Lecture 8

POINT ESTIMATION WITH BINOMIAL AND POISSON DISTRIBUTIONS

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

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

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

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

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

8.1. A Simple Bayesian Estimate: Quantitative Prior Information.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

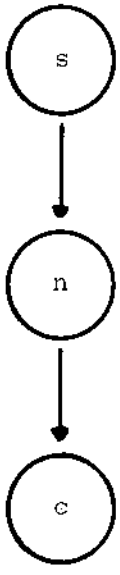

Figure 8.1. Direction of causal influences.

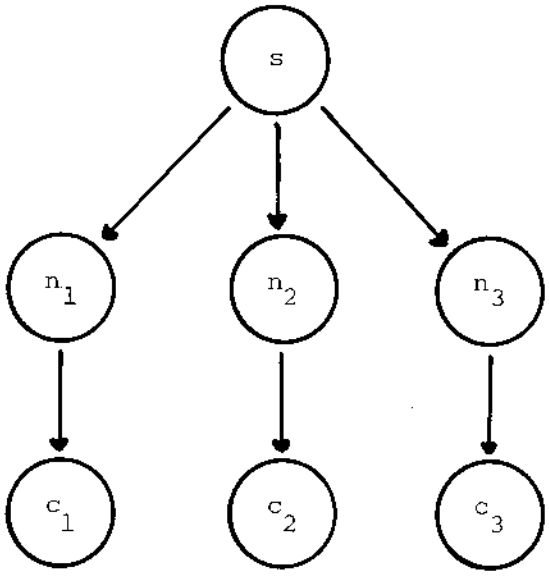

Figure 8.2. Causal influences in successive measurements.

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

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

or,

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

This is a Poisson distribution with mean value

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

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

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

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

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

have to substitute:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

or

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

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

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

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

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

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

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

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

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

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

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

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

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

8.2. Effect of Qualitative Prior Information.

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

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

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

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

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

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

been the point of controversy for a good long time, for in any frequency theory of probability, we certainly have no basis at all for assigning the probabilities $(n_1|X_A)$.

Now, of course, Mr. A is going to assign a uniform prior probability here, and our statistician friend will object on the grounds that this is a completely unwarranted assumption. He will say, "How do you know that all values of n_1 are equally likely? They might not be equally likely at all. You just don't know, and you have no basis for applying Bayes' theorem until you have found the correct prior probability distribution." Note that this is not because our friend has any particular dislike for a uniform distribution; for he would object just as strongly (and in fact, I suspect, even more strongly) to any other prior probability assignment we might propose to use. It would always seem, to him, like an unwarranted assumption which would invalidate all our conclusions.

I am belaboring this point because it lies at the heart of the most persistently held misconception about the Laplace-Bayes theory. Unless we understand clearly what we're doing when we assign a uniform prior probability, we're going to be faced with tremendous conceptual difficulties from here on. This is what Mr. A replies to the statistician:

"Your objection shows that the word 'probability' has entirely different meanings to you and me. When you say that I cannot apply Bayes' theorem until I have determined the 'correct' prior probability distribution, you are implying that the event n_1 possesses some intrinsic 'absolute' probability. I deny this. n_1 is what it is; simply an unknown number. The only meaning of the word 'probability' which makes any sense at all to me, is simply the best indication of the truth of a proposition, based on whatever evidence we do <u>in fact</u> have. To me, a probability assignment is not an assertion about experience, real or potential. When I say, 'the probability of event E is p,' I

am not describing any property of the event. I am describing my state of knowledge concerning the event.

"Now, evidently, each of us believes that the other is suffering from a very fundamental and dangerous confusion about the proper use of probability theory. But we can never settle this by philosophical arguments about the meaning of words. The only real way of settling the question, which of these conceptions of probability is best, is to put them to the test in specific problems. You say that my uniform prior probability assignment is foolish. If so, then it ought to lead to at least one foolish result. So I'm just going to ignore your warning and go ahead with my calculation. If I get a foolish result, then from studying how it happened, I can learn something. But if I get a sensible result, then maybe you are the one who can learn something.

"According to Bayes' theorem, I need to find the probability assignment $(n_1|X_A)$ which represents my state of knowledge before I observed that $c_1 = 10$ counts. At that time, n_1 might have been 0, 1, 137, 2069, or 10^5 for all I knew. There was nothing in my prior knowledge which would justify saying that any one of those was more likely than any other, and assigning the same probability to all of them is simply my way of stating that fact. n_1 might easily have been as large as 10^7, for all I knew. But there is some upper limit N, for which I knew that $n_1 < N$. For example, if n_1 had been $10^{10^{10}}$, then not only the laboratory, but our entire galaxy, would have been vaporized by the energy in the beam. I could justify a considerably lower value of N than that, and if it turns out to make a difference in my conclusions, I'll have to think harder about just how low I could take it. But before going to all that work, I'd better find out whether it does make any difference. So, I'll just take

$$(n_1|X_A) = \begin{cases} \frac{1}{N}, & 0 \le n_1 < N \\ 0, & N \le n_1 \end{cases} \tag{8-16}$$

and see what Bayes' theorem gives me."

Well, Mr. A turns out to be lucky, for nicely enough, the 1/N cancels out of Equations (8-14), (8-15), and we are left with

$$(n_1|c_1\ X_A) = \begin{cases} \dfrac{(c_1|n_1)}{\sum_{n_1=0}^{N-1} (c_1|n_1)} & ,\quad 0 \le n_1 < N \\ 0 & ,\quad N \le n_1 \end{cases} \tag{8-17}$$

We have noted, in Equation (8-11), that as a function of n, $(c|n)$ attains its maximum at $n = c/a$ (=100, in this problem). For n large compared to this, $(c|n)$ falls off like $n^c(1-a)^n \simeq n^c\, e^{-an}$. Therefore, the sum in (8-17) converges so rapidly that if N is as large as a few hundred, there is no appreciable difference between

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

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

$$(n_1|c_1X_A) = a\ (c_1|n_1) = \binom{n_1}{c_1} a^{c_1+1}\ (1-a)^{n_1-c_1}\ . \tag{8-18}$$

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

$$(\hat{n}_1)_A = \frac{c}{a} = 100 \tag{8-19}$$

while the mean value estimate is calculated as follows:

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

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

The sum is equal to

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

$$= [1 - (1-a)]^{-c_1-2} = \frac{1}{a^{c_1+2}} \tag{8-20}$$

and, finally, we get

$$(\bar{n}_1)_A = c_1 + (c_1+1)\,\frac{1-a}{a} = \frac{c_1+1-a}{a} = 109 \quad . \tag{8-21}$$

Now, how about the other robot, Mr. B? Does his extra knowledge help him here? He knows that there is some definite source strength s. And, because the laboratory is not being vaporized, he knows that there is some upper limit S_o. Suppose that he assigns a uniform prior probability density for $0 \leq s < S_o$. Then he will obtain

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

$$= \frac{1}{S_o} \int_0^{S_o} \frac{s^{n_1} e^{-s}}{n_1!}\, ds \quad . \tag{8-22}$$

Now, if n_1 is appreciably less than S_o, the upper limit of integration can for all practical purposes, be taken as infinity, and the integral is just unity. So, we have

$$(n_1|X_B) = (s|X_B) = \frac{1}{S_o} = \text{const., if } n_1 < S_o \quad . \tag{8-23}$$

In putting this into Bayes' theorem with $c_1 = 10$, the significant range of values of n_1 will be of the order of 100, and unless S_o is lower than about 200, we will have exactly the same situation as before; Mr. B's extra knowledge didn't help him at all, and he comes out with exactly the same distribution and the same estimates:

$$(n_1|c_1X_B) = (n_1|c_1X_A) = a\,(c_1|n_1) \quad . \tag{8-24}$$

Jeffreys (1939; Chap. 3) has proposed a different way of handling this problem. He suggests that the proper way to express "complete ignorance" of

a continuous variable known to be positive, is to assign uniform prior probability to its logarithm; i.e. the prior probability density is

$$(s|X_J) = \frac{1}{s} \tag{8-25}$$

Of course, you can't normalize this, but that doesn't stop you from using it, because when we expand the denominator of Bayes' theorem as in (8-15), we see that the prior probability appears in both numerator and denominator [the same reason that N cancelled out of (8-17)]. So, in applying Bayes' theorem, it doesn't really matter whether the prior probabilities are normalized or not.

Jeffreys justified (8-25) on the grounds of invariance under certain changes of parameters; i.e. instead of using the parameter s, what prevents us from using $t \equiv s^2$, or $u \equiv s^3$? Evidently, to assign a uniform prior probability density to s, is not at all the same thing as assigning a uniform prior probability to t; but if we use the Jeffreys prior, we are saying the same thing whether we use s or any power s^m as the parameter. There is the germ of an important principle here; but it was only recently that the situation has been fairly well understood. When we take up the theory of transformation groups later on, we will see that the real justification of Jeffreys' rule cannot lie merely in the fact that the parameter is positive; but that our desideratum of consistency in the sense (b) of Lecture 2 (p. 26) uniquely determines the Jeffreys rule in the case when s is a "scale parameter." The question then reduces to whether s can properly be regarded as a scale parameter in this problem. However, this takes us far beyond the present topic, so I don't want to spend a lot of time now arguing either for or against (8-25); but, in the spirit of this problem, we can put it to the test and see what it gives. The calculations are all very easy, and we find these results:

$$(n_1|X_J) = \frac{1}{n_1}, \qquad (c_1|X_J) = \frac{1}{c_1}$$

$$(n_1|c_1X_J) = \frac{c_1}{n_1}(c_1|n_1) \quad . \tag{8-26}$$

This leads to the most probable and mean value estimates:

$$(\hat{n}_1)_J = \frac{c_1 - 1 + a}{a} = 91 \qquad (8\text{-}27)$$

$$(\bar{n}_1)_J = \frac{c}{a} = 100 \quad . \qquad (8\text{-}28)$$

The amusing thing emerges that Jeffreys' prior probability rule just lowers the most probable and mean value by 9 each, bringing the mean value right back to the maximum likelihood estimate!

This comparison is valuable in showing us how little difference there is numerically between the consequences of different prior probability assignments which are not sharply peaked, and helps to put arguments about them into proper perspective. We made a rather drastic change in the prior probabilities, in a problem where there was really very little information contained in the result of the random experiment, and it still made less than 10 per cent difference in the result. This is, as we will see in the next lecture, small compared to the probable error in the estimate which was inevitable in any event. In a more realistic problem where a random experiment is repeated many times to give us a good deal more information, the difference would be very much smaller still. So, from a pragmatic standpoint, the arguments about which prior probabilities correctly express a state of "complete ignorance" usually amount to quibbling over pretty small peanuts.* From the standpoint of principle, however, they are very important and have to be thought about a great deal.

Now we are ready for the interesting part of this problem. For during the next second, we see $c_2 = 16$ counts. What can Mr. A and Mr. B now say about the numbers n_1, n_2, of particles responsible for c_1, c_2? Well, Mr. A has no reason to expect any relation between what happened in the two time

*This is most definitely not true if the prior probabilities are to describe a definite piece of prior knowledge, as the next example shows.

intervals, and so to him the increase in counting rate is evidence only of an increase in the beam intensity. His calculation for the second time interval is exactly the same as before, and he will give as the most probable value

$$(\hat{n}_2)_A = \frac{c_2}{a} = 160 \tag{8-29}$$

and his mean value estimate is

$$(\bar{n}_2)_A = \frac{c_2 + 1 - a}{a} = 169 \quad . \tag{8-30}$$

Knowledge of c_2 doesn't help him to get any improved estimate of n_1, which stays the same as before.

But now, Mr. B is in an entirely different position than Mr. A; his extra qualitative information suddenly becomes very important. For knowledge of c_2 enables him to improve his previous estimate of n_1. Bayes' theorem now gives

$$(n_1|c_2c_1X_B) = (n_1|c_1X_B)\,\frac{(c_2|n_1c_1X_B)}{(c_2|c_1X_B)}$$

$$= (n_1|c_1X_B)\,\frac{(c_2|n_1X_B)}{(c_2|c_1X_B)} \tag{8-31}$$

Again, the denominator is just a normalizing constant, which we can find by summing the numerator. We see that the significant thing is $(c_2|n_1X_B)$. Using our trick of resolving c_2 into mutually exclusive alternatives, this is

$$(c_2|n_1X_B) = \int_0^\infty (c_2s|n_1X_B)\,ds = \int_0^\infty (c_2|sn_1)(s|n_1)\,ds$$

$$= \int_0^\infty (c_2|s)(s|n_1)\,ds \quad . \tag{8-32}$$

We have already found $(c_2|s)$ in Equation (3-7), and we need only

$$(s|n_1) = (s|X_B)\,\frac{(n_1|s)}{(n_1|X_B)} = (n_1|s) \quad , \quad \text{if } n_1 \ll S_o \tag{8-33}$$

where we have used Equation (8-23). We have found $(n_1|s)$ in Equation (8-3), so we have

$$(c_2|n_1X_B) = \int_0^\infty \left[\frac{e^{-sa}(sa)^{c_2}}{c_2!}\right]\left[\frac{e^{-s}s^{n_1}}{n_1!}\right] ds = \binom{n_1+c_2}{c_2} \frac{a^{c_2}}{(1+a)^{n_1+c_2+1}} . \quad (8\text{-}34)$$

Now we just substitute (8-24) and (8-34) into (8-31), carry out an easy summation to get the denominator, and the result is

$$(n_1|c_2c_1X_B) = \frac{(2a)^{c_1+c_2+1}}{(c_1+c_2)!\ (1-a)^{c_1}\ (1+a)^{c_2+1}} \frac{(n_1+c_2)!}{(n_1-c_1)!} \left(\frac{1-a}{1+a}\right)^{n_1} . \quad (8\text{-}35)$$

Note that we could also have derived this by direct application of our trick:

$$(n_1|c_2c_1X_B) = \int_0^\infty (n_1s|c_2c_1X_B)\, ds = \int_0^\infty (n_1|sc_1)(s|c_2c_1)\, ds . \quad (8\text{-}36)$$

We have already found $(n_1|sc_1)$ in (8-8), and it is easily shown that $(s|c_2c_1)$ = (const.) x $(c_2|s)(c_1|s)$, which is therefore given by (8-5). This, of course, leads to the same result (8-35); this provides another test of the consistency of our rules, which we sought to ensure by the functional equation arguments in Lecture 3.

To find Mr. B's new most probable value of n_1, we set

$$\frac{(n_1|c_2c_1X_B)}{(n_1-1|c_2c_1X_B)} = \frac{n_1 + c_2}{n_1 - c_1}\frac{1 - a}{1 + a} = 1 ,$$

or,

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

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

$$= 127 \quad (8\text{-}37)$$

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

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

$$= \frac{c_1 + c_2 + 1 - a}{2a} + \frac{c_1 - c_2}{2}$$

$$= 131.5 \quad . \tag{8-38}$$

You see that both estimates are considerably raised, and the difference between most probable and mean value is only half what it was before. If we want Mr. B's estimates for n_2, then from symmetry we just interchange the subscripts 1 and 2 in the above equations. This gives for his most probable and mean value estimates, respectively,

$$(\hat{n}_2)_B = 135 \tag{8-39}$$

$$(\bar{n}_2)_B = 137.5 \tag{8-40}$$

Now, can we understand what is happening here? Intuitively, the reason why Mr. B's extra qualitative prior information makes a difference is that knowledge of both c_1 and c_2 enables him to make a better estimate of the source strength s, which in turn is relevant for estimating n_1. The situation is indicated more clearly by the diagrams, Fig. (8.2). To Mr. A, each sequence of events $n_i \rightarrow c_i$ is entirely independent of the others, so knowledge of one doesn't help him in reasoning about any other. In each case, he must reason from c_i directly to n_i, and no other route is available. But to Mr. B, there are two routes; he can reason directly from c_1 to n_1 as Mr. A does, as described by $(n_1|c_1X_A) = (n_1|c_1X_B)$; but because of his knowledge that there is a fixed source strength s "presiding over" both n_1 and n_2, he can also reason along the route $c_2 \rightarrow n_2 \rightarrow s \rightarrow n_1$. If this were the only route available to him (i.e., if he didn't know c_1), he would obtain the distribution

$$(n_1|c_2X_B) = \int_0^\infty (n_1|s)(s|c_2X_B)\, ds$$

$$= \frac{a^{c_2+1}}{c_2!\ (1+a)^{c_2+1}} \ \frac{(n_1 + c_2)!}{n_1!\ (1+a)^{n_1}} \tag{8-41}$$

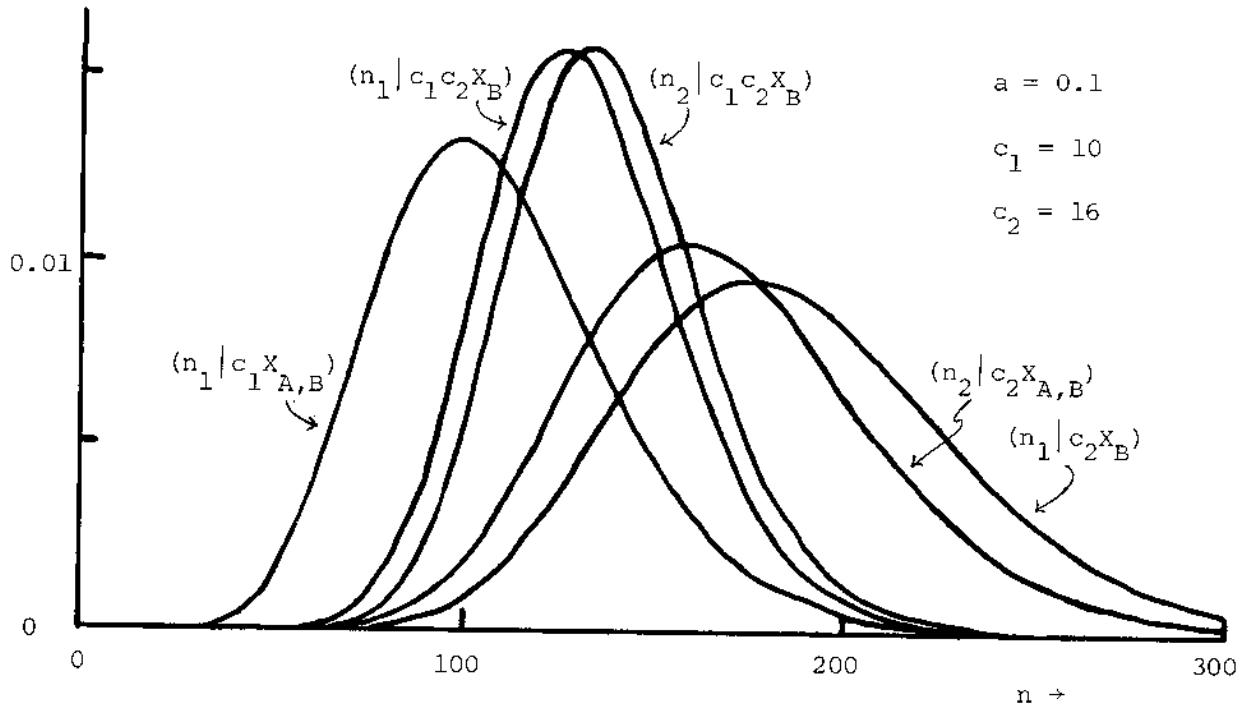

Figure 8.3. The various distributions (8-18), (8-35), (8-41), showing the effect of varying prior information.

and, comparing the above relations, we see that Mr. B's final distribution (8-35) is, except for normalization, just the product of the ones found by reasoning along his two routes:

$$(n_1|c_1c_2X_B) = (\text{const.}) \cdot (n_1|c_1X_B)(n_1|c_2X_B) \tag{8-42}$$

The information (8-41) about n_1 obtained by reasoning along the new route $c_2 \to n_2 \to s \to n_1$ thus introduces a "correction factor" in the distribution obtained from the direct route $c_1 \to n_1$, enabling Mr. B to improve his estimates.

This suggests that, if Mr. B could obtain the number of counts in a great many different seconds, c_3, c_4, ..., c_m, he would be able to do better and better; and perhaps in the limit $m \to \infty$ his estimate of n_1 might become as good as the one we found from Eq. (8-8), in which the source strength was considered known exactly. In the next Lecture we will check this surmise by working out the degree of reliability of these estimates, and by generalizing these distributions to arbitrary m, from which we can obtain the asymptotic forms.