BAYES' THEOREM [1]

By

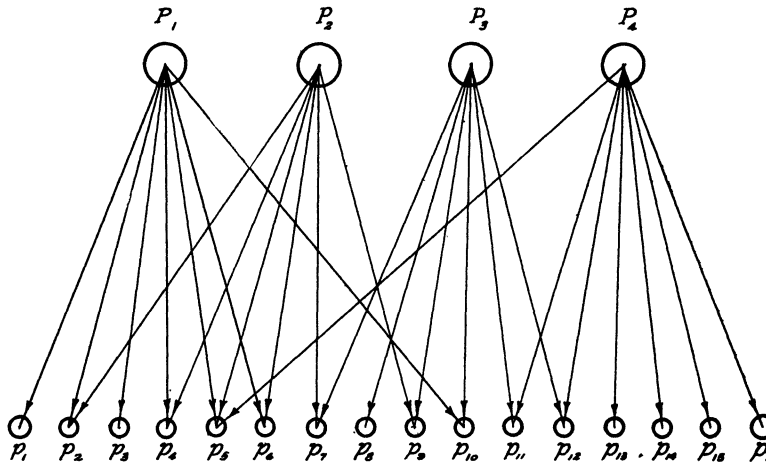
JOSEPH BERKSON

As for all established sciences, the typical problems of practical statistics have become inveterately attached to their several neat and convenient formulary solutions. To recall consideration of the basic reasoning underlying every-day statistical practice that applies to an elementary question may appear in the nature of an unnecessary disturbance of prevailing peace. If the experience of the writer is typical, however, vagueness or dubiousness of the premises inherent in a rule applied by rote will emerge to plague one in the conclusions, and a periodic return to fundamentals is as salutary for mental comfort as for the integrity of science itself. In what follows, an attempt will be made to go over the ground covered by Bayes' Theorem, and to point out its import for sound statistical reasoning. No claim is laid to mathematical originality at any specific points, but in the approach and synthesis will be found, we hope, a measure of instructive novelty.

A large class of statistical problems is typified in the following. A standard machine is known, from long experience, to produce a certain fraction P of imperfect products. What is the probability that in the next issue of n products, a fraction p will be imperfect?

We now present a related but not identical question. There is no available knowledge concerning the general practice of a machine; n products are examined and a fraction p found to be imperfect. What is the probability that the machine turns out generally a fraction P of imperfect products? The distinction between the two questions may be schematized as in Figure 1.

1. From the Department of Biometry and Vital Statistics of the School of Hygiene and Public Health (Paper No. 125); and the Institute for Biological Research of the Johns Hopkins University.



The values P_1 , P_2 , P_3 , P_4 represent serially all the various fractions of imperfect products which might characterize particular machines, each one, let us say, determined by some definite combination of mechanical defects. Values p_1 , p_2 , p_3 , etc., are the fluctuating fractions of imperfect products that might appear in the samples produced by these machines. Connected by arrows with P_1 are the randomly varying values of p that might result from P_1 , with P_2 those that might result from P_2 , etc., the weight of the arrows being proportional to the probability of the particular p concerned. It is to be noticed that each P may give rise to any of a number of p 's and that some of the p 's may result from any of a number of P 's.

The first question in terms of the diagram is: "Given P_1 , how probable is it that p_3 shall result?" The second is: "Given p_5 , how probable is it that P_1 has been its source?" Answering the first, we calculate in the realm of the p 's connected with P_1 . In the second we calculate in the realm of the P 's connected with p_5 .

An answer to the first is given directly in terms of our every-day statistical reasoning. We say that the p 's which result from P_1 can be adequately described as a normal distribution with $\sigma = \sqrt{\frac{P(1-P)}{n}}$, and from this the probability of any particular p calculated. The answer to the second is more difficult, and was given in general terms first by Bayes (1) in the theorem known by his name. Bayes' Theorem is not frequently used in applied statistics; yet the problems that

arise in practical situations would often seem to demand just such an answer as it provides. More often than not do we have a specific sample and inquire about the probable character of the universe from which it was drawn, in contra-distinction to the situation in which the universe is known, and the questions concern the possible samples.

The method of presenting the theorem here given will not follow rigidly any historical demonstration. Actually the calculation quantitatively of an "inverse probability" or the "probability of causes," was first given by Bayes. But he considered a purely geometric set-up and his solution was in terms of this conception. By implication he utilized a general principle first clearly stated later by Laplace, and furthermore, Laplace generalized the solution still more by arguing from the probability of a cause given by a particular sample, to the probability of the next sample. With this realized, then, that Bayes is to be credited with the original demonstration and Laplace for an important extension, we may proceed to a demonstration which is not exactly that of either.

I. *Problem.* We have an urn containing three balls. Each ball is colored black or white, and each color is equally likely. We draw one ball and it is black. What are the probable contents of the urn? We argue—the following are the possibilities:

I	II	III	IV
w w w	w w b	w b b	b b b

All of these possibilities, we say, are equally likely *a priori* and we have for the probabilities of the sample the following:

P_s I, the probability of a black sample from I = 0
P_s II, the probability of a black sample from II = 1/3
P_s III, the probability of a black sample from III = 2/3
P_s IV, the probability of a black sample from IV = 3/3

where P_s I is the probability of the sample s being drawn from urn I, P_s II from urn II, etc. We say now that the relative probabilities of the various urns are in proportion to the probabilities of the sample drawn, and we have

(a) $$P\,\mathrm{I} : P\,\mathrm{II} : P\,\mathrm{III} : P\,\mathrm{IV} = 0 : 1/3 : 2/3 : 3/3$$

where P I is the probability that, having drawn the ball, urn I was its source, P II that urn II was the source, etc.

Also, since the ball must have been drawn from some one of the urns, the total probability of one or another of the urns is unity and we have

(b) $$P\ \text{I} + P\ \text{II} + P\ \text{III} + P\ \text{IV} = 1$$

From (a) and (b) we have therefore

$$\begin{aligned} P\,\text{I} &= 0 \\ P\,\text{II} &= 1/6 \\ P\,\text{III} &= 2/6 \\ P\,\text{IV} &= 3/6 \end{aligned}$$

We now extend the problem to the case where the *a priori* probabilities of the various possible urns are not equal.

Suppose we say that there are many urns of the description I, II, III, IV in a large chamber, and that these are in proportion I : II : III : IV = 1 : 2 : 3 : 4. We now pick an urn at random and draw from it a ball, which turns out to be black. What is the probability that the urn is of some particular description? Proceeding as before, we have for the probabilities of the sample being drawn from the various urns the following:

p_s I = 1/10 × 0 = 0 (Probability of urn × probability of sample)
p_s II = 2/10 × 1/3 = 2/30
p_s III = 3/10 × 2/3 = 6/30
p_s IV = 4/10 × 3/3 = 12/30

where p_s I is the probability that such a sample s be drawn from urn I, etc.

And again on the principle that the probabilities of the urns are in proportion to the probabilities of the sample drawn, we have

$$P\ \text{I} : P\ \text{II} : P\ \text{III} : P\ \text{IV} = 0 : 2/30 : 6/30 : 12/30$$

and as preceding

$$P\,\mathrm{I} + P\,\mathrm{II} + P\,\mathrm{III} + P\,\mathrm{IV} = 1.$$

Therefore

$$\begin{aligned} P\,\mathrm{I} &= 0 \\ P\,\mathrm{II} &= 2/20 \\ P\,\mathrm{III} &= 6/20 \\ P\,\mathrm{IV} &= 12/20 \end{aligned}$$

We shall now generalize this solution.

Let π_1, π_2, π_3, etc. be the *a priori* probabilities of the various possible universes from which a sample is to be drawn. Let p_1, p_2, p_3, etc., be the probability of the sample being drawn from the respective universes. Then, a sample s having been drawn, the probability that its source is universe r is given by

$$P_r = \frac{\pi_r p_r}{\sum \pi p}$$

If all the universes are equally likely (our first case above), $\pi_1 = \pi_2 = \pi_3 = \pi_4 \cdots$ and we have

$$P_r = \frac{p_r}{\sum p} \tag{1}$$

If the equally probable universes are infinite in number, the P's varying by infinitesimal gradations from zero to unity, and p may assume any positive value less than 1, we may extend the last formula (1) by use of the calculus as follows:

Let x = any possible P between 0 and 1. From a universe x I draw a sample containing $r + s$ individuals, designated hereafter as a sample (r, s). The probability that it will contain r successes and s failures is given by

$$P_{(r,s)} = E_{r,s}\, x^r (1-x)^s$$

where $P_{(r,s)}$ is the probability that the sample (r, s) coefficient of the $(r+1)$th term in the Bernoulli expansion $= \frac{(r+s)!}{r!\,s!}$.

The probability of the sample of (r, s) coming from a universe

the P of which lies between x and $(x + dx)$ is therefore

$$ {}_{x}^{x+dx}P_{(r,s)} = E_{r,s}\, x^r(1-x)^s dx $$

where ${}_{x}^{x+dx}P_{(r,s)}$ is the probability that the sample (r, s) emanates from a universe whose P lies between x and $(x + dx)$. If the universe from which the sample is drawn may have a P anywhere between a and b, the probability of the sample (r, s) is

$$ (2) \qquad {}_{a}^{b}P_{(r,s)} = E_{r,s} \int_{a}^{b} x^r(1-x)^s dx $$

and the probability that x is between a and b is therefore as in (1)

$$ (3) \qquad {}_{a}^{b}P = \frac{\int_{a}^{b} x^r(1-x)^s dx}{\int_{0}^{1} x^r(1-x)^s dx} $$

where ${}_{a}^{b}P$ is the probability that the universe from which the sample (r, s) was drawn has a P between a and b. This is Bayes' Theorem in terms of the integral calculus.

Now, we ask the further question, what is the probability of a second sample containing m successes and n failures[1] being drawn?

If x be the p of the universe from which the sample (m, n) is drawn, and if P may vary from 0 to 1 we have analogously with (2)

$$ (4) \qquad {}_{0}^{1}P_{(m,n)} = E_{m,n} \int_{0}^{1} x^m(1-x)^n dx $$

where ${}_{0}^{1}P_{(m,n)}$ is the probability that a sample (m, n) be drawn from universes whose P's vary between 0 and 1, and

$$ E_{m,n} = \frac{(m+n)!}{m!\,n!} $$

1. Designated hereafter as the sample (m, n).

The probability of the event (m, n) occurring from any particular universe is given by the product of the probability of that universe and the probability of the event. The total probability of the event (m, n), i. e., the probability that the event (m, n) occurs at all from any universe, is, therefore, given by the product of form (3) with 0 and 1 substituted for a and b and (4), as follows:

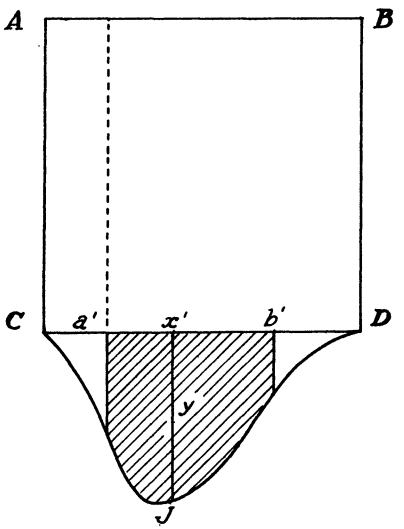
(5)
$$P_{(m,n):(r,s)} = \frac{(m+n)!}{n!\,m!} \frac{\int_0^1 x^{r+m}(1-x)^{s+n}\,dx}{\int_0^1 x^r(1-x)^s\,dx}$$

where $P_{(m,n):(r,s)}$ is the probability of a second sample (m, n) after a first sample (r, s) has been drawn.

This is Laplace's extension of Bayes' Theorem, somewhat modified.

Bayes' Solution.

It will be illuminating to derive this result by the method of Bayes. We shall follow his proof except to simplify his notation and to use the integral calculus where he used geometric demonstration.



ABCD is a square billiard table. A ball is thrown and comes to rest at a', through which a line is drawn parallel to AC. A second ball is thrown; if it stops to the left side of the line a', we designate a success, to the right, a failure. Before the first ball is thrown, what is the probability of the second ball succeeding r and failing s times in r plus s trials?

If the first ball comes to rest at x', the probability of a successful second throw is $\frac{Cx'}{CD} = p$ and of failure $\frac{Dx'}{CD} = q$. The probability of r successes and s failures with the first ball at x is then $\frac{(r+s)!}{r!s!} p^r q^s$.

Let us erect at each point x' along CD a distance y', so that

$$\frac{y'}{CD} = \frac{(r+s)!}{r!s!} p^r q^s \tag{6}$$

and connect the summits forming a figure as shown in Figure 2. At each point, of course, y' will be different because $p = \frac{Cx'}{CD}$, and $q = \frac{Dx'}{CD}$ will be different, but for any particular case, r and s remain constant.

The probability that the first ball shall fall between a and $(a+dx)$ is $\frac{dx'}{AD}$ and that the second ball shall therefore succeed r and fail s times is $\frac{y'}{CD}$. That both shall happen is therefore

$$\frac{y'}{CD} \times \frac{dx'}{CD}$$

and if x is to be between a' and b', the total probability is

$${}_{a'}^{b'}P_{(r,s)} = \frac{1}{CD^2} \int_{a'}^{b'} y' dx'$$

where ${}_{a'}^{b'}P_{(r,s)}$ is the probability that the first ball fall between a' and b' and that a ball thrown subsequently $r+s$ times, succeed r and fail s times.

But CD^2 = Area of AD and $\int_{a'}^{b'} y' dx'$ = Area of the shaded portion, $a'Jb'$. Therefore

$$(7) \qquad {}_{a'}^{b'}P_{(r,s)} = \frac{\text{Area } a'Jb'}{\text{Area } AD}$$

The probability that the first ball fall between C and D and thereafter there occur r successes and s failures is similarly $\frac{\text{Area } CJD}{\text{Area } AD}$. But the first ball must fall somewhere between C and D; therefore the total probability of the second throws having r successes and s failures is given by

$$(8) \qquad P_{(r,s)} = \frac{\text{Area } CJD}{\text{Area } AD}$$

With this established, the analysis proceeds.

Given the result of a series of throws to be r successes and s failures, what is the probability that the first ball has fallen between a' and b'? This we may obtain by the use of the solution already derived and the principle of compound probability[1].

Let x be the desired probability that the first ball fell between a' and b'. We have seen that the probability of r successes and s failures in the second series of throws is

$$\frac{\text{Area } CJD}{\text{Area } AD} \qquad \text{from (8)}$$

therefore the probability of the first falling between a' and b' and the experience (r, s) following is

$$x \cdot \frac{\text{Area } CJD}{\text{Area } AD}$$

But we have shown that this combined probability is equal to

$$\frac{\text{Area } a'Jb'}{\text{Area } AD} \qquad \text{from (7)}$$

Therefore

$$(9) \qquad x = \frac{\text{Area } a'Jb'}{\text{Area } CJD}$$

1. This step is very elaborately proved in Bayes' original paper by a circuitous demonstration.

This is Bayes' Theorem, as its author gives it. The additional part of his work is concerned with the quantitative estimate of the ratio.

We may now show that his solution is the same as that given in (3), as follows:

$$y' = CD \times E_{r,s}\left(\frac{x'}{CD}\right)^r\left(1-\frac{x'}{CD}\right)^s \qquad \text{from (6)} \tag{10}$$

where

x' = distance from C to x'

$$E_{r,s} = \frac{(r+s)!}{r!s!}$$

Now

$$a' = a \times CD$$
$$b' = b \times CD$$

a and b having the meaning of equation (3). Assume the relationship

$$x' = CD \times x \tag{11}$$

$$dx' = CD \times dx \tag{12}$$

Then

$$\text{Area } aJb = \int_{x'=a'}^{x'=b'} y'dx'$$

$$= CD^2 \times E_{r,s} \int_{x=a}^{x=b} x^r(1-x)^s dx$$

(Substituting from (11) and (12)).

Similarly

$$\text{Area } CJD = CD^2 \times E_{r,s} \int_{x=0}^{x=1} x^r(1-x)^s\,dx$$

Therefore

$$Area\ \frac{aJb}{CJD} = \frac{\int_a^b x^r(1-x)^s\,dx}{\int_0^1 x^r(1-x)^s dx}$$

which is the same as formula (3) previously derived.

To be directly applicable to statistical problems formula (5) must be numerically evaluated. This is accomplished exactly for most practical instances only with a great amount of labor, and methods of approximation have been resorted to. For a few simple special cases the solution may be easily derived as follows:

An event has been tried N times with p successes and q failures. What is the probability that in the next single trial it will succeed?

Applying formula (5) to this instance, we have

$$r = p \qquad m = 1$$
$$s = q \qquad n = 0$$

and the desired probability is given by

$$P = \frac{\int_0^1 x^{(p+1}(1-x)^q dx}{\int_0^1 x^p(1-x)^q dx}$$

Now

$$\int_0^1 x^a(1-x)^b\,dx = \frac{a!\,b!}{(a+b+1)!}$$

From which we have

$$P = \frac{m+1}{m+n+2} = \frac{m+1}{N+2}$$

So that if nothing is known concerning an event except that it has been tried three times and succeeded twice, the probability that it will

succeed in the next trial is 3/5, not 2/3 as the more usual procedure would indicate. Again, if an event has occurred a thousand times without a failure, and we know concerning it nothing except that fact, the probability that it will fail next instance is 1/1002. If an event has never been tried at all, the probability that it will succeed on the first trial is 1/2.

An event has been tried N times and succeeded each instance. What is the probability that in the next d trials it will again succeed each time? Here

$$r = N \qquad m = d$$

$$s = 0 \qquad n = 0$$

and the desired probability is given by

$$P = \frac{\int_0^1 x^{N+d}\,dx}{\int_0^1 x^{N}\,dx}$$

$$= \frac{N+1}{N+d+1}$$

From this we conclude that if an event has succeeded 25 times and never failed, the probability that in 25 further trials it will again not fail even once is 26/51, or in general if an event has never failed in N trials, the probability that N further trials will yield no failure is about 1/2.

Discussion.

To precisely what position in the methodology of applied statistics Bayes' Theorem will eventually become adjusted, it is impossible at this point in its development to say with certainty. The literature on the subject, as soon as it leaves the realm of purely hypothetical situations, is rife with disagreement, and clarification remains a contemporary problem. In this brief presentation, no attempt can be made to adequately summarize the various views concerning the questions at issue. We may, however, consider a few points that have disciplinary value for statistical thinking rather than any immediate practical utility.

It is basic to the aims of statistical calculations to estimate the

probability of given experiences from assumptions of pure random variation. A consideration of the logic involved in the development of Bayes' Theorem is useful in bringing out the inadequacy of the reasoning by which our most ordinary statistical procedures attempt to accomplish this. If, having observed a probability p, we estimate the standard deviation of succeeding samples of n by $\sqrt{\frac{pq}{n}}$ we imply tacitly that in the universe from which the sample was drawn, the chance of a success is the p of our observation. The reasoning leading to, and formula (3) itself, indicate how unwarranted this is. Our knowledge of the universe which generated the sample is never given with certainty by the sample. Indeed, formula (3) states a probability for any particular universe that may be assumed. With only a sample as the source of knowledge, and without Bayes' Theorem, we have no clue as to the nature of the generating universe. But, if we do not know the universe, how are we to calculate the character of its samples? One answer is to take refuge in formula (5), i. e. use Bayes' Theorem. As a practical solution of the difficulty this has two major objections: first, there are no existing tables for making the necessary calculations without prohibitive arithmetic labor; second, even if the evaluation could be effected there are reasons to doubt the validity of its application. For the formula in question rests on the assumption that all the probabilities from zero to unity which might characterize the universes from which we draw samples are *a priori* equally likely, the socalled assumption of the equal distribution of ignorance. Now this is an exceedingly questionable assumption, and it is partly on these grounds that Keynes rejects outright the possibility of applying probability to actual experience. It must be admitted, we think, that it is difficult to see what there is to justify the assumption that every sort of general universe from which arise the events of experience is equally likely. Would it not appear the more reasonable hypothesis that these universes are themselves "events," samples of some larger universe; and why should this be extremely different in the distribution of its probabilities from the universes that we ordinarily meet? There are writers, however, who, admitting that the assumption is to be questioned, believe it may be subjected to experimental test, and have essayed to actually sample at random the probabilities that characterize the universes of our experience. It would be impertinent to assert that an experimental investigation is bound to be futile, but the utility of this sort of procedure seems to us exceedingly dubious. We doubt indeed that any clear meaning can be assigned to the concept of "the universes of our experience," of which random samples are to be obtained. But granting the existence of such a

distribution of *a priori* probabilities we doubt the relevancy of its estimation to any practical problem. In any actual investigation, we deal with a definite slice of possible experience; an anthropologist is not concerned with the universes dealt with in the investigation of an economist or an epidemiologist. If *a priori* probabilities are of interest to him, they are those that obtain in his peculiar world of observation. It appears to us quite as wide of the mark aimed at, to call in a formula which obtains its *a priori* probability from experience in general, as to obtain it from the unique experience at hand, and indeed it may be argued that, as between the two, the latter is the more reasonable.

What then does all this come to? Does it mean that the entire structure of established statistical procedure rests on quicksand, to be toppled over by anyone armed with a reading of Bayes' Theorem? We are inclined to the belief held by Keynes that, so far as *logic* is concerned, this is substantially true. As regards this, however, it is at bottom in no worse plight than any current scientific procedure when its fundamental assumptions are hard pressed. But we do not rest the matter here. All this admits is that applied statistics, like all applied science, is not founded on unquestionable premises and invulnerable logic. It is perfectly consistent to add that *in general* its formulae are good *approximations*. How good? This is a question permitting no dogmatic comprehensive answer. Differently good for different situations. Some idea of the degree of approximation may be obtained for given assumed conditions by direct calculation. It may be shown, for instance, that under certain conditions results obtained by way of Bayes' Theorem or the more usual "normal" distribution render not very different results, and these conditions, indeed, approach the ones we most frequently encounter. But, in general, a more satisfactory answer is furnished in the pragmatic consideration that our formulae have in fact been widely used and experience has not violated their anticipations. This is the fact that we would stress, because it throws into relief the experimental as opposed to the mathematical foundation of statistics. Comforted on the one hand that experience in general supports our procedures, the considerations we have elicited in this discussion will emphasize equally their shifting approximation. The clear minded and careful worker will keep this constantly in mind and shun literal interpretation of conclusions drawn from formulae applied to extreme cases. No scientist worth his salt will permit himself the use of formulae the premises of which he has not examined. But the statistician, because of the great variability of

the data with which he is likely to deal, stands in special need of this precaution. Where statistics run counter to what appears to be the general experience, it is a wise rule to re-examine the statistics rather than to indict forthwith the dependability of the experience. Such an attitude would modify considerably much that is found in current statistical literature and it would modify it in the direction of greater soundness.

REFERENCES

1. Bayes, Thomas. Phil. Trans., 1763, LIII, 370; 1764, LIV, 296.
2. Coolidge, Julian L. Probability, Chapter VI.
3. DeMorgan, Augustus. An Essay on Probabilities, Chapter III.
4. Keynes, Maynard. Treatise on Probability, Chapter XXX.
5. Pearson, Egon. Biometrika, 1925, XVII, 388.
6. Pearson, Karl. Phil. Mag. 1927, 13, 365.
7. Todhunter, I. A History of the Mathematical Theory of Probability, Chapter XIV.
8. Wishart, John. Biometrika, 1927, XIX, 1.