Lecture 9

# INTERVAL ESTIMATION AND ASYMPTOTIC PROPERTIES

There is still an essential feature missing in the comparison of Mr. A and Mr. B in our particle-counter problem. We would like to have some measure of the degree of reliability which they attach to their estimates, especially in view of the fact that their estimates are so different. Clearly, the best way of doing this would be to draw the entire probability distributions

$$(n_1 | c_2 c_1 X_A) \qquad \text{and} \qquad (n_1 | c_2 c_1 X_B)$$

and from this make statements of the form, "90 per cent of the posterior probability is concentrated in the interval $\alpha < n_1 < \beta$." But, for present purposes, we will be content to give the standard deviations [i.e., the square root of the variance as defined in Eq. (8-10)] of the various distributions we have found. An inequality due to Tchebycheff then asserts that, if $\sigma$ is the standard deviation, then the amount p of probability concentrated between the limits $(\bar{n}_1 \pm t\sigma)$ satisfies

$$p \geq 1 - \frac{1}{t^2} \tag{9-1}$$

This tells us nothing when $t \leq 1$, but it tells us more and more as t increases beyond unity. For example, at least 3/4 of the probability must be assigned to the nange $\bar{n} \pm 2\sigma$, and at least 8/9 to the range $\bar{n} \pm 3\sigma$.

## 9.1. Calculation of Variance.

The variances $\sigma^2$ of all the distributions we have found in the last

lecture are readily calculated. In fact, the calculation of any moment of these distributions is easily performed by making use of the general formula

$$\sum_{m=0}^{\infty} \binom{m+a}{m} m^n x^m = \left(x \frac{d}{dx}\right)^n \frac{1}{(1-x)^{a+1}}, \quad |x| < 1, \tag{9-2}$$

which we have already used in calculation of the mean value in (8-21). For Mr. A and Mr. B, and the Jeffreys prior probability distribution, we find the variances

$$\text{Var } (n_1 | c_1 X_A) = \frac{(c_1+1)(1-a)}{a^2} \tag{9-3}$$

$$\text{Var } (n_1 | c_2 c_1 X_B) = \frac{(c_1+c_2+1)(1-a^2)}{4a^2} \tag{9-4}$$

$$\text{Var } (n_1 | c_1 X_J) = \frac{c_1(1-a)}{a^2} \tag{9-5}$$

and the variances for $n_2$ are found from symmetry.

This has been a rather long discussion, so let's summarize all our results so far in a table. I'll give, for problem 1 and problem 2, the most probable values of number of particles as found by Mr. A and Mr. B, and also the (mean value) ± (standard deviation), which provides a reasonable interval estimate.

From this table we see that Mr. B's extra information not only has led him to change his estimates considerably from those of Mr. A, but it has enabled him to make an appreciable decrease in his probable error. Prior information which has nothing to do with frequencies can greatly alter the conclusions we draw from a random experiment, and their degree of reliability.

It is also of interest to ask how good Mr. B's estimate of $n_1$ would be if he knew only $c_2$; and therefore had to use the distribution (8-41) representing reasoning along the route $c_2 \to n_2 \to s \to n_1$ of Fig. (8.2). From (8-41) we find the most probable, and the (mean) ± (standard deviation) estimates

| | | Problem 1 $c_1 = 10$ | Problem 2 $c_1 = 10$ $c_2 = 16$ | |
|---|---|---|---|---|
| | | $n_1$ | $n_1$ | $n_2$ |
| A | most prob. | 100 | 100 | 160 |
| | mean ± s.d. | 109 ± 31 | 109 ± 31 | 169 ± 39 |
| B | most prob. | 100 | 127 | 133 |
| | mean ± s.d. | 109 ± 31 | 131.5 ± 26 | 137.5 ± 26 |
| J | most prob. | 91 | | |
| | mean ± s.d. | 100 ± 30 | | |

$$\hat{n}_1 = \frac{c_2}{a} = 160 \qquad (9\text{-}6)$$

$$\text{mean} \pm \text{s.d.} = \frac{c_2+1}{a} \pm \frac{\sqrt{(c_2+1)(a+1)}}{a} = 170 \pm 43.3 \qquad (9\text{-}7)$$

In this case he would obtain slightly poorer estimate (i.e. a larger probable error) than Mr. A even if the counts $c_1 = c_2$ were the same, because the variance (9-3) for the direct route contains a factor (1-a), which gets replaced by (1+a) if we have to reason over the indirect route. Thus, if the counter has low efficiency, the two routes give nearly equal reliability for equal counting rates; but if it has high efficiency, a ≃ 1, then the direct route $c_1 \rightarrow n_1$ is far more reliable. I think your common sense will tell you that this is just as it should be.

## 9.2. Generalization and Asymptotic Forms.

Now in the last lecture we conjectured that Mr. B might be helped a good deal more in his estimate of $n_1$ by acquiring still more data $\{c_3, c_4, \ldots, c_m\}$. Let's investigate that further. The standard deviation of the distribution (8-8) in which the source strength was known exactly, is only $\sqrt{s(1-a)} = 10.8$ for $s = 130$; and from the table, Mr. B's standard deviation for his estimate of $n_1$ is now about 2.5 times this value. What would happen if we gave him more and more data from other time intervals, such that his estimate of s approached 130? To answer this, note that, if $1 \leq k \leq m$, we have (now dropping the $X_B$ because we will be concerned only with Mr. B from now on):

$$(n_k|c_1\ldots c_m) = \int_0^\infty (n_k s|c_1\ldots c_m)\, ds$$

$$= \int_0^\infty (n_k|sc_k)(s|c_1\ldots c_m)\, ds \qquad (9\text{-}8)$$

in which we have put $(n_k|sc_1\ldots c_m) = (n_k|sc_k)$ because, from Fig. (8.2), if s is known, then all the $c_i$ with $i \neq k$ are irrelevant for inference about $n_k$. The second factor in the integrand of (9-8) can be evaluated by Bayes' theorem:

$$(s|c_1\ldots c_m) = (s|X_B)\,\frac{(c_1\ldots c_m|s)}{(c_1\ldots c_m|X_B)}$$

$$= (\text{const.})\cdot(s|X_B)(c_1|s)(c_2|s)\ldots(c_m|s)$$

Using (8-5) and normalizing, this reduces to

$$(s|c_1\ldots c_m) = \frac{(ma)^{c+1}}{c!}\, s^c\, e^{-msa} \qquad (9\text{-}9)$$

where $c \equiv c_1 + \ldots + c_m$ is the total number of counts in the m seconds.

Let's note in passing the properties of this distribution. The most probable, mean, and variance of the distribution (9-9) are respectively

$$\hat{s} = \frac{c}{ma} \qquad (9\text{-}10)$$

$$\bar{s} = \frac{c+1}{ma} \qquad (9\text{-}11)$$

$$\text{var}(s) = \overline{s^2} - \bar{s}^2 = \frac{c+1}{m^2 a^2} = \frac{\bar{s}}{ma} \tag{9-12}$$

So it turns out, as we might have expected, that as $m \to \infty$, the distribution $(s|c_1 \ldots c_m)$ becomes sharper and sharper, the most probable and mean value estimates of s get closer and closer together, and in the limit we would have just a $\delta$-function:

$$(s|c_1 \ldots c_m) \to \delta(s-s')$$

where

$$s' \equiv \lim_{m\to\infty} \frac{c_1 + c_2 + \ldots + c_m}{ma} \tag{9-13}$$

So, in the limit, Mr. B does acquire exact knowledge of the source strength.

Returning to (9-8), both factors in the integrand are now known from (8-8) and (9-9), and so

$$(n_k|c_1 \ldots c_m) = \int_0^\infty \frac{e^{-s(1-a)}[s(1-a)]^{n_k-c_k}}{(n_k-c_k)!} \frac{(ma)^{c+1}}{c!} s^c e^{-msa}\, ds$$

or

$$(n_k|c_1 \ldots c_m) = \frac{(n_k-c_k+c)!}{(n_k-c_k)!\,c!} \frac{(ma)^{c+1}(1-a)^{n_k-c_k}}{(1+ma-a)^{n_k-c_k+c+1}} \tag{9-14}$$

which is the promised generalization of (8-35). In the limit $m \to \infty$, $c \to \infty$, $(c/ma) \to s' = \text{const.}$, this goes into the Poisson distribution

$$(n_k|c_1 \ldots c_m) \to \frac{e^{-s'(1-a)}}{(n_k-c_k)!}[s'(1-a)]^{n_k-c_k} \tag{9-15}$$

which is identical with (8-8). We therefore confirm that, given enough additional data, Mr. B's standard deviation can be reduced from 26 to 10.8, compared to Mr. A's 31.

For finite m, the mean value estimate of $n_k$ from (9-14) is

$$\bar{n}_k = c_k + \bar{s}(1-a) \tag{9-16}$$

where $\bar{s} = (c+1)/ma$ is the mean value estimate of s from (9-11). Equation (9-16), which is to be compared to (8-9), includes (8-21) and (8-38) as special cases. Likewise, the most probable value of $n_k$ according to (9-14), is

$$\hat{n}_k = c_k + \hat{s}(1-a) \tag{9-17}$$

where $\hat{s}$ is given by (9-10).

Note that Mr. B's revised estimates in problem 2 still lie within the range of reasonable error assigned by Mr. A. It would be rather disconcerting if this were not the case, as it would then appear that probability theory is giving Mr. A an unduly optimistic picture of the reliability of his estimates. There is, however, no theorem which guarantees this; for example, if the counting rate had jumped to $c_2 = 80$, then Mr. B's revised estimate of $n_1$ would be far outside Mr. A's limits of reasonable error. But in this case, Mr. B's common sense would lead him to doubt the reliability of his prior information $X_B$; we would have another example like that in Lecture 6, of a problem where one of those alternative hypotheses down at -100 db, which we don't even bother to formulate until they are needed, is resurrected by very unexpected new evidence.

## 9.3. Comparison of Bayesian and Orthodox Results.

Well, in the last lecture I said I was going to compare the results of Bayes' theorem with those obtained by the orthodox statistician's methods in this problem. I have already done that in the case of Mr. A; for his most probable values of $n_1$ and $n_2$ were in all cases just the same as the direct maximum likelihood estimates. The statistician accepts Bayes' theorem in the initial example where the source strength was known. He rejects it in the problem where the source strength was unknown, and says that (Wald, 1941): "These problems cannot be solved by any theorems of the calculus of probabi-

lities alone. Their solution requires some additional principles besides the axioms on which the calculus of probabilities is based." The new principle which he introduces is maximum likelihood; but mathematically, he ends up doing exactly what he would have done if he had stayed with Bayes' theorem. In order to form some idea of the degree of reliability of the estimate, he introduces still another ad hoc principle, the confidence interval. Our robot obtains all of these results automatically, by application of a single principle which is contained in the calculus of probabilities, as formulated by Laplace.

But how does this comparison look in the case of Mr. B? We have seen how Bayes' theorem automatically "digests" his qualitative prior information: $X_B$ = "there is a constant but unknown source strength s," and how it enables him to improve his estimates and lower his probable error. How would the orthodox statistician make use of this information? In the first place, his ideology forbids him to use any of the equations (8-22), (8-23), (8-32), (8-36), (8-41), (9-8), (9-9) which formed the backbone of our various derivations, for he contends that "Probability statements can be made only about random variables. It is meaningless to speak of the probability that s lies in a certain interval, because s is not a random variable, but only an unknown constant." According to his doctrines, the distinction between a "random" and a "nonrandom" quantity is very essential; the methods he will use for inference, (and the conclusions he will arrive at,) depend on his decision as to which quantities are random, which are not.

I want to point out some difficulties with this position in a minute; however, right now our job is not to criticize the orthodox statistician's methods, but to describe them. If he refuses to use Bayes' theorem in the way our robot did, how would he handle it? I can't really be sure; and in fact I'll wager that different statisticians would handle it in different ways, because orthodox teaching has just not produced any unique method for such

problems. But I think I can suggest one ad hoc procedure that he might invent, and which most of his colleagues would accept. Consider the problem where we know that $c_1 = 10$, $c_2 = 16$. If anyone were to refuse to use the prior information $X_B$, on the grounds that it does not consist of frequency data, then he would have little choice but to estimate $n_1$ and $n_2$ by direct maximum likelihood, i.e., by maximization of $(c_1|n_1)$ and $(c_2|n_2)$; and it would collapse back to the problem of Mr. A. But, as I said in Lecture 4, if we do have prior information which is clearly relevant to the problem, common sense will tell all but the most pedantic not to use direct maximum-likelihood estimation. Without departing from orthodox principles, one can use the prior information $X_B$ to formulate the problem in a different way. Here is one possible line of reasoning that he might use.

"The unknown constant s determines the objective statistical properties of n and c; i.e., the relative frequencies with which the random variables n and c would assume various values in the long run. Therefore, if I knew the value of s, it would be perfectly legitimate to use Bayes' theorem in the form

$$(n_1|c_1c_2s) = (n_1|s)\,\frac{(c_1c_2|n_1s)}{(c_1c_2|s)} \tag{9-18}$$

since every probability here has a clear frequency interpretation. Furthermore, since

$$(c_1c_2|n_1s) = (c_1|c_2n_1s)(c_2|n_1s) = (c_1|n_1)(c_2|s)$$

and

$$(c_1c_2|s) = (c_1|c_2s)(c_2|s) = (c_1|s)(c_2|s) \quad , \tag{9-19}$$

the calculation would reduce to

$$(n_1|c_1c_2s) = \frac{(n_1|s)(c_1|n_1)}{(c_1|s)} = (n_1|c_1s) \tag{9-20}$$

i.e., if s is known, then knowledge of $c_2$ is not relevant for estimation of $n_1$.

This leads, according to equation (8-9), to the mean-value estimate

$$\bar{n}_1 = c_1 + s(1-a) \ . \qquad (9\text{-}21)$$

Now if I had a reasonable estimate of s, then substituting it into (9-21) should give me an estimate of $n_1$ which is in some sense equally reasonable. So, instead of estimating $n_1$ by direct maximum-likelihood, I'll use an indirect method: first estimate s by maximum-likelihood, and use the result in (9-21)."

From (9-19) and (8-5) we have

$$\log (c_1 c_2 | s) = (c_1 + c_2) \log s - 2sa + (\text{const.})$$

where the (const.) is independent of s. So, the maximum-likelihood estimate of s, given $c_1$ and $c_2$, is found from $\frac{d}{ds} \log (c_1 c_2 | s) = 0$, or

$$(s)_{\text{max. likelihood}} = \frac{c_1 + c_2}{2a} = \frac{10 + 16}{2 \times 0.1} = 130$$

and his estimate of $n_1$ is then

$$\bar{n}_1 = 10 + 130(1 - 0.1) = 127 \ , \qquad (9\text{-}22)$$

which is the same as Mr. B's most probable value (8-37)! The fact that these estimates turn out exactly the same is, of course, fortuitous; but we see from equations (9-10) and (9-17) that in this problem the agreement would still hold no matter how many counts $\{c_1, c_2, \ldots, c_m\}$ had been observed.

This comparison shows how, in practice, the orthodox statistician who uses a little common sense in formulating the problem, can often manage to get very acceptable results and make use of his prior information without ever using a probability for a "nonrandom" quantity. But if now we asked him to make some definite statements about the reliability of the estimate (9-22), he would be faced with a quite sticky problem. He would probably set up a confidence interval to describe the uncertainty in s; but then he would have to find some way of "folding" this uncertainty with the uncertainty in $n_1$, inherent in (9-21) even when s is known exactly. I will not presume to guess how he would do this; again, since orthodox teaching has produced no unique

way of handling such problems, we can be pretty sure that different workers would do it in different ways, and come out with different conclusions. With reasonable common sense, however, the orthodox conclusions would not differ greatly from the ones our robot obtains from the posterior probability distribution $(n_1 | c_1 c_2 X_B)$. From a purely pragmatic standpoint, which sees no value in the fact that the robot's method comes from a more general and unified set of basic principles, the robot's procedure still has the advantage that he obtains all of these results from a single elementary calculation.

There is a further point which should be made on these estimation problems. We have seen that the most probable value and the mean value estimates are not the same in general. Which is best? The answer, evidently, depends on the use to be made of the theory, and on the form of the posterior probability distribution. For example, in Figure (9.1a) we have a distribution for which the most probable value is not only intuitively a poorer estimate than the mean value, but is also very unstable; very small changes in the data could tilt the curve the other way, making a large change in the estimate, which seems like a clear violation of common sense. But in Figure (9.1b) we have a case where the most probable value is quite likely to be the correct one, while the mean value is known to be an impossible one. In all cases, however, the mean value is the estimate which minimizes the expected square of the error. Generally, if the distribution has a single peak, the mean value would seem preferable. At any rate, any principle which denies us the choice between them cannot possibly be the best in all cases. We are concerned here with value judgments rather than inference; this will be studied in more detail when we consider decision theory.

In summary, what can we now say about the principle of maximum likelihood? If you ask a statistician about these things, one answer you are likely to get is that the real justification of maximum likelihood is not found in problems

of the sort just examined, but in its asymptotic properties, as we accumulate more and more random data. But, of course, in that limit the various "laws of large numbers" guarantee that all these methods approach the same thing. Indeed, in the "large sample" limit the evidence stares you in the face, and anybody can see what general conclusions are indicated, with hardly any need for a formal statistical theory. Scientists and engineers have been getting along fairly well for a long time without statistical training, for just that reason. It is in the small and medium sample case we considered here, that our unaided common sense lacks sufficient discrimination, and we need the guidance of a mathematical theory in order to make definite and defensible judgments.

In any event, whatever desirable properties maximum likelihood might have, asymptotic or otherwise, are also enjoyed by Bayes' theorem with uniform prior probabilities, because mathematically they amount to the same thing. But it

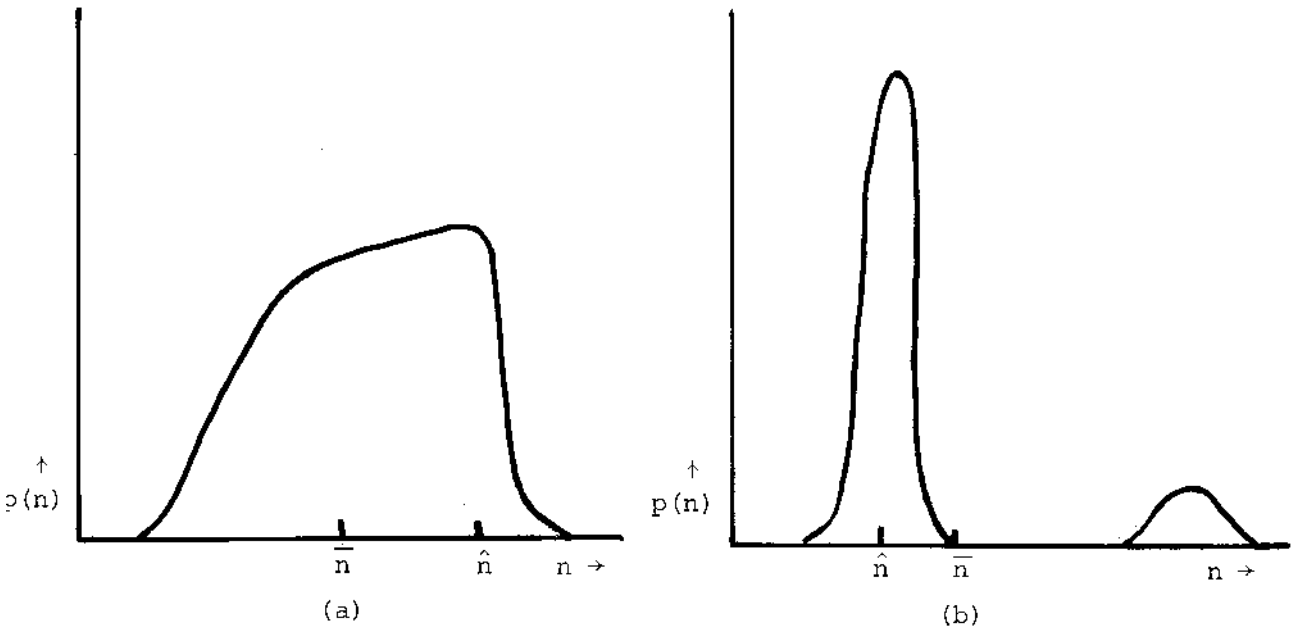


Figure 9.1. (a) A likelihood function for which the maximum-likelihood estimate is not a reasonable one. (b) A case where the maximum-likelihood estimate is more reasonable than the mean value estimate.

is still best to interpret the rules as an application of Bayes' theorem, for the following reason. Statisticians are well aware that the maximum-likelihood estimate may be very poor in the small-sample case. But these are just the cases in which situations like that depicted in Figure (9.1a) arise. In the small sample case, the Bayesian mean-value estimate (i.e., the "center of gravity" of the likelihood function) is often far more reasonable than the maximum-likelihood estimate.

It seems to me that we have to conclude from this that there is no sound reason for ever introducing the notion of maximum likelihood as a separate principle. It is already contained in Bayes' theorem as a special case, and whenever it is the appropriate method to use, Bayes' theorem will reduce to maximum likelihood. From this point of view, we will see later (when we take up decision theory) that it is possible to define precisely the conditions in which maximum likelihood is the optimal procedure [see Sec. 13.5.].

## 9.4. The Trouble with "Random" Variables.

Now let's take a glimpse at some of the difficulties that face the orthodox statistician because of his belief that use of probability theory requires us to distinguish between random and nonrandom quantities. In the example just studied, he didn't face any serious impediment because in this problem there was really no difficulty in deciding intuitively that s is a "constant", while n and c are "random variables". There is little danger that anyone would make a different decision. But there are other problems of inference in which it is not at all clear how this distinction is to be drawn. We will study some cases of this in detail when we take up linear regression (which means simply: fitting the best straight line to a plot of experimental points). This is probably the most common of all statistical problems faced by the experimental scientist; yet it is in just this problem that the distinction

between random and nonrandom quantities is so obscure that you sometimes have to resort to black magic to draw any distinction at all.

This situation has led to some really hilarious proposals for data reduction, solemnly advocated in the orthodox literature. Here is one way it can happen: the abscissa of our graph represents some physical quantity that has a true value X; but this is unknown because the value x actually read from a meter suffers from some experimental error $\varepsilon = x - X$. Nobody ever doubts that $\varepsilon$ is random; but then which of the quantities x, X is random?

To change from one value of X to another, the experimenter typically turns a knob on his apparatus. According to some orthodox writers (Berkson, 1950; Mandel, 1964; Chap. 12), if he turns it without particularly noticing just where the "x-meter" ends up, then X is an unknown constant, and x a random variable. Orthodox theory then tells us how to analyze the data.

But another experimenter, even though he turns the knob in exactly the same way and stops at exactly the same place, does so with the conscious intention of stopping when the meter reads the value x. In this case, we are told, x is the "constant," and X the "random variable". Although there is absolutely no difference in the physical conditions of the experiment, orthodox teaching then tells us that we should analyze the data in an entirely different way, which can lead to different estimates of the slope and intercept of that line, and to widely different conclusions about the reliability of those estimates. If this isn't black magic, I would like to know what it is.

If, now, the second experimenter flips a coin to help him decide at what value of x to set the knob, then both x and X become random variables; and orthodox theory says we should use still a third method of data analysis, leading to a third set of conclusions!

I think most of us are persuaded that the import of the experiment ought to depend on this: how were the knobs actually turned, and what data resulted?

It does not depend on what thoughts flitted through anybody's imagination while turning the knobs; a given experimental procedure and resulting data have exactly the same import whether the knobs were turned by a man or a chimpanzee.

Orthodox theory fails to meet this rather elementary desideratum; if you give an orthodox statistician only the actual procedure and the actual data, plus one of the usual hypotheses about the errors, he has no definite way of getting started on the problem, because for him it is taboo to write down any probability distribution $p(x)$ unless it has been established that $x$ is random; and this information gives him no basis for deciding which quantities are random. Although common sense tells you it cannot be relevant, he wants to know also something about the "state of mind" of the experimenter; and his final conclusions will depend on this. The fact that orthodox practice has to invoke psychokinesis in order to set up some problems hardly supports the claim (Bross, 1963) that orthodox methods, unlike Bayesian, are "objective" and "fact-oriented."

The Bayesian analysis does conform to our desideratum, because it is liberated from that taboo, and therefore has no need to draw artificial distinctions which have nothing to do with the physical conditions of the experiment. Given the above information, our robot can proceed immediately with definite calculations; he is not afraid to introduce probability distributions for any quantity about which he needs inference, and the question whether it is or is not "random" just never comes up at all. Because of his liberation from a taboo that has no justification and serves no purpose, probability theory is, for our robot, an enormously more powerful mathematical reasoning device than it is for one whose ideology forbids the use of that mathematics in its full generality. We will see some spectacular examples of this later when we compare Bayesian and orthodox significance tests and inter-

val estimation methods.

But orthodox taboos can lead to even worse consequences. They force one to attach such supreme importance to this random-nonrandom distinction that, in addition to introducing irrelevancies, many writers will not hesitate to throw away practically all the relevant data of a problem, in order to achieve the situation of "independent random errors" which their theory presupposes. For example, in the problem of fitting a straight line to experimental points, if there is cumulative error (i.e., the error in one value $x_i$ is propagated into all subsequent $x_j$, $j > i$) Mandel (1964; Chap. 12) advocates that we estimate the slope of the line using only the first and last points; and simply throw away all the intermediate ones! To our robot--and also to the poor experimenter who labored to get the data--this is a far graver offense against reason than merely dabbling in a little black magic. As we will see later, throwing away the highly relevant evidence of the intermediate points can increase the probable error of your estimate by more than an order of magnitude in real problems.

The Bayesian analysis never requires us to do such absurd things, because it contains no artificial presuppositions about "randomness". If there is cumulative error, that is just an additional mathematical detail that Bayes' theorem takes into account without any difficulty, while retaining all of the relevant evidence.

Yet in spite of all this emphasis on the necessity of specifying the "random" quantities, no worker in probability theory, orthodox or otherwise, has produced any definition of "random variable" which could actually be applied in real life situations. Here is, for example, a quotation from the book of Savage (1954; p. 45): "The concept of a random variable enters into almost any discussion of probability. Experts are fairly well agreed on the following definition. A random variable is a function x attaching a value

x(s) in some set X to every s in a set S on which a probability measure P is defined." Definitions essentially equivalent to this can be found in most of the modern books on statistics. While this may be fine for setting up an abstract mathematical theory, the most obvious thing about it is that the definition is absolutely useless in helping us decide whether some specific quantity, such as the number of beans in a can, is or is not "random".

If you read the literature carefully, I think you will see that whenever the orthodox statistician gets down to a very specific problem, he uses the word "random" merely as shorthand for "likely to be different in different situations." In Laplace's theory there is no need to emphasize, or even to define, any sharp distinction between random and nonrandom quantities, for the commonsense reason that in the specific problem at hand, the quantity I am reasoning about (in the problem just discussed, $n_1$) is <u>always</u> simply a definite, but unknown number. Whether this number would or would not be the same in some other situation that I am <u>not</u> reasoning about, is just not relevant to my problem; to adopt different methods on such grounds is to commit the most obvious inconsistency of reasoning.

All right, I hope this little excursion into polemics has given you a clearer understanding of why, in the theory we are developing, the word "random" just doesn't appear; and of the kind of troubles we would get into if we did try to use it. In the next lecture, I want to return to the constructive development of the theory.