Lecture 6

MULTIPLE HYPOTHESIS TESTING

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

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

6.1. Admitting an Unlikely Hypothesis.

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

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

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

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

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

where

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

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

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

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

- 10 db for A

+ 10 db for B

- 60 db for D

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and so finally we arrive at

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

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

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

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

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

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

These probabilities can be factored by Rule 1:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

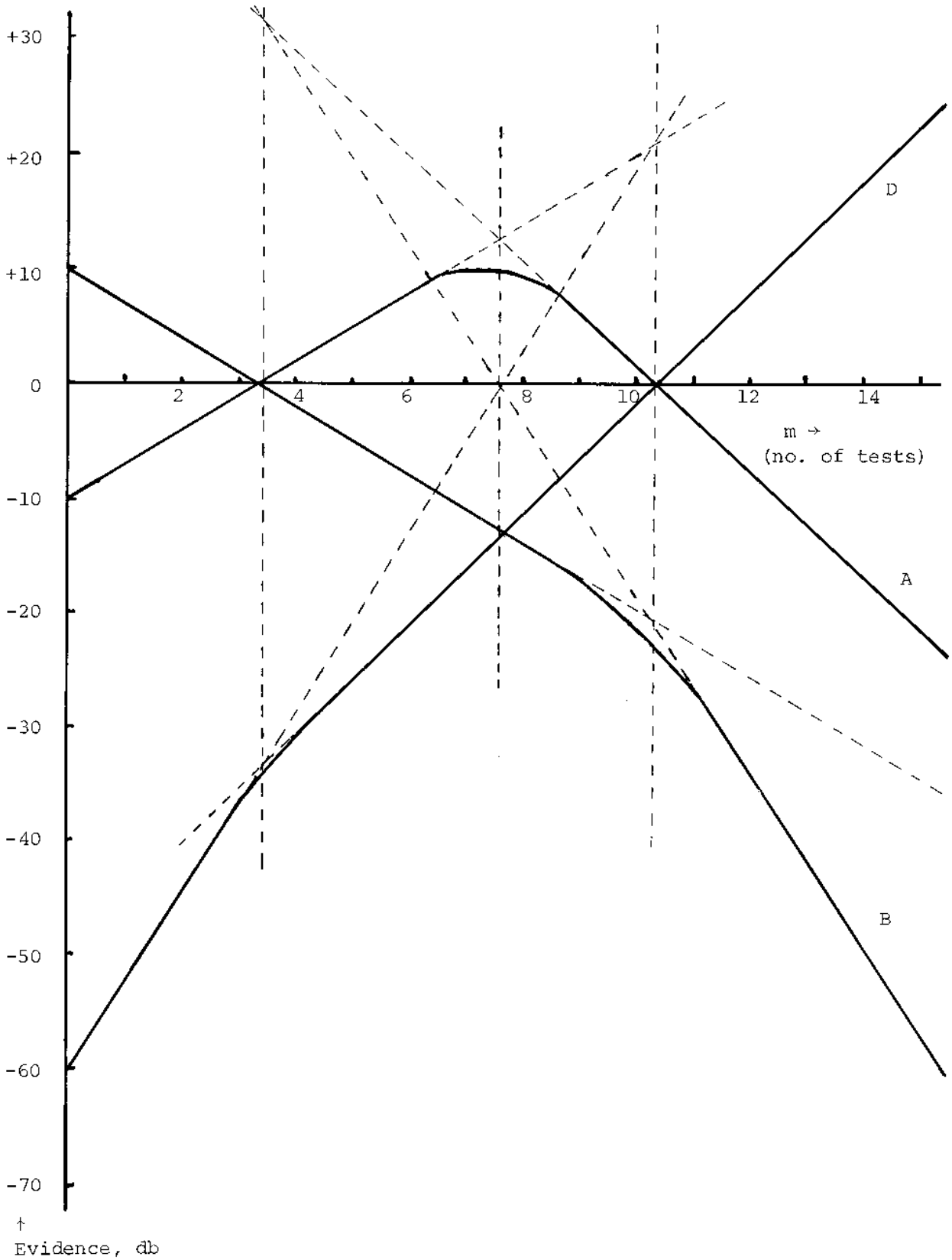

Figure 6.1. Course of a multiple hypothesis test.

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

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

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

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

Good Bad

Figure 6.2. Plausibility flow diagrams.

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

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

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

6.2. Testing an Infinite Number of Hypotheses.

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

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

There is a prior probability density

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

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

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

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

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

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

or,

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

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

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

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

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

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

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

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

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

corresponding to the particular choice of prior density:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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