \ldots x_n\}$. Equation (13-40) is the fundamental integral equation which determines the best estimator.

Taking the second variation, we find the condition that (13-40) shall yield a true minimum is

$$\int d\alpha\ f(\alpha)\ \frac{\partial^2 L}{\partial \beta^2}\ (x_1 \ldots x_n | \alpha) > 0 \qquad (13\text{-}41)$$

Thus a sufficient condition for a minimum is simply

$$\frac{\partial^2 L}{\partial \beta^2} \geq 0 \qquad (13\text{-}42)$$

but this is far stronger than necessary.

If we take the quadratic loss function $L(\alpha,\beta) = K(\alpha - \beta)^2$, equation (13-40) reduces to

$$\int d\alpha\ f(\alpha)\ (\alpha - \beta)\ (x_1 \ldots x_n | \alpha) = 0$$

or, the optimal estimator for quadratic loss is

$$\beta(x_1 \ldots x_n) = \frac{\int d\alpha\ f(\alpha)\ \alpha\ (x_1 \ldots x_n | \alpha)}{\int d\alpha\ f(\alpha)\ (x_1 \ldots x_n | \alpha)} \qquad (13\text{-}43)$$

But, you see, this is just the mean value over the posterior distribution of α:

$$(\alpha | x_1 \ldots x_n) = \frac{f(\alpha)\ (x_1 \ldots x_n | \alpha)}{\int d\alpha\ f(\alpha)\ (x_1 \ldots x_n | \alpha)} \qquad (13\text{-}44)$$

given by Bayes' theorem if we interpret $f(\alpha)$ as a prior probability density!

This example shows us, perhaps more clearly than any I have given so far, why the mathematical form of Bayes' theorem is always going to be the fundamental principle behind parameter estimation, independently of all philosophical arguments about the "meaning of probability", or about "random variables".

Let's see what happens for some other loss functions. If we take as a loss function the absolute error, $L(\alpha,\beta) = |\alpha - \beta|$, then the fundamental equation (13-40) becomes

$$\int_{-\infty}^{\beta} d\alpha\ f(\alpha)\ (x_1 \ldots x_n|\alpha) = \int_{\beta}^{\infty} d\alpha\ f(\alpha)\ (x_1 \ldots x_n|\alpha)$$

which states that $\beta(x_1 \ldots x_n)$ is to be taken as the median over the posterior distribution of α:

$$\int_{-\infty}^{\beta} d\alpha\ (\alpha|x_1 \ldots x_n) = \int_{\beta}^{\infty} d\alpha\ (\alpha|x_1 \ldots x_n) = \frac{1}{2} \tag{13-45}$$

Likewise, if we take a loss function $L(\alpha,\beta) = (\alpha - \beta)^4$, equation (13-40) leads to an estimator $\beta(x_1 \ldots x_n)$ which is the real root of

$$f(\beta) = \beta^3 - 3\bar{\alpha}\beta^2 + 3\overline{\alpha^2}\beta - \overline{\alpha^3} = 0 \tag{13-46}$$

where

$$\overline{\alpha^n} = \int d\alpha\ \alpha^n\ (\alpha|x_1 \ldots x_n) \tag{13-47}$$

is the n'th moment of the posterior distribution of α. [That (13-46) has only one real root is seen on forming the discriminant; the condition $f'(\beta) \geq 0$ for all real β is just $(\overline{\alpha^2} - \bar{\alpha}^2) \geq 0$.]

If we take $L(\alpha,\beta) = |\alpha - \beta|^k$, and pass to the limit $k \to 0$, or if we just take

$$L(\alpha,\beta) = \begin{cases} 0, & \alpha = \beta \\ 1, & \text{otherwise} \end{cases} \tag{13-48}$$

Eq. (13-40) tells us that we should choose $\beta(x_1 \ldots x_n)$ as the most probable value, or mode of the posterior distribution $(\alpha|x_1 \ldots x_n)$. If $f(\alpha) = \text{const.}$, this is just Fisher's maximum likelihood estimate.

In this result we finally see just what maximum likelihood accomplishes, and under what circumstances it is the optimal method to use. The maximum likelihood criterion is the one in which we care only about the chances of being exactly right; and if we are wrong, we don't care how wrong we are. This is just the situation we have in shooting at a small target, where "a miss is as good as a mile". But it is clear that there aren't very many other situations where this would be a rational way to behave; almost always, the amount of error is of some concern to us, and so maximum likelihood is not the best estimation criterion.

Note that in all these cases it was the posterior distribution $(\alpha|x_1 \ldots x_n)$ that was involved. That this will always be the case is easily seen by noting that our "fundamental integral equation" (13-40) is not so profound after all. It can equally well be written as

$$\frac{\partial}{\partial \beta} \int d\alpha \; f(\alpha) \; L(\alpha,\beta) \; (x_1 \ldots x_n|\alpha) = 0.$$

but if we interpret $f(\alpha)$ as a prior probability density, this is identical with (13-16), which we had already derived from much simpler reasoning! Likewise the condition (13-41) for a true minimum is identical with (13-17).

13.6. "Objectivity" of Decision Theory.

Decision Theory occupies a unique position in discussion of the logical foundations of statistics, because, as we have seen in (13-16) and (13-40), its procedures can be derived from either of two diametrically opposed viewpoints about the nature of probability theory, and it thus forms a kind of bridge between them. While there appears to be universal agreement as to the actual procedures that should be followed, there remains a fundamental disagreement as to the underlying reason for them, having its origin in the old issue of frequency vs. non-frequency definitions of probability.

From a pragmatic standpoint, such considerations may seem at first to be unimportant. However, in the attempt to apply decision-theory methods in real problems one learns very quickly that these questions intrude in the initial stage of setting up the problem in mathematical terms. In particular, our judgment as to the generality and range of validity of decision-theory methods depends on how these conceptual problems are resolved. My aim is to expound the viewpoint according to which these methods have the greatest possible range of application. Now we find that the main source of controversy here is on the issue of prior probabilities; on the orthodox viewpoint, if the problem involves use of Bayes' theorem then these methods are just not

applicable unless the prior probabilities are known frequencies. But to maintain this position consistently would imply an enormous restriction on the range of legitimate applications. Therefore, let's see whether the mathematical form of our final equations can shed any light on this issue.

Notice that only the product $f(\alpha)L(\alpha,\beta)$ is involved in (13-40) or (13-16); thus whether we interpret the problem as:

(A) Prior probability $f(\alpha)$, loss function $L(\alpha,\beta) = (\alpha - \beta)^2$ or as

(B) Uniform prior probability, loss function $L(\alpha,\beta) = f(\alpha)(\alpha - \beta)^2$ or as

(C) Prior probability $g(\alpha)$, loss function $f(\alpha)(\alpha - \beta)^2/g(\alpha)$, the solution is just the same. This is equally true for any loss function.

I emphasize this rather trivial mathematical property because of a curious psychological phenomenon. In expositions of decision theory written from the orthodox viewpoint, the writers are always very reluctant to introduce the notion of prior probability. They postpone it as long as possible, and finally give in only when the mathematics forces them to recognize that prior probabilities form the only basis for choice among the different admissible decisions. Even then, they are so unhappy about the use of prior probabilities that they feel it necessary always to invent a situation--often highly artificial--which makes the prior probabilities appear to be frequencies; and they will not use this theory for any problem where they don't see how to do this. But these same writers do not hesitate to pull a completely arbitrary loss function out of thin air without any basis at all, and proceed with the calculation!

The equations show that if your final decision depends strongly on which particular prior probability assignment you use, it is going to depend just as strongly on which particular loss function you use. If you worry about arbitrariness in the prior probabilities, then in order to be consistent, you ought to worry just as much about arbitrariness in the loss functions. If

you claim (as most writers on this subject have been doing for decades) that uncertainty as to the proper choice of prior probabilities invalidates the Laplace-Bayes theory, then in order to be consistent, you must also claim that uncertainty as to the proper choice of loss functions invalidates Wald's decision theory.

The reason for this strange lopsided attitude is closely connected with a certain philosophy variously called behavioristic, or positivistic, which wants us to restrict our statements and concepts to objectively verifiable things. Therefore the observable decision is the thing to emphasize, while the process of inductive reasoning and the judgment described by a prior probability must be swept under the rug. But I see no need to do this, because it seems to me obvious that rational action can come only as the result of rational thought.

If we refuse to consider the problem of rational thought merely on the grounds that it is not "objective", the result will not be that we obtain a more "objective" theory. The result will be that we have lost the possibility of getting any satisfactory theory at all, because we have denied ourselves any way of describing what is actually going on in the decision process. And, of course, the loss function is just the expression of a purely subjective value judgment, which can in no way be considered any more "objective" than the prior probabilities.

In fact, I claim that the prior probabilities are usually more objective than the loss function, both in the mathematical theory and in the everyday decision problems of "real life". In the mathematical theory we have two quite general formal principles--maximum entropy and transformation groups--that completely remove the arbitrariness of prior probabilities for a large class of important problems, which includes most of those discussed in statistical text books. Of course, these principles will not solve all problems,

and undoubtedly there are more such principles waiting to be discovered. I hope that one result of these talks will be to encourage others to seek them. To the best of my knowledge, there are as yet no general principles for determining loss functions--not even where the criterion is purely economic, because the utility of money remains ill-defined.

In "real life" decision problems, we have a similar situation. Each man knows, pretty well, what his prior probabilities are; and because his beliefs are based on all his past experience, they are not easily changed by one more experience, so they are fairly stable things. But in the heat of argument he may lose sight of his loss function; or he may never have bothered to reason out the consequences of his actions. Thus the labor mediator must deal with parties with violently opposing ideologies; policies considered noble by one party are regarded as reprehensible by the other. The successful labor mediator realizes that mere talk will not alter prior beliefs; and so his role must be to turn the attention of both parties away from this area, and explain clearly to each what his loss function is. In this sense, I think we can claim that in real life decision problems, the loss function is often far more "subjective" (in the sense of being less well fixed in our minds) than the prior probabilities.

Of course, we have to concede this much to the behaviorists--the final criterion by which we judge the soundness of any theory must be on the objective, pragmatic level. After a theory has been constructed, the ultimate test we apply to it is not whether its premises are philosophically satisfying, but how it works out in practice. Indeed, a major objective of these talks is to show you, in detail, just how the Laplace-Bayes theory does work out in practice and how its results compare with those of the orthodox methods; because that is something you very seldom find in any of the literature written from the orthodox viewpoint.

But in the process of constructing a theory, we must demand the right to invent and use any concepts we please, whether or not these concepts are themselves "objectively verifiable". If we deny ourselves this freedom on the grounds of some philosophical dogma, we are putting ourselves in a strait-jacket which effectively prevents further progress. In the case of physical theories, this point has been stressed repeatedly and strongly by Einstein; his own work is, of course, the perfect example of what can be accomplished through the free invention of new concepts.

Now let's see the extent to which varying loss functions lead to varying decisions, by some numerical examples.

13.7. Effect of Varying Loss Functions.

Suppose that on the basis of the observed sample x, a parameter α has the posterior distribution

$$(\alpha|x) = k\, e^{-k\alpha} \quad , \quad 0 \le \alpha < \infty \tag{13-49}$$

This has the n'th moment

$$\langle \alpha^n \rangle = \int_0^\infty \alpha^n \, (\alpha|x) \, d\alpha = n!\, k^{-n} \quad . \tag{13-50}$$

With loss function $(\alpha - \beta)^2$, the best estimator is the mean value

$$\beta = \langle \alpha \rangle = k^{-1} \quad . \tag{13-51}$$

With loss function $|\alpha - \beta|$, the estimator is the median, determined by

$$\frac{1}{2} = \int_0^\beta (\alpha|x) \, d\alpha = 1 - e^{-k\beta} \tag{13-52}$$

or

$$\beta = k^{-1} \ln 2 = 0.693 \, \langle \alpha \rangle . \tag{13-53}$$

To minimize $\langle (\alpha - \beta)^4 \rangle$, we should choose β to satisfy equation (13-46), which becomes, in this case,

$$y^3 - 3y^2 + 6y - 6 = 0 \tag{13-54}$$

with $y = k\beta$. The root of this is at $y = 1.59$, so the optimal estimator

with loss function $|\alpha - \beta|^4$ is

$$\beta = 1.59 \langle\alpha\rangle. \tag{13-55}$$

For the loss function $(\alpha - \beta)^{s+1}$ with s an odd integer, the fundamental equation (13-40) is

$$\int_0^\infty (\alpha - \beta)^s \, e^{-k\alpha} \, d\alpha = 0 \tag{13-56}$$

which reduces to

$$\sum_{m=0}^{s} \frac{(-k\beta)^m}{m!} = 0 \tag{13-57}$$

of which (13-54) is a special case with s = 3. In the case s = 5, loss function $(\alpha - \beta)^6$, we find

$$\beta = 2.025 \langle\alpha\rangle. \tag{13-58}$$

As $s \to \infty, \beta$ also increases without limit. But the maximum-likelihood estimate, which corresponds to the loss function

$$L(\alpha,\beta) = -\delta(\alpha - \beta)$$

or equally well to

$$\lim_{k\to 0} |\alpha - \beta|^k$$

is $\beta = 0$!

These numerical examples merely illustrate what was already clear intuitively; when the posterior distribution $(\alpha|x)$ is not sharply peaked, the best estimate of α depends very much on which particular loss function we use.

You might suppose that a loss function must always be a monotonically increasing function of the error $|\alpha - \beta|$. In general, of course, this will be the case, but there is nothing in this theory which restricts us to such functions. You can think of some rather frustrating situations in which, if you are going to make an error, you would rather make a large one than a small one. William Tell was in just that fix. If you study our equations for this case, you will see that there is really no very satisfactory decision

at all; and nothing can be done about it.

Our noting that the final decision depends only on the product of prior probability and loss function also helps to clear up a mystery which has long been puzzling to Bayesians. As we noted in Lecture 8, Jeffreys (1939) proposed that, in the case of a continuous parameter α known to be positive, we should express prior ignorance by assigning, not uniform prior probability, but a prior density proportional to $(1/\alpha)$. The theoretical justification of this rule was long unclear; but it yields very sensible-looking results in practice, which led Jeffreys to adopt it as fundamental in his significance tests. We saw in Lecture 12 that, in the case that α is a scale parameter, the Jeffreys prior is uniquely determined by invariance under the transformation group; but now we can see a still more general justification of it.

From the decision-theory viewpoint the thing that matters is not the prior or loss function separately; only their product enters into the final decision. If we use the absolute error loss function $|\beta - \alpha|$ when α is known to be positive, then to assign $f(\alpha) = \text{const.}$ in (13-45) amounts to saying that we demand an estimator which yields, as nearly as possible, a constant <u>absolute</u> accuracy for all values of α in $0 < \alpha < \infty$. That is clearly asking for too much in the case of large α; and we must pay the price in a poor estimate for small α. But we now see that the median of Jeffreys' posterior distribution is mathematically the same thing as the optimal estimator for uniform prior and loss function $|\beta - \alpha|/\alpha$; we ask for, as nearly as possible, a constant <u>percentage</u> accuracy over all values of α. This is, of course, exactly what we do want in most cases where we know that $0 < \alpha < \infty$. The reason for the superior performance of Jeffreys' rule is thus made apparent; and the mystery disappears if we re-interpret it as saying that the $(1/\alpha)$ factor is part of the loss function.

13.8. General Decision Theory.

In the foregoing, I have developed decision theory only in terms of one particular application; parameter estimation. But we really have the whole story already; the criterion (13-16) for constructing the optimal estimator generalizes immediately to the criterion for finding the optimal decision of any kind. The final rules are simply:

(1) Enumerate the possible states of nature θ_j, discrete or continuous, as the case may be.

(2) Assign prior probabilities $(\theta_j|X)$ which maximize the entropy subject to whatever prior information X you have.

(3) Digest any additional evidence E by application of Bayes' theorem, thus obtaining the posterior probabilities $(\theta_j|EX)$.

(4) Enumerate the possible decisions D_i.

(5) Specify the loss function $L(D_i,\theta_j)$ that tells what you want to accomplish.

(6) Make that decision D_i which minimizes the expected loss $\langle L\rangle_i = \sum_j L(D_i,\theta_j)(\theta_j|EX)$.

That is all there is to it; after all is said and done, the final rules of calculation to which the theorems of Cox, Wald, and Shannon lead us are just the ones which had already been developed by Bayes, Laplace, and Daniel Bernoulli in the 18'th century, except that the entropy principle generalizes the principle of indifference.

These rules either include, or improve upon, practically all known statistical methods for hypothesis testing and point estimation of parameters. If you have mastered them, then you have just about the entire field at your fingertips. The most outstanding thing about them is their simplicity--if we

sweep aside all the polemics and false starts that have cluttered up this field in the past and consider only the constructive arguments that lead directly to these rules, it is clear that the underlying rationale could be fully developed in a one-semester undergraduate course.

However, in spite of the utter simplicity of the rules themselves, really facile application of them involves intricate mathematics, and fine subtleties of concept; so much so that several generations of workers in this field misused them and concluded that the rules were all wrong! So, we still need a good deal of leading by the hand in order to develop facility in using them. It is a good deal like learning how to play a musical instrument--anybody can make noise with it, but you will not play this instrument well without years of practice.

As an example--although a rather trivial one--of the little tricks that help in applying this theory, note that the decision rule is invariant under any proper linear transformation of the loss function; i.e. if $L(D_i,\theta_j)$ is one loss function, then the new one

$$L'(D_i,\theta_j) = a + b\, L(D_i,\theta_j)$$

where $-\infty < a < \infty$, $0 < b < \infty$, will lead to the same decision, whatever the prior probabilities $(\theta_j|X)$ and new evidence E. Thus, in a binary decision problem, given the loss matrix

$$L_{ij} = \begin{pmatrix} 10 & 100 \\ 19 & 10 \end{pmatrix}$$

we can equally well use

$$L'_{ij} = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}$$

corresponding to a = -10/9, b = 1/9. This may simplify the calculation of expected loss quite a bit.