Lecture 5

SEQUENTIAL HYPOTHESIS TESTING

Our second example of statements made about Bayes' theorem in the literature has been provided by Professor Wm. Feller. On page 85 of his book (Feller, 1950) he writes: "Unfortunately Bayes' rule has been somewhat discredited by metaphysical applications of the type described above.* In routine practice, this kind of argument can be dangerous. A quality control engineer is concerned with one particular machine and not with an infinite population of machines from which one was chosen at random. He has been advised to use Bayes' rule on the grounds that it is logically acceptable and corresponds to our way of thinking. Plato used this type of argument to prove the existence of Atlantis, and philosophers used it to prove the absurdity of Newton's mechanics. In our case it overlooks the circumstance that the engineer desires success and that he will do better by estimating and minimizing the sources of various types of errors in predicting and guessing. The modern method of statistical tests and estimation is less intuitive but more realistic. It may be not only defended but also applied."

Well, that gives us a pretty clear idea of one common attitude toward Bayes' theorem, at least for problems of quality control. Now what are the procedures referred to as the "modern method of statistical tests?" I can't tell of course from reading, but ever since the early days of World War II

---

*The reference is to Laplace's law of succession, about which we will have a lot to say later on in Lecture 16.

when he invented it, Wald's sequential testing procedure (Wald, 1947) has been generally considered the optimum one available, optimum according to several different criteria.

Let's illustrate the problem by considering manufacture of some small item. Suppose we take crystal diodes. One of the important things about a crystal diode is the maximum inverse peak voltage it can stand without damage. Clearly, the way to find out just how good our diodes are is to test each one and measure the voltage at which damage occurs. The trouble is that once we've done this the diode is ruined, so we can't test every one this way. We can test only some fraction of the batch and we would not want to test a very large fraction. So the problem of quality control in this case is to find some method of plausible reasoning which lets us do the best possible job of deciding whether we have a good batch or not, with the smallest number of diodes ruined in testing. I think all statisticians agree that Wald's method is the optimum one in this sense of requiring, on the average, fewer tests than any other for a given probability of error. Wald, in a footnote in his book, says that he conjectures that it's an optimum test in this sense but didn't succeed in proving it. We'll come back to that statement a little later.

Just for variety, let's go first into the way the robot would handle this problem. We will simply ignore Feller's warning, and see for ourselves whether Bayes' theorem can be "applied." After the final comparisons are at hand, we will also see whether it can be "defended."

5.1 Logarithmic Form of Bayes' Theorem.

First, let's manipulate Bayes' theorem a little bit in a manner suggested by I. J. Good (Good, 1950). Instead of calculating the probability, it would be just as good if we'd calculate any monotonic function of the

probability, if we know what function we've got. So, let's do a little rebuilding on Bayes' theorem. I'll use E to stand for new evidence.

$$(A|EX) = (A|X)\,\frac{(E|AX)}{(E|X)}\quad . \tag{5-1}$$

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

$$(a|EX) = (a|X)\,\frac{(E|aX)}{(E|X)}\quad , \tag{5-2}$$

and we can take the ratio of the two equations:

$$\frac{(A|EX)}{(a|EX)} = \frac{(A|X)\,(E|AX)}{(a|X)\,(E|aX)}\quad . \tag{5-3}$$

In this case, one of our terms will drop out. This doesn't look like any particular advantage. But the quantity that we have here, the ratio of the probability that A is true to the probability that it's false, has a technical name. We call it the "odds" on the proposition A. So if I write the "odds of A, given E and X," as the symbol

$$O(A|EX) \equiv \frac{(A|EX)}{(a|EX)} \tag{5-4}$$

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

$$O(A|EX) = O(A|X)\,\frac{(E|AX)}{(E|aX)} \tag{5-5}$$

The odds on A are equal to the prior odds multiplied by the ratio of the probability that E would be seen if A was true, to the probability that E would be observed if A was false. The odds are, of course, a monotonic function of the probability, so we could equally well calculate these quantities.

In some applications it is even more convenient to take the logarithm of the odds because of the fact that we can then add up terms--the same reason the logarithm was invented in the first place. Now we could take logarithms to any base we want. What I'm after here is something which is handy for numerical work, and the base 10 turns out to be easier to use

than the base e for that purpose, even though it makes our equations look less elegant. And so I'm going to define a new function which I'll call the evidence for A given E:

$$e(A|EX) \equiv 10 \log_{10} O(A|EX) \ . \qquad (5\text{-}6)$$

This is still a monotonic function of the probability. By using the base 10 and putting the factor 10 in front, we've now reached the condition where we're measuring evidence in decibels! Now what does Bayes' theorem look like? The evidence for A, given E, is equal to the prior evidence plus the number of db provided by working out the probability ratio in the second term below:

$$e(A|E) = e(A|X) + 10 \log_{10}\left[\frac{(E|A)}{(E|a)}\right] . \qquad (5\text{-}7)$$

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

$$E = E_1 \, E_2 \, E_3 \ldots$$

In that case, we could expand this a little more by successive applications of Rule 1:

$$e(A|E) = e(A|X) + 10 \log_{10}\left[\frac{(E_1|A)}{(E_1|a)}\right] + 10 \log_{10}\left[\frac{(E_2|E_1A}{(E_2|E_1a}\right] + \ldots \qquad (5\text{-}8)$$

In a lot of cases, it turns out that the probability of E_2 is not influenced by knowledge of E_1. For example, in the case where one says technically the probability is a chance; say the tossing of a coin, where knowing the result of one toss (if you know the coin is honest) doesn't influence the probability you would assign for the next toss. In case these several pieces of evidence are independent, the above equation becomes:

$$e(A|E) = e(A|X) + 10 \sum_i \log_{10}\left[\frac{(E_i|A)}{(E_i|a)}\right] \ , \qquad (5\text{-}9)$$

where the sum is over all the extra pieces of information we get.

Now it would be a good idea for us to get some feeling for numerical values here. So, I'd like to give a table and a graph. We have here three different ways we can measure plausibility; evidence, odds, or probability; they're all monotonic functions of each other. Zero db of evidence corresponds to odds of 1 or to a probability of 1/2. Now every electrical engineer knows that 3 db means a factor of 2 and 10 db is a factor of 10, and so if we just go up in steps of 3 db, or 10, why we can write down this table pretty fast.

e	O	p
0	1:1	1/2
3	2:1	2/3
6	4:1	4/5
10	10:1	10/11
20	100:1	100/101
30	1000:1	0.999
40	10^4:1	0.9999
-e	1/O	1-p

You see here why giving evidence in db is nice. When probabilities get very close to one or very close to zero, our intuition doesn't work very well. Does the difference between the probability of 0.999 and 0.9999 mean a great deal to you? It certainly doesn't to me. But after living with this for a while, the difference between evidence of plus 30 db and plus 40 db <u>does</u> mean something to me. It's now in a scale which my mind can comprehend. This is just another example of the Weber-Fechner law. Now let's draw a graph showing reasonably well the numerical values of evidence versus probability. This graph is shown in Figure (5.1). The graph is symmetric about

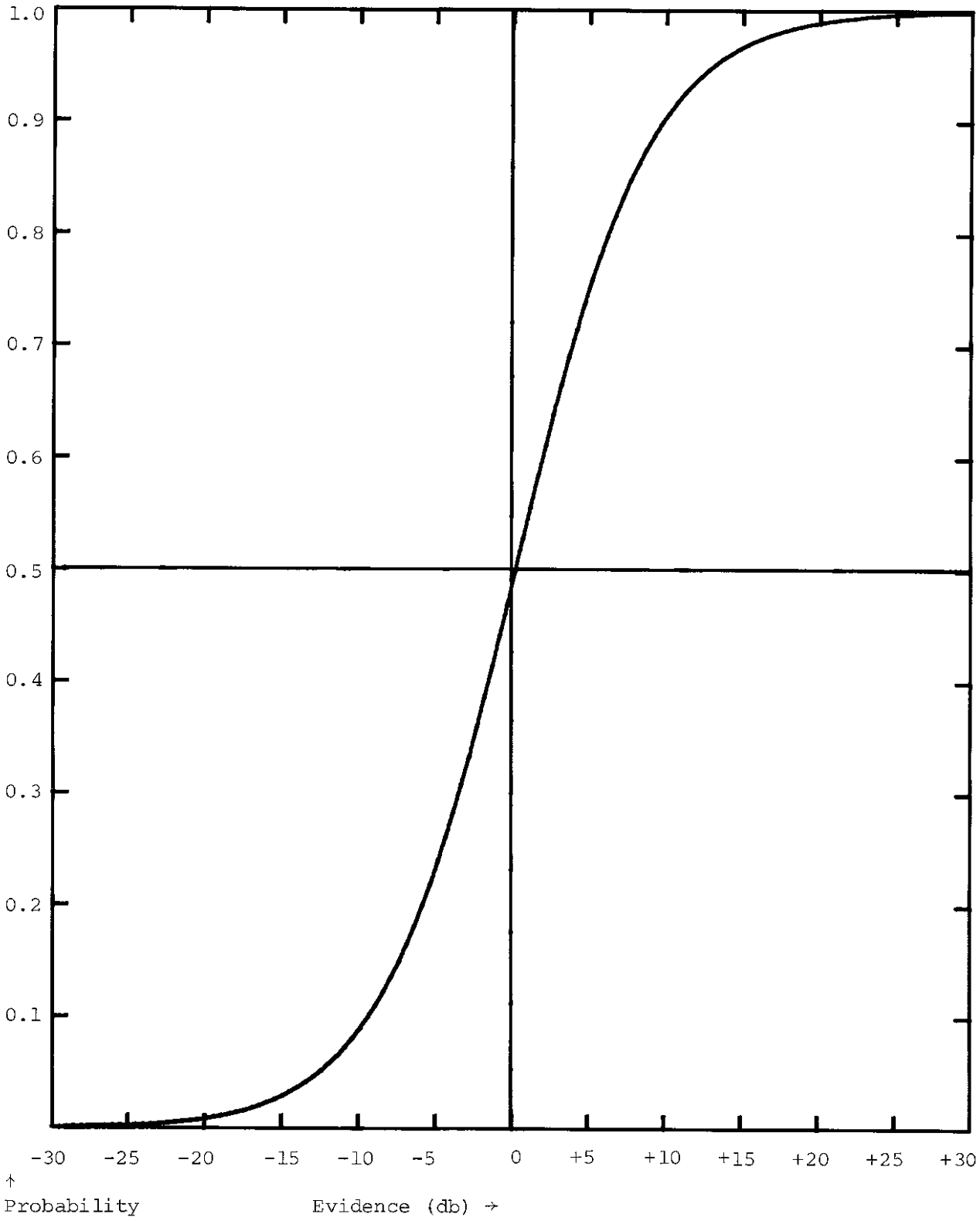

Figure 5.1. Probability vs. Evidence.

the center.

Now let's take our specific example of quality control. I'll assume numbers which are not at all realistic in order to bring out some points a little bit better. We have eleven automatic machines which are turning out crystal diodes. This example corresponds to a very early stage in the development of crystal diodes, because ten of the machines produce, on the average, one in six defective. The eleventh machine is even worse; it makes one in three defective. The output of each machine is collected in an unlabeled box and stored in the warehouse. We choose one of the boxes and we test a few of the diodes. Our job is to decide whether we got a box from the bad machine or not; that is, whether we're going to accept this batch or reject it. Now we're going to turn this job over to our robot and see how he handles it.

He says: "If we want to make judgments about whether we have the box of defective diodes, the way to do this is to calculate the probability that we have the box of defective diodes, conditional on all the evidence available." Let's say the proposition A shall stand for the statement "we chose the bad box." All right, what is the initial evidence for proposition A? The only initial evidence is that there are eleven machines and we don't know which one we got; so by Rule 4 $(A|X) = 1/11$, and by Rule 2 $(a|X) = 1 - (A|X) = 10/11$. Therefore,

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

$$= -10 \text{ db} \qquad (5\text{-}10)$$

Evidently, the only property of X that's going to be relevant to this problem is just this number, - 10 db. Any other kind of prior evidence which led to the same initial probability assignment would give us exactly the same mathematical problem from this point on. So, it isn't really necessary

to say we're talking only about a problem where there are eleven machines, and so on. There might be only one machine, and the prior evidence consists of our previous experience with it. My reason for stating the problem in terms of eleven machines was just that we have, so far, only one principle, Rule 4, by which we can convert raw information into numerical values of probability. I mention this here only because of Professor Feller's remark about a single machine. To our robot, it doesn't make any difference how many machines there are; the only thing that counts is the prior probability, however arrived at.

Now from this box we take out a diode and test it to see where it breaks down. Every time we pull out a bad one, what will that do to the evidence? That will add to this the number

$$10 \log_{10} \frac{(\text{bad}|A)}{(\text{bad}|a)} \tag{5-11}$$

where $(\text{bad}|A)$ represents the probability of getting a bad diode, given A, etc. We have, then, to determine these probabilities.

If we have the box in which one in three are bad, what is the probability that we will draw a bad one? The final answer is obvious to all of us without any calculation, and the argument showing this from the principles of probability theory is almost trivial. Nevertheless, I want to give that argument in full because there is a very important general principle lurking here, which will apply in countless other applications of probability theory.

5.2. Sampling With and Without Replacement.

Consider first the traditional "urn" of probability theory, in which we have placed N balls, all of the same size, weight, surface texture, etc., labeled 1, 2, ..., N. Balls 1, 2, ..., n are black, and the remaining (N-n) are white. What is the probability of drawing blindfolded any parti-

cular ball, say the i'th? Rule 4 answers this, for there are N mutually exclusive possibilities, and the information given provides no justification for expecting any one of them in preference to any other. In this state of knowledge, therefore, the probability sought must be $p_i = 1/N$.

Let us recall clearly just what this means. The probability assignment $p_i = 1/N$ is not an assertion of any physical property of the balls; it is merely a means of describing the state of knowledge of the robot prior to the drawing. It is, therefore, utterly meaningless to speak of "verifying" this probability assignment by performing any experiment on the balls; that would be exactly like trying to verify a boy's love for his dog by performing experiments on the dog. What it does mean was explained in our derivation of Rule 4; the assignment $p_i = 1/N$ is uniquely determined by the requirement that the robot's reasoning be consistent in the sense that, in two problems where he has the same state of knowledge, he must assign the same probabilities. If he were to assign anything different from the uniform distribution, then merely by a permutation of labels we could exhibit a second problem in which the robot's state of knowledge is exactly the same; but in which he is assigning a different probability distribution. I have repeated this argument for emphasis, because to the best of my knowledge, this point is not recognized in any other work on probability theory.

Now, what is the probability that we shall draw a black ball? Since different balls are mutually exclusive possibilities, Rule 3 as extended to Eq. (3-21) applies, and the probability of drawing a black one is the sum

$$(\text{black}|X) = \sum_{i=1}^{n} p_i = n/N \tag{5-12}$$

i.e., it is just the fraction of black balls in the urn. It is, therefore, also equal to the relative frequency with which we would draw black balls, if we took them all out; or as it is usually stated, if we "sampled the entire population."

We have here one of the many different connections between probability and frequency. In spite of the triviality of its derivation, I ask you to note carefully just how it came about; because today most writers on probability and statistics deny that probability theory has anything to do with plausible reasoning, and insist that the only proper meaning of probability is that of relative frequency in some "random experiment." According to this school of thought, if a probability is not a frequency, then it is not "objective," and its use is just not scientifically respectable.

On the other hand, I maintain that, as its derivation shows, the relation (5-12) has absolutely nothing to do with the definition of probability; on the contrary, it is an almost trivial mathematical consequence of probability theory interpreted as the "calculus of inductive reasoning." In fact, by this broader interpretation of the theory, we lose none of the usual connections between probability and frequency; as will become clear gradually in the remaining lectures, every connection between probability and frequency that is actually used in applications, is deducible in a similar way as a consequence of our "inductive reasoning" form of the theory.

At this point, you might ask, "Aren't you making a tempest in a teapot? Since on either viewpoint we end up writing down the same equation (5-12), which was obvious intuitively without any derivation at all, what difference do these philosophical questions make? It seems like pedantic nit-picking." Well, it is true that in many problems the connection between probability and frequency is so close that the notions are easily confused, and this confusion does no harm in the pragmatic sense that we end up writing down the same equations. Usually, the importance of my nit-picking does not lie at all in the actual equations used; it lies in our judgment about the range of validity of those equations.

The point is that many of the most important problems of current science and engineering are just problems of inductive reasoning, in which no "random experiment" is involved in any way. If you insist that a probability is not respectable unless it is also a frequency, then you will have to conclude that probability theory is just not applicable to these problems. But I am going to insist in these lectures that the relations of probability theory are perfectly valid when used in the Laplace sense of the "calculus of inductive reasoning," whether or not there is any connection between probability and frequency. By using the theory in just this sort of problem, where the "frequentist" would deny the validity of probability theory, I hope to show that we can not only obtain important, useful, and nontrivial results; we can also clear up some of the paradoxes surrounding present communication theory, statistical mechanics, and quantum mechanics.

In fact, the problem of quality control, which led us into this little excursion, provides one of the most striking examples of the value of this nit-picking. However, I want to postpone discussion of the history of this problem until we have the full comparisons at hand; then we will be able to see how much statistical practice has suffered from the other kind of nit-picking, which restricts the apparent range of validity of the theory.

Before returning to the quality-control problem, let's extend the result (5-12) to get the general relations in sampling from a finite population. For this, we need a little more notation; let B_k stand for the proposition, "black ball at the k'th draw," whereupon $b_k \equiv W_k$ will stand for, "white ball at the k'th draw." And, let's indicate the prior information more explicitly. What I called X in (5-12) contained the statement that we have a total of N balls, of which n are black, and (N-n) white; to remind us of this, I will now write Eq. (5-12) in the form

$$(B_1|N,n) = n/N. \tag{5-13}$$

Now, what is the probability of drawing two black balls in two draws? This is, by Rule 1,

$$(B_1B_2|N,n) = (B_1|N,n)(B_2|B_1,N,n) \tag{5-14}$$

First, we suppose that a ball drawn is not replaced before drawing the next one. So, in evaluating the last factor, the fact that one black one has already been drawn means that at the second draw we are sampling from a population of (N-1) balls, of which (n-1) are black; and so

$$(B_1B_2|N,n) = \frac{n(n-1)}{N(N-1)} = \frac{n!\ (N-2)!}{(n-2)!\ N!} \tag{5-15}$$

Continuing in this way, we see that the probability of drawing r black balls in succession without replacement, is

$$(B_1\ldots B_r|N,n) = \frac{n!\ (N-r)!}{(n-r)!\ N!}\quad ,\ r \le n \tag{5-16}$$

The restriction $r \le n$ isn't necessary if we understand that we define factorials by the gamma function relation: $n! \equiv \Gamma(n+1)$; for then the factorial of a negative integer is infinite, and (5-16) automatically gives zero when $r > n$.

Likewise, the probability of drawing s white balls in succession without replacement is given by a relation of the same form, except that the roles of n and (N-n) are interchanged:

$$(W_1\ldots W_s|N,n) = \frac{(N-n)!\ (N-s)!}{(N-n-s)!\ N!} \tag{5-17}$$

Next, we ask for the probability that in m draws without replacement we shall obtain r black balls and (m-r) white ones, in a specified order. Suppose first that black balls are drawn on the first r trials, and white ones on the remaining (m-r) trials. Then Rule 1 gives

$$(B_1\ldots B_rW_{r+1}\ldots W_m|N,n) = (B_1\ldots B_r|N,n)(W_{r+1}\ldots W_m|B_1\ldots B_r,N,n) \tag{5-18}$$

of which the first factor is given by (5-16), and the second by (5-17), if we note that after r black balls have been drawn, we are then sampling

from a population of (N-r) balls (instead of N), of which (n-r) are black (instead of n). Also, the quantity denoted by s in (5-17) is equal to (m-r). So, we have

$$(B_1 \cdots B_r W_{r+1} \cdots W_m | N,n) = \frac{n!\ (N-r)!}{(n-r)!\ N!} \frac{(N-n)!\ (N-m)!}{(N-n-m+r)!\ (N-r)!} \tag{5-19}$$

Although this result was derived for a particular order of drawing black and white balls, the probability actually depends only on the numbers r, (m-r) drawn; and not on the particular order in which black and white appeared. To see this, write out the expression (5-19) more fully, in the manner

$$\frac{n!}{(n-r)!} = n(n-1)(n-2)\cdots(n-r+1) \tag{5-20}$$

and similarly for the two other ratios of factorials in (5-19). It then becomes

$$\frac{n(n-1)\cdots(n-r+1)\,(N-n)\,(N-n-1)\cdots(N-n-m+r+1)}{N(N-1)\cdots(N-m+1)} \tag{5-21}$$

Now suppose that r black balls and (m-r) white ones are drawn, in any other order. The probability of this is the product of m factors; every time a black one is drawn there appears a factor: (number of black balls in urn)/(total number of balls); and similarly for drawing a white one. The total number of balls in the urn decreases by one at each drawing; therefore, for the k'th drawing a factor (N-k+1) appears in the denominator, whatever the colors of the first k draws. Just before the k'th black ball is drawn, whether this occurs on the k'th trial or any later one, there are (n-k+1) black balls in the urn; so drawing the k'th black one places a factor (n-k+1) in the numerator. Just before the k'th white ball is drawn, there are (N-n-k+1) white balls in the urn; and so drawing the k'th white one places a factor (N-n-k+1) in the numerator regardless of whether this occurs on the k'th trial or any later one. Therefore, by the time all m balls have been drawn, one has accumulated exactly the same factors in numerator and

and denominator as in (5-21); different orders of black and white correspond only to different permutations of the order of factors in the numerator. The probability of drawing r black balls in <u>any</u> specified order in m trials, without replacement, is therefore given by (5-19).

Finally, we ask: what is the probability of drawing exactly r black balls in m trials without replacement, regardless of their order? Different orders of drawing are mutually exclusive events, so we must sum over all possible orders. But since all orders have the same probability (5-19), this means that we must multiply (5-19) by the binomial coefficient

$$\binom{m}{r} \equiv \frac{m!}{r!\ (m-r)!} \tag{5-22}$$

which represents the number of different possible orders of drawing r black balls in m trials. [Question for you to ponder: why isn't this factor just m!? After all, we started this discussion by saying that all the balls, in addition to being either black or white, also carried individual labels i = 1, 2, ..., N, so permutations of black balls among themselves are distinguishable events. A little private thought will enable you to answer this, unless you have had the misfortune of studying Bose and Fermi statistics in quantum theory from the usual textbook discussions; in that case you may have some unlearning to do first. <u>Hint</u>: In (5-19) we are not specifying <u>which</u> black balls and <u>which</u> white ones are to be drawn; if we did, (5-19) would collapse to (N-m)!/N!].

Taking the product of (5-22) and (5-19), the many factorials appearing can be reorganized into three binomial coefficients, and the probability of r black balls in m trials without replacement becomes

$$(r|m,N,n) = \frac{\binom{n}{r}\binom{N-n}{m-r}}{\binom{N}{m}} \tag{5-23}$$

This is our main result, and it is called the hypergeometric distribution, because the right-hand side of (5-23) is closely related to the coefficients in the power series representation of the hypergeometric function. As an aid to memory, we can put this into a more symmetrical form by adopting a new notation; the probability of drawing b black and w white balls, without replacement, from a population of B black and W white ones, is

$$(bw|BW) = \frac{\binom{B}{b}\binom{W}{w}}{\binom{B+W}{b+w}} \tag{5-24}$$

and in this form we can generalize still further. We have been considering an urn with only two kinds of balls: black and white. Suppose there are also red, green, brown, etc. balls present; in all, m different colors. I leave it for you to verify that the probability of drawing n_1 balls of type 1, n_2 of type 2, etc., without replacement, from a population of N_1 of type 1, N_2 of type 2, etc., is

$$(n_1 \ldots n_m | N_1 \ldots N_m) = \frac{\binom{N_1}{n_1} \cdots \binom{N_m}{n_m}}{\binom{\Sigma N_i}{\Sigma n_i}} . \tag{5-25}$$

The hypergeometric distribution (5-23) is rather complicated in its most general form, but it goes into a simpler distribution in the limit where the numbers n, (N-n) become very large compared to the number m sampled. Intuitively, this is clear; since then the proportions of black and white balls in the urn change only negligibly due to the small number drawn, so the probability of getting a black ball is essentially the same at each drawing. To see this mathematically, note that (5-21) can be written as

$$\frac{n^r (N-n)^{m-r}}{N^m} \left\{ \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{r-1}{n}\right)\left(1-\frac{1}{N-n}\right)\left(1-\frac{2}{N-n}\right)\cdots\left(1-\frac{m-r-1}{N-n}\right)}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{m-1}{N}\right)} \right\} \tag{5-26}$$

Now let $N \to \infty$, $(N-n) \to \infty$ in such a way that the ratio $p \equiv n/N$ remains constant.

All the factors in curly brackets in (5-26) tend to unity, and so (5-26) reduces in the limit to

$$\frac{n^r (N-n)^{m-r}}{N^m} = p^r \, (1-p)^{m-r} \tag{5-27}$$

This is the probability of drawing r black, (m-r) white balls in a specified order, and you see that it corresponds to a constant probability p of getting a black ball, (1-p) of getting a white one, at each trial. The probability of getting r black in m draws regardless of the order, again requires the combinatorial factor (5-22); and so in the limit the hypergeometric distribution goes into

$$(r|m,p) = \lim_{\substack{N\to\infty \\ N-n\to\infty \\ n/N\to p}} (r|m,N,n) = \binom{m}{r} p^r \, (1-p)^{m-r} \tag{5-28}$$

This is the binomial distribution, so called because the function

$$f(s) \equiv \sum_{r=0}^{m} s^r \, (r|m,p) = \sum_{r=0}^{\infty} \binom{m}{r} (sp)^r \, (1-p)^{m-r}$$

$$= (sp + 1 - p)^m \tag{5-29}$$

is just a representation of Newton's binomial theorem. F(s) is called the generating function of the binomial distribution; we will see later that generating functions provide a powerful tool for carrying out certain advanced calculations, as was first shown in Laplace's "Theorie Analytique." Note that the evident relation f(1) = 1 is just a verification that the probabilities in (5-28) are correctly normalized; i.e.

$$\sum_{r=0}^{m} (r|m,p) = 1 \tag{5-30}$$

We can carry out a similar limiting process on the generalized hypergeometric distribution (5-25). Again, I leave it for you to verify that in the limit where all the $N_i \to \infty$ in such a way that the fractions

$$p_i \equiv \frac{N_i}{\Sigma N_i}$$

tend to constants, (5-25) goes into the multinomial distribution

$$(n_1 \cdots n_m | p_1 \cdots p_m) = \lim_{N_i \to \infty} (n_1 \cdots n_m | N_1 \cdots N_m)$$

$$= \frac{n!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m} \qquad (5\text{-}32)$$

where $n \equiv \Sigma n_i$. And, as in (5-29), you can define a generating function of (m-1) variables, from which you can prove that (5-32) is correctly normalized.

Up to now, we have considered only the case where we sample without replacement; and that is obviously appropriate to our quality-control problem, where each diode drawn is tested to destruction. But suppose now that we sample balls, and after noting the color of each, we replace it in the urn before drawing the next ball. This case, of sampling with replacement, is enormously more complicated conceptually, but with some assumptions usually made, ends up being simpler mathematically, than sampling without replacement. For, let's go back to the probability of drawing two black balls in succession:

$$(B_1 B_2 | N,n) = (B_1 | N,n)(B_2 | B_1,N,n) \qquad (5\text{-}33)$$

Evidently, we still have (n/N) for the first factor; but what is the second one? Answering this would be, in general, an enormously difficult problem, requiring a vast amount of additional data before it could be solved. Because, what happens to that black ball that we put back in the urn? If we merely dropped it into the urn, and immediately drew another ball, then it was left lying on the top of the other balls, (or in the top layer of balls); and so it is more likely to be drawn again than any other specified ball, whose location in the urn is unknown. But this upsets the whole basis of our calculation, because the probability of drawing any particular (i'th) ball is no longer given by Rule 4, which led to (5-12).

Evidently, the probability of drawing any particular ball now depends on such details as the exact size and shape of the urn, the size of the balls,

the exact way in which the first one was tossed back in, the elastic properties of balls and urn, the coefficients of friction between balls and between ball and urn, the exact way you reach in to draw the second ball, etc. Even if all these data were at hand, I don't think that a team of the 1,000 best mathematicians in the world, backed up by all the computing facilities in the world, would be able to solve the problem; or would even know how to get started on it. Still, I don't think it would be quite right to say that the problem is unsolvable in principle; only so complicated that it just isn't worth anybody's time even to think about it.

So, what do we do? Well, there's a very clever trick for handling problems that become too difficult. As far as I know, it originated in probability theory; but it produces such euphoria that it has already spread to physics, and there is some danger that it may spread also to other fields.

In probability theory, when a problem becomes too hard to solve, we solve it anyway by:

(1) making it still harder;

(2) redefining what we mean by "solving" it, so that it becomes something we can do;

(3) inventing a dignified and technical-sounding word to describe this procedure, which has the psychological effect of concealing the real nature of what we have done, and making it appear respectable.

In the case of sampling with replacement, we apply this strategy by (1) supposing that after tossing the ball in, we shake up the urn. However complicated the problem was initially, it now becomes many orders of magnitude more complicated, because the solution now depends on every detail of the precise way we shake it, in addition to all the factors mentioned above; (2) assert that the shaking has somehow made all these details irrelevant,

so that the problem reverts back to the simple one where Rule 4 applies; (3) inventing the dignified-sounding word randomization to describe what we have done. This term is, evidently, a euphemism whose real meaning is: deliberately throwing away relevant information when it becomes too complicated for us to handle.

I have described this procedure in laconic terms, because an antidote is needed for the impression created by some writers on probability theory, who attach a kind of mystical significance to it. For some, declaring a problem to be "randomized" is an incantation with the same purpose and effect as those uttered by a Priest to convert ordinary water into Holy Water; i.e., it sanctifies their subsequent calculations and renders them immune to criticism. We agnostics often envy the sense of security that the True Believer thus acquires so easily; but which is forever denied to us.

However, in defense of this procedure, we have to admit that it often leads to a useful approximation to the correct solution; i.e., that the complicated details, while undeniably relevant, might nevertheless have little numerical effect on the answers to certain particularly simple questions, such as the probability of drawing r black balls in m trials when m is sufficiently small.

From the standpoint of principle, however, an element of vagueness necessarily enters at this point; for while we may feel intuitively that this leads to a good approximation, nobody has ever produced a proof of this, much less a reliable estimate of the accuracy of the approximation, which presumably improves with more shaking. The vagueness is particularly evident in the fact that different people have widely divergent views about exactly how much shaking is required to justify step (2). [Witness the minor furor surrounding a recent Government-sponsored and nationally televised game of chance, when someone objected that the procedure for drawing numbers from a

fish bowl to determine the order of call-up of young men for Military Service was "unfair" because the bowl hadn't been shaken enough to make the drawing "truly random," whatever that means. Yet if anyone had asked the objector: "To whom is it unfair?" he could not have given any answer except, "To those whose numbers are on top; I don't know who they are." But after any amount of further shaking, this will still be true!]

Again, you may accuse me of nit-picking, because you know that after all these polemics, I am just going to go ahead and use the randomized solution like everybody else does. Note, however, that my objection is not to the procedure itself, provided that we frankly acknowledge what we are doing; i.e., instead of solving the real problem, we are making a practical compromise and being, of necessity, content with an approximate solution of unknown accuracy. That is something we have to do in all areas of applied mathematics, and there is no reason to expect probability theory to be any different in this respect.

My objection is to this mystical belief that by "randomization" we have somehow washed away all our sins, and from that point on we proceed with exact relations--so exact that we can then subject our solution to all kinds of extreme conditions and believe the results. The most serious and most common error resulting from this belief is in the derivation of limit theorems (i.e., when sampling with replacement, nothing prevents us from passing to the limit $m \to \infty$ and obtaining the usual "laws of large numbers"). If we don't recognize the approximate nature of our starting equations, we delude ourselves into believing that we have "proved" things (such as the rigorous identity of probability and limiting frequency) that are just not true in real random experiments.

Returning to the equations, what answer can we now give to the question

posed after Eq. (5-33)? The probability $(B_2|B_1,N,n)$ of drawing a black ball on the second draw, clearly depends not only on N and n, but also on the fact that a black one has already been drawn and replaced. But this latter dependence is just so complicated that we can't, in real life, take it into account; so we shake the urn to "randomize" the problem, and then declare B_1 to be irrelevant: $(B_2|B_1,N,n) \simeq (B_2|N,n) = n/N$. After drawing and replacing the second ball, we again shake the urn, declare it "randomized", and set $(B_3|B_2,B_1,N,n) \simeq (B_3|N,n) = n/N$, etc. In this approximation, the probability of drawing a black one at <u>any</u> trial, is (n/N), and (N-n)/N is the probability, at every trial, of drawing a white ball. This leads us to write the probability of drawing exactly r black balls in m trials regardless of order, as

$$(r|m,N,n) = \binom{m}{r}\left(\frac{n}{N}\right)^r\left(\frac{N-n}{N}\right)^{m-r} \tag{5-34}$$

which is just the binomial distribution (5-28) with p = n/N.

Evidently, for small m, this approximation will be quite good; but for large m these small errors can accumulate (depending on exactly how we shake the urn, etc.) to the point where (5-34) is utterly useless. However, I think that some workers in probability theory would deny this; so let's demonstrate it explicitly by a simple, but realistic, extension of the problem.

Suppose that drawing and replacing a black ball actually increases the probability of a black one at the next draw by some small amount $\varepsilon>0$, while drawing and replacing a white one decreases the probability of a black one at the next draw by a (possibly equal) small quantity $\delta>0$; and that the influence of earlier draws than the last one is negligible compared to ε or δ. Then

$$\begin{aligned} (B_k|B_{k-1},N,n) &= p + \varepsilon \quad , & (B_k|W_{k-1},N,n) &= p - \delta \\ (W_k|B_{k-1},N,n) &= 1 - p - \varepsilon, & (W_k|W_{k-1},N,n) &= 1 - p + \delta \end{aligned} \tag{5-35}$$

where $p \equiv n/N$. The probability of drawing r black, (m-r) white balls in any specified order, is easily seen to be:

$$p(p+\varepsilon)^{b}(p-\delta)^{b'}(1-p+\delta)^{w}(1-p-\varepsilon)^{w'} \tag{5-36}$$

if the first draw is black, while if the first is white, the first factor in (5-36) should be (1-p). Here b is the number of black draws preceded by black ones, b' the number of black preceded by white, w the number of white draws preceded by white, and w' the number of white preceded by black. Evidently,

$$b + b' = \begin{Bmatrix} r-1 \\ r \end{Bmatrix} , \qquad w + w' = \begin{Bmatrix} m-r \\ m-r-1 \end{Bmatrix} \tag{5-37}$$

the upper case and lower cases holding when the first draw is black or white, respectively.

Now it is clear that, when r and (m-r) are small, the presence of ε and δ in (5-36) makes little difference, and it reduces for all practical purposes to

$$p^{r}(1-p)^{m-r} \tag{5-38}$$

as in the binomial distribution (5-34). But as these numbers increase, we can use relations of the form

$$\left(1 + \frac{\varepsilon}{p}\right)^{b} \simeq \exp\left(\frac{\varepsilon b}{p}\right) \tag{5-39}$$

and (5-36) goes into

$$p^{r}\ (1 - p)^{m-r}\ \exp\left\{\frac{\varepsilon b - \delta b'}{p} + \frac{\delta w - \varepsilon w'}{1-p}\right\} \tag{5-40}$$

The probability of drawing r black, (m-r) white balls now depends on the order in which black and white appear, and for a given ε, when the numbers b, b', w, w' become sufficiently large, the probability can become arbitrarily large (or small) compared to (5-38).

We see this effect most clearly if we suppose that $N = 2n$, $p = 1/2$, in which case we will surely have $\varepsilon = \delta$. The exponential factor in (5-40) then

reduces to:

$$\exp\{2\varepsilon[(b-b') + (w-w')]\} \tag{5-41}$$

This shows that, (1) as the number m of draws tends to infinity, the probability of results containing "long runs"; i.e. long strings of black (or white) balls in succession, becomes arbitrarily large compared to the value given by the "randomized" approximation; (2) this effect becomes appreciable when the numbers (εb), etc., become of order unity. Thus, if $\varepsilon = 10^{-3}$, the "randomized" approximation can be trusted up to about $m \sim 1000$; beyond that, you are deluding yourself by using it. In the limit $m \to \infty$, it cannot be trusted for <u>any</u> $\varepsilon > 0$.

All right, we've had a first glimpse at some of the principles and pitfalls of standard sampling theory, so let's turn back to the quality-control problem in which the question came up.

5.3. The Robot's Procedure

You recall, we were trying to use Bayes' theorem in the form of the evidence function:

$$e(A|E) = e(A|X) + 10 \log_{10} \frac{(E|A)}{(E|a)} \tag{5-42}$$

to test hypothesis A ≡ "we have a batch in which 1/3 are bad" against a single alternative B ≡ "we have a batch in which 1/6 are bad." The prior evidence for A was, by (5-10), $e(A|X) = -10$ db, and we had reached the "problem" of evaluating the other terms $(E|A)$, $(E|a)$ in (5-9) for the case that the experimental result was E ≡ "we draw a bad one on the first draw." What is the probability of this happening if A is true? Well, if 1/3 of them are bad, then we are sampling from a population of unknown total N, in which $n = N/3$ are bad, $(N-n) = 2N/3$ good. By (5-12), the probability of drawing a bad one on the first draw, given A, is of course $(\text{bad}|A) = n/N = 1/3$, as was obvious to all from the start. To evaluate $(E|a) = (\text{bad}|a)$, note

that in this problem it is part of the prior information X that either proposition A or B must be true; no other hypothesis about the batch is to be considered (we will see in Lecture 6 what happens if we change this condition). So, in this problem, a = B; if A is false, then B must be true; i.e. there are 1/6 bad, and $(E|a) = 1/6$. Thus, if we draw a bad one on the first draw, this will increase the evidence for A by

$$10 \log_{10} \frac{(E|A)}{(E|a)} = 10 \log_{10} \frac{(1/3)}{(1/6)} = 10 \log_{10} 2 = 3 \text{ db} \qquad (5\text{-}43)$$

What happens now if we draw a second bad one? We are sampling without replacement, so in the notation of (5-14), this contributes further evidence of

$$10 \log_{10} \frac{(B_2|B_1A)}{(B_2|B_1a)} \qquad (5\text{-}44)$$

But $(B_2|B_1A) = (n-1)/(N-1)$ now depends on the number N in a batch. To avoid this complication, let's suppose that N, while unknown, is at least known to be very much larger than any number that we contemplate testing; i.e. we are going to test such a negligible fraction that the proportion of bad and good ones in the batch is not changed appreciably by the drawing. Then the limiting form of the hypergeometric distribution (5-23) will apply, namely the binomial distribution (5-28). Or, you can say equally well that in this case sampling without replacement is practically the same thing as sampling with replacement, leading again to the binomial distribution (5-34). In any event, the result is that the probability of drawing a bad one is the same at every draw, regardless of what has been drawn previously; so Eq. (5-43) now applies for every draw in which we get a bad one. Every bad one we draw will provide +3 db of evidence in favor of hypothesis A, the proposition that we had a bad batch. Now suppose we find a good diode. We'll get evidence for A of

$$10 \log_{10} \frac{(\text{good}|A)}{(\text{good}|a)} = 10 \log_{10} \frac{2/3}{5/6} = -0.97 \text{ db}, \tag{5-45}$$

but let's call it -1 db. Again, this will hold for any draw, if the number in the batch is sufficiently large. If we have inspected n diodes, of which we found n_b bad ones and n_g good ones, the evidence that we have the bad batch will be

$$e(A|E) = -10 + 3n_b - n_g \ . \tag{5-46}$$

You see how easy this is to do once we've set up the machinery. For example, if the first twelve we test show up five bad ones, then we'd end up with

$$e(A|E) = -10 + 15 - 7 = -2 \text{ db} \tag{5-47}$$

or, from Figure (5-1), the probability of a bad batch is brought up to

$$(A|E) \simeq 0.4 \ . \tag{5-48}$$

In order to get at least 20 db worth of evidence for proposition A, how many bad ones would we have to find in a certain sequence of tests? Well, that's not a hard question to answer. If the number of bad ones satisfies

$$n_b \geq 5 + \frac{n}{4} \tag{5-49}$$

then we have at least 20 db of evidence for the bad batch above where we started. Which shows that if we make enough tests, if just slightly more than a quarter of the ones tested turn out to be bad, that will give us 20 db of evidence that we have the batch in which 1 in 3 are bad.

Now all we have here is the probability or plausibility or evidence, whatever you wish to call it, of the proposition that we got the bad batch. Eventually, we have to make a decision. We're going to accept it or we're going to reject it. How are we going to do that? Well, evidently we have to decide beforehand: if the probability of proposition A reaches a certain level than we'll decide that A is true. If it gets down to a certain value, then we'll decide that A is false. There's nothing in probability theory

which can tell us where to put these threshold levels at which we make our decision. This has to be based on our judgment as to what are the consequences of making wrong decisions, and what are the costs of making further tests. For example, making one kind of error (accepting a bad batch) might be very much more serious than making the other kind of error (rejecting a good batch). That would have an obvious effect on where we place our threshold. So we have to give the robot some instructions such as "if the evidence for A gets greater than +0 db, then we'll reject this batch. If it goes down as low as - 15, then we'll accept it."

Let's say that we'd set some threshold limits: we arbitrarily decided that we will reject the batch if the evidence reaches the upper level, and we will accept it if the plausibility goes down to the lower one. We start doing the tests, and every time we find a bad one the evidence for the bad batch goes up 3 db; every time we find a good one, it goes down 1 db. The tests terminate as soon as we get into either the accept or reject region for the first time. This would be the way our robot would do it if we told him to reject or accept on the basis that the posterior probability of proposition A reaches a certain level.

We could describe this in terms of a "control chart," where we start at -10 db at zero number of tests, and plot the result of each test (Fig. 5.2).

5.4. Wald's Probability-Ratio Test.

Now, how does Wald do this? He (Wald, 1947) does not mention Bayes' theorem. But what he actually does is just the same with the one characteristic difference which we find in all these comparisons. Like Fisher in the case of maximum likelihood, he always starts out by throwing away his prior information. His graphs always start out at 0 db.

Wald's probability ratio test involves the calculation of just the last term of Equation (5-9), except that he uses natural logarithms. The

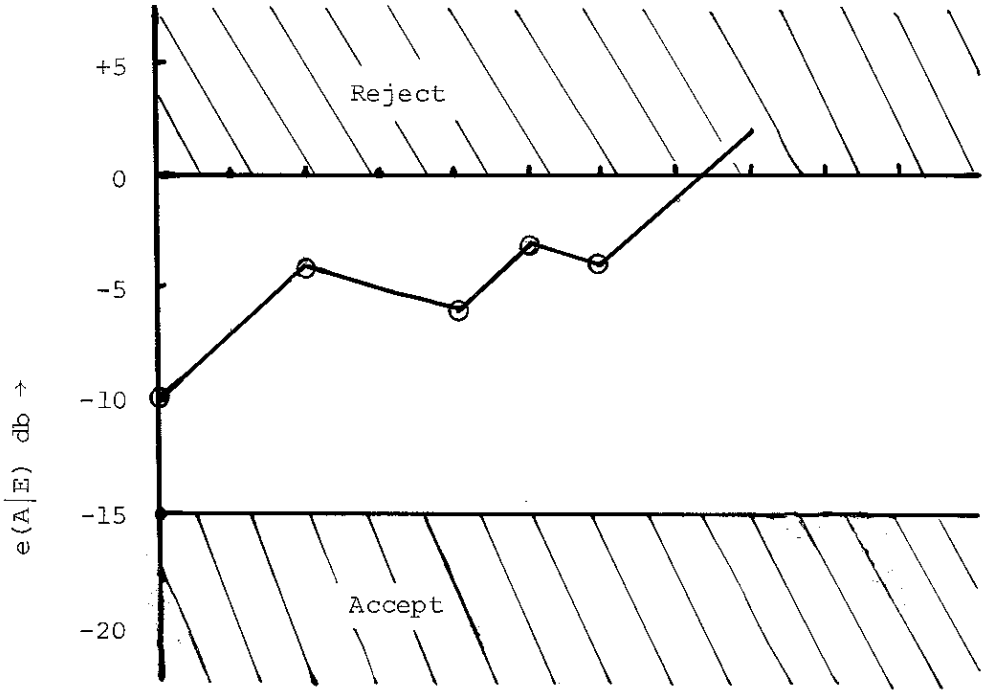

Figure 5.2. A control chart for sequential testing.

name "sequential" describes the fact that the number of tests is not determined in advance, but depends on what is observed. Thus, at each step of the sequence of tests we choose one of three alternatives: (1) accept; (2) reject; (3) make another test. This is the procedure which he conjectured represents an optimum procedure in the sense of requiring on the average fewer tests than any other, but he didn't succeed in proving it. Several years later, such a proof was offered, by Wald and Wolfowitz. We can well imagine how much mathematical effort has been expended on this problem. But how does it look to our robot? Well, to the robot this problem doesn't exist at all; it is only a "Scheinproblem." To him the fact that we have derived it from Bayes' theorem is already the proof that the probability ratio test is the optimum calculation to do, by any sensible criterion of "optimum." Any criterion which required us to reason in a manner not reducible to Bayes' theorem would also require us to be inconsistent in the

sense discussed earlier, or to violate qualitative common sense. Our robot would say this: "When you have calculated the probability of proposition A, conditional on all the available evidence, then you have got everything bearing on the truth of A that is to be had from the evidence. No method of analyzing the data can give you more than this, and there is nothing more to be said."

Does anyone incur any serious error by starting out at zero db? In principle, this is bad in the sense that it is inconsistent if we do have prior information. But, of course, in practice the person using the test still has his common sense; and if he has prior information he will use that information in deciding where to put the boundaries of the accept and reject regions. We cannot remove all the arbitrariness in location of these boundaries, but we can remove some of it, by taking into account prior probability. In practice, the orthodox statistician would use his common sense to take account of his prior information, without ever having to admit that there is any such thing as a "prior probability."

A particularly frank admission of the relevance of prior information is given by Lehman (1959; p. 62) in his well-known work on hypothesis testing according to the "orthodox" viewpoint. He writes: "Another consideration that frequently enters into the specification of a significance level [this is something essentially equivalent to choosing the threshold values in our problem] is the attitude toward the hypothesis before the experiment is performed. If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low." Exactly so! But it is just the prior probability that shows quantitatively how this is to be done.

Of course, there is a great deal more to sequential testing theory than just applying the probability ratio test. There are many questions about the procedure that the manufacturer and customer would ask, and would want the statistician to answer. For example, if all batches have a certain fraction f defective, and we use a sequential test with specified threshold levels, α, β what is the expected number of diodes tested per batch? How does this average sample number depend on $\{f,\alpha,\beta\}$? Or if a fraction g of the batches is in fact bad, what fraction do we expect to be rejected on the average if certain threshold levels are used?

Questions of this type can be answered by straightforward extensions of this analysis and there is an extensive literature on them. In these talks we are concerned only with showing that the rules for plausible reasoning which we have built into the robot's brain will, if applied to this problem, lead to the same actual procedures as the newest methods developed by statisticians. Their conceptual basis is, however, entirely different. To the orthodox statistician, the justification of the sequential probability-ratio test would probably lie in considerations of average sample numbers for given probability of errors. To the robot, this is only an incidental consequence of the fact that this procedure is the one that makes full use of the available data, <u>because</u> it is derivable from Bayes' theorem.

We see that the robot's prediction has been borne out in one more example. We are warned not to use Bayes' theorem for quality-control tests, because it was associated with some metaphysical nonsense 150 years ago. But so was everything else in science. It is even insinuated that Bayes' theorem cannot be "applied." But the simple fact is that the most powerful known method of quality control, only recently discovered by statisticians, is nothing but an application of Bayes' theorem, in exactly the way Laplace would have handled this problem.

5.5. The Value of Nit-Picking.

So now we are in a position to discuss the value of my "nit-picking" about the meaning of Eq. (5-12), and see why the problem of quality control provides a good example of the situation. Basically, of course, it is a problem of testing one hypothesis (we got a bad batch) against a single alternative (we got a good batch); and the mathematics we have developed applies just as well to any such problem of hypothesis testing, such as testing two rival theories in physics, or biology, or economics, against each other.

Now the procedures we are developing in this and the next three lectures, were used by Laplace in just such problems (although not in the logarithmic form, which is only a convenient mathematical detail) from about 1774, and they have been available to anyone who had the sense to use them since the appearance of Laplace's Theorie Analytique in 1812. Yet generations of statisticians were taught that these methods were wrong, and it was only in the early 1940's--130 years later--that statisticians rediscovered the procedure in this lecture from an entirely different view point without at first recognizing it. It was then hailed as a major new advance in statistical practice, and several more years elapsed before it was generally realized (Good, 1950; Wald, 1950) that it was mathematically identical with application of Bayes' theorem in exactly the manner that had long been rejected as wrong.

What caused this procedure to be lost to science for 130 years? Just the point about which I was nit-picking earlier in this lecture; stubborn adherence to a belief, for which there is no supporting evidence, that the notion of probability can be used only in the sense of "frequency in a random experiment". From this one concluded that it is meaningless to speak of the probability that an hypothesis is true, because that is not a "random variable." On such grounds statistical workers denied themselves use of

the proper statistical methods, and worked instead with a great variety of ad hoc approximate methods.

In his later book, "Statistical Decision Functions" (Wald, 1950), Wald developed this theory very much further, and here we have one of those ironical situations where years of the most careful and painstaking work leads right back to the very thing one had been trying to refute. Wald sought to develop a general theory of decision making in the face of uncertainty in a way which avoids the supposed mistakes of Laplace and Bayes, who with Daniel Bernoulli had already developed such a theory in the 18'th century. In order to keep the theory completely "objective," the notion of inductive reasoning, which to Laplace was the central problem of the theory, was suppressed, and attention was concentrated on the decision itself. After long mathematical arguments to impose various conditions of consistency, it finally developed that a class of "admissible" decision rules, which consists, roughly speaking, of all those any sane person would ever consider adopting, is identical with the class derivable by the methods of Bayes and Laplace, and the only basis for a choice among them lay in the prior probabilities! Wald called this class of rules, very properly, "Bayes strategies." As a final irony it was shown (Chernoff & Moses, 1959; Chap. 6), that in practical applications it is only the fact that these decision rules can be found by repeated application of Bayes' theorem that makes it feasible to use this theory at all in nontrivial problems, where the number of conceivable strategies is astronomical. We will come back to these topics when we take up Decision Theory in Lectures 13, 14.