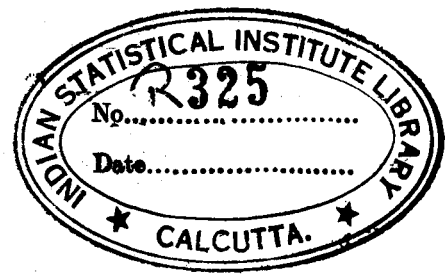
R 325
WAS

PROBABILITY AS A BASIS FOR ACTION

By

W. A. Shewhart

Bell Telephone Laboratories



W. A. SHEWHART'S COLLECTION

Paper presented at the joint meeting of the American Mathematical Society and Section K of the American Association for the Advancement of Science at Atlantic City, December 27, 1932.

PROBABILITY AS A BASIS FOR ACTION

Introduction

Applied science involves the discovery of so-called physical properties and laws and the use of these in the fabrication of things to satisfy human wants. I shall assume that this is to be done at a minimum of human effort. Does probability theory give us a rational basis for planning and managing our actions to this end? I propose to discuss this question, and, subject to specified limitations, to give an affirmative answer.

Our present age has been characterized[1] as "the reign of probability". It is generally admitted that inference about the world is inherently of the nature of probability inference. Our every-day as well as our scientific judgments are never more than probable. Asked if it is going to rain tomorrow, we cannot be sure of our answer. Asked what is the magnitude of the charge on an electron or to give some physical law, we likewise cannot speak with certainty. Confronted with any question, all that we can rationally answer is something like this: "I think that such and such is true", "I am quite sure that it is", or "I am almost certain that it is".

From the earliest recorded beginning of the human race, Man has sought truth, partly perhaps for its own sake but certainly because he wanted to use the results of his search in guiding his future actions.

It is desirable to keep in mind from the beginning the distinction between the search for truth and the use of a judgment assumed to be true. As we proceed, we shall see that there are two distinct probability concepts: one, the probability or degree of rational belief that a given judgment is true; the other a measure of the expected frequency of occurrence of a phenomenon within specified limits if the given judgment is true. An extensive critical record of human endeavors toward these ends is contained in the history of the development of theories of knowledge and in treatises on the use of knowledge. One of the outstanding developments holding our attention today is considered under the heads, pure and applied science. Both in the technical and lay

---

1. Weaver, Warren, "The Reign of Probability", Scientific Monthly, Nov. 1930, Volume 31, pp.457-466.

literature one finds much said about Scientific Method. There has been as it were a veritable scramble to apply this so-called scientific method in one field after another. One of the first fields to be cultivated in this way was that of the natural sciences. Now, we hear of the science of history, the science of education, the science of aesthetics, the science of economics, and one of the latest of these "new sciences", the science of management.

When one tries to put his finger on scientific method as a technique, he is reminded of his childhood experience in trying to put salt on the bird's tail. I feel much that way today after having read several treatises bearing on the subject. Thanks to some of the recent writers and in particular, to Keynes, Nicod, Ramsey, Broad, Johnson, Lewis, Whitehead, Russell, Eddington, Eaton, Jeans and Dibble, I have at least been led to a faint appreciation of the nature of the problems involved in establishing a scientific method. Some of these writers have done more for me: they have led me to focus my attention upon the problems involved in the use of deductive and inductive logic - particularly the latter.

Admitting the possibility of intuitive induction of facts and laws of nature but dismissing it on pragmatic grounds, there are left for our consideration the proposed techniques of rational induction. I shall try to indicate the practical significance of the generally accepted conclusion that the use of these techniques in a given problem can only lead to a probability judgment[1]. This assumption will condition what I have to say about the establishment of guides to human action in discovery.

We have still, however, to consider the scientific basis for action involved in the management of human efforts in the application of probability judgments. At least in the limited field which I shall consider, the guiding principle is that which has played so important a rôle in scientific work, namely, the use of hypothesis. It seems that efficient human action in this field depends upon the inherent ability of the individual to establish in a given case an hypothesis of the form: If such and such is true, such and such

---

1. By this I mean a judgment to which we can attach a rational degree of belief less than certainty.

a sequence of events may be expected to fall within certain limits derivable by deductive use of mathematical probability theory. It thus becomes possible to choose limits to the fluctuations within the sequence upon the basis of the given assumption such that, if the underlying assumption is true, we shall not look for causes of variation in the sequence more than some expected number of times which under existing conditions is considered economical.

To get any place we must not take in too much territory. Perhaps some of you remember what happened to the cowboy who sauntered into a Texas barroom and announced that he could lick anybody in the room. Nothing happened. So the bully, somewhat non-plussed, added that he could lick anybody in town. Again nothing happened, and he took in a little more territory with the remark: "I can lick anybody in the whole United States". A moment later, while slowly picking himself up from the corner of the barroom, he was heard to remark, "That's all right, boys, I just took in a little too much territory"

I shall try to profit from the experience of the cowboy in this story and confine my remarks to some of the uses of probability theory as a basis for action in the restricted field of engineering and manufacturing. What I shall say has been prompted by a study of a key problem in this field; namely, the establishment and maintenance of economic standards of quality.

Some Types of Probability Judgment at the Basis of Engineering Action

We shall concern ourselves here with the following four types of judgment:

I. Thing B is of standard quality A.

II. X is the cause of Y.

III. The expected value and standard deviation of a statistical variable, such as a physical property, are such and such, respectively.

IV. The law of relationship between Y and X is such and such.

The first type of judgment is of almost universal significance. The judgment that B is of standard quality is perhaps but the outgrowth of one of the oldest forms of judgment lying at the basis of common knowledge. In other

words, it is the kind of judgment that is at the basis of representative symbolism in which we attempt to characterize for common use any one of a number of things by some particular symbol. Schematically, of course, we may represent the situation somewhat as in Fig. 1, in which A is assumed to be the symbolic representation for the standard of a given kind of thing. Along side of this we conceive of a thing B which is or is not the same as the standard, or as we say in the present connection, is or is not of standard quality. Thus A might be the distance between two notches on the platinum iridium bar kept in Paris and accepted as the standard length of the meter and B might then be a thing produced in an attempt to reproduce this standard of length A.

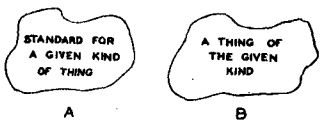


Fig. 1

Similarly, A might be the accepted standard of a piece of a given kind of apparatus or of some kind of manufactured product as, for example, a foodstuff, whereas B would then be something supposedly of the same kind.

The practical judgment at the basis of the use of such a standard is: B is of standard quality A.

A typical example of the second kind of judgment in which X is judged to be the cause of Y would be that the chemical impurities present in corrugated iron are the cause of the corrosion of the iron when used in culverts in a given locality. Tables of values of the so-called physical constants contain many judgments of the third kind, whereas all of the so-called physical laws whether statistical or functional are illustrations of judgments of the fourth kind.

As a starting point, it is necessary for us to consider briefly the meaning of the elements which enter into judgments of the character just described. In connection with the type of judgment involving the statement that B is of standard quality A, we are apt to think of this statement at first as applying to the real parts of B and A. In other words, upon first approach to this problem, we are likely to try to conceive of the "reality" of B as being of standard quality just as physicists and natural scientists have often conceived of physics as dealing with the realities of the physical universe.

We need but read a few of the comprehensive discussions of the subject of appearance and reality by such men as Bradley, Broad, Whitehead, Russell, Joad and others to see that there has been and still exists considerable confusion in the minds of philosophers and scientists alike as to the nature of the difference between appearance and reality. There does seem to be, however, general agreement among many modern logicians and natural scientists that there is a certain givenness in an object which gives rise to the sensations which we experience and interpret in the form of concepts which are then introduced into empirical judgments. As to the exact nature of this givenness, there are again differences of opinion, as such writers as Joad, Eddington, Jeans, Dibble, Broad, Lewis, Stace, and others clearly indicate. Fortunately so far as we are concerned in this elementary and practical discussion, such uncertainties need not cause serious difficulty, because we shall content ourselves with a consideration of judgments involving concepts which arise from the direct experience of physical objects.

Even here, however, we meet with one rather serious difficulty in that the awareness may not always be reduced - at least at the present time - to what Eddington and others refer to as pointer readings, or as an engineer might say, measurable quality characteristics of the thing. This is particularly important in the consideration of the first type of judgment. The judgment, B is of standard quality A may, and usually does, refer not only to those characteristics which are measurable in the physical sense at the present time but also to the characteristics such as beauty, aesthetic value, etc., which play such an important rôle in our appreciation or evaluation of many manufactured products. We need only turn to the discussion of measures of value as given so admirably by such mean as Laird, Perry, Carritt and others to see how far we are at the present time from being able to establish the quantitative measures of certain aspects of the thing of which we are aware. In other words, it may perhaps be safely assumed that judgments in such cases are at least partially subjective. Hence I shall limit my consideration here to probability judgments in which meter readings or functions of meter readings become the elements of the judgment. I think that with this limitation we

can safely assume that there is comparatively general agreement that such aspects of reality have an objective existence.

Confining our attention to this aspect of reality as used in the four types of judgments to which our discussion is limited, there seems to be quite general agreement among logicians and scientists alike that the judgments must be inherently of the nature of probability judgments. Without further discussion I shall start by making the following three assumptions which incidentally I believe are consistent with much of the recent literature on the subject.

Assumption 1

Judgments or inferences of the types previously designated as I, II, III, IV, are never certain.

Assumption 2

If the inference or judgment P is connected to the evidence Q though some probability relation, then there is an objective rational degree p_b' of belief in P upon the evidence Q.

By convention we shall take unity corresponding to certainty as a measure of rational belief. If we did not assume the objective existence of p_b' in a given case, it is somewhat difficult to see how far we could proceed toward developing a methodology of arriving at judgments and of interpreting these in daily use.

Assumption 3

The objective degree of belief p_b' in an inference or judgment P is not an intrinsic property like truth but inheres in the inference or judgment through some relation to evidence Q.

To point out the significance of this 3rd assumption, which appears to be consistent with the beliefs of most students of the subject, we need only call to mind a simple illustration: We assume that there exists an objective value of a physical constant such as the charge e' on an electron or the velocity c' of light. We may find estimates of these constants in tables of physical constants. For example, Birge[1] (1929) gives for the electron charge e':

$$e' = (4.770 \pm 0.005) \times 10^{10} \text{ abs. es. units} \qquad (1)$$

---

1. Birge, Raymond T. - "Probability Values of the General Physical Constants as of June, 1929", The Physical Review Supplement, Volume 1, No. 1, July 1929, pp. 1-73.

All that we can say is that his estimate rests upon the evidence which he takes into consideration. Upon the basis of such evidence and Assumption 2, there is some objective degree p_b' of rational belief in this estimate. Let us compare with judgment (1) the following proposition:

The true charge on an electron is e' abs. es. units. (2)

Here e' is the unknown magnitude of the charge. It is obvious that judgment (1) is only of interest as expressing a result based upon given evidence, whereas judgment (2) is apriorily certain and requires no statement of enidence. Of course, (2) like the first law of thought

A is A

may be claimed to be tautological.

From a practical viewpoint, the mere difference between certain and probable inference is significant in that it indicates that we should always consider a probability judgment in relation to the evidence brought forth in support of that judgment. In spite of this, how often we find articles discussing the most probable values of physical constants: how often we find tables of such constants with no indication as to the source of evidence used as a basis of the judgment!

Furthermore, it is generally assumed that as more and more pertinent evidence Q is acquired and taken into account, the degree p_b' of rational belief in a given inference or judgment P may increase or decrease, as shown schematically in Fig. 2. The degree of rational belief p_b' will not, in general, be a uniformly increasing or decreasing function of the amount of evidence. Referring to Fig. 2, we conceive of a degree of belief p_{b_1}' corresponding to a quantity Q_1 of evidence. At least to

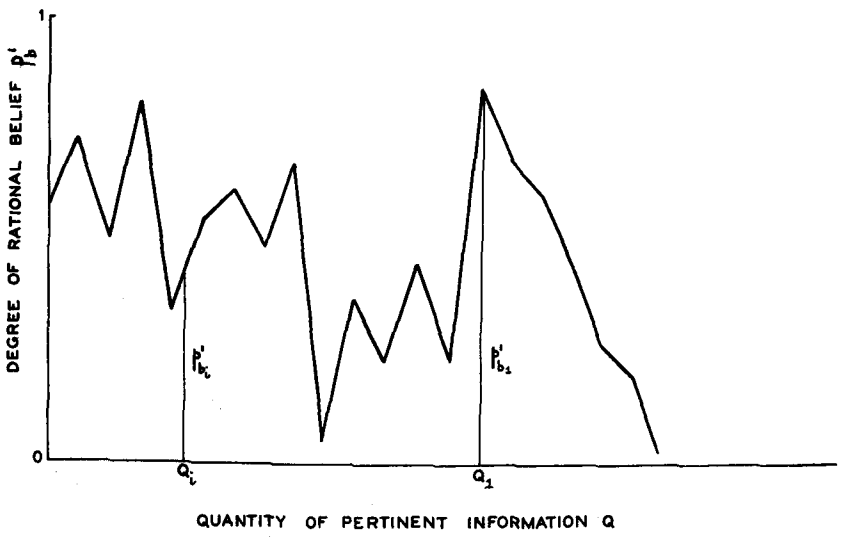


Fig. 2

myself it is a somewhat shocking commentary on our ability to proceed towards certainty to find that this procedure, in respect to degree of rational belief, may behave as indicated schematically in Fig. 2. For example, we may approach as close to certainty as p'_{b_1} only to find that with the addition of a comparatively small amount of evidence the corresponding degree of rational belief may be very much different.

This is of fundamental importance from the viewpoint of action. Thus Keynes[1] emphasizes the point that there must come a time when it is no longer worthwhile to spend trouble in acquiring more pertinent information, although there is no available principle to determine how far we should go in increasing the amount of information. In this sense it appears, therefore, that probability theory does not tell us how close we are to truth in a given induction and does not tell us how far we should go in collecting data.

Three Kinds of Probability Concepts

We have stated four types of judgments and have assumed that each of these is such that the judgment itself bears a certain probable relation to an accepted quantity of information. We have noted the difference between the certain inference of formal logic involved in the proposition, "If Q, then P," and what might be termed the corresponding proposition in uncertain or probability inference, "If Q, then P with a certain degree p'_b of probability". Just as there is a formal or normative science of logic dealing with certain inference, so also is there a partially developed structure of formal or normative logic dealing with uncertain inference. The degree p'_b of rational belief or probability is, as I take it, a definite concept belonging to the formal or normative logic of partial belief. I shall refer to this kind of probability as degree of belief probability. So far as I have been able to discover, this kind of probability is almost universally accepted by logicians. It appears to have something of the formal standing of such concepts as terms, propositions, variables, and so on, of formal logic and mathematics.

Furthermore, there appears to be agreement among many, although not all, logicians that theoretically p'_b is a quantitatively measurable entity.

---

1. Keynes, J.M., A Treatise on Probability, Macmillian Co., 1921, p. 77.

I have so assumed in Assumption 2. There also appears to be agreement that it is theoretically possible to develop a mathematics using this quantitative degree of belief. Ramsey and Keynes are among those who have made interesting attempts to construct such a mathematics. There also appears to be universal agreement among logicians at the present time that we do not know how to measure this degree of belief.

Finally it is of interest to note that there is, so far as I am aware, general agreement that the truth of a probability inference is just the same as that of a certain inference. In other words, it is assumed that, upon the basis of information Q, the inference P may be drawn with the degree of belief p_b' with just as much certainty as a formal certain inference can be drawn, even though it turn out that additional evidence may lead to certain inference of the negative judgment.

In the beginning we mentioned the search for truth as one of the compelling objectives of the human race. Perhaps it is reasonable to believe that the formal processes of mathematics, symbolic logic, and formal logic, including both certain and probable inference, have developed largely as a result of this search for truth. One gets this impression on reading parts of a treatise on formal logic, such as J. N. Keynes' classic. In other words, it seems that we desire to make judgments about the physical world in which we live and we want to make these true or certain judgments. As I have already emphasized, however, we are also concerned with the use of such judgments and hence we are concerned with the meaning of the formal processes of logic and mathematics.

We need go back only a comparatively short time to find at least some mathematicians believing that their subject had meaning. It appears that many comparatively recent philosophers, as for example Kant, accepted as did the contemporary mathematicians the intuitive character of the "axioms" of mathematics and therefore accepted the conclusions or deductions from these axioms as indicating the way in which things necessarily happen in the world. Of course, most mathematicians since then have gradually swung over to the belief so forcefully stated by men like Hilbert and Russell that mathematics is merely

a game played with meaningless marks on paper. During the same period, we find physicists accepting this formal and meaningless character of mathematics.

When we turn, however, to the field of classical formal logic, including the discussion of probable inference, it appears that it is only within the last few decades that attention has been given to the meaning of logic. In fact, one needs only to read the Proceedings of the Aristotelian Society for 1931 to find that there is still being waged a battle between those who do and those who do not believe that formal logic is meaningless in the same sense that Hilbert and Russell consider mathematics to be meaningless.

For example, Schiller[1] in discussing the tautological and meaningless character of the law of thought, A is A, of formal logic makes this comment: "As actually used it is by no means a tautology. For it is used to guarantee the transition from what is taken to be one case of A to another. Hence it should be formulated 'A' is A. This is so far from being a tautology that we are disposed to reject it as a monstrous assumption."

It is significant that in the report of a symposium on formal logic in the Proceedings of the Aristotelian Society for 1931 there is indicated a partial agreement at least to the separation of logic into formal logic and useful logic. It is of even greater interest to note that five papers by the men taking part in this symposium, published in Mind for the current year, indicate a still closer agreement on the need for some such division. When one considers the stimulating critical discussions on this subject by such men as Lewis, Keynes, Ramsey, Nicod and Russell, for example, it is difficult to see how one can even consider the question of the need for such a division as being open to further discussion.

When, however, we consider some of the recent discussions of men like Jeans, Eddington, and Dibble from the field of natural science, and Joad, Cohen, and others from the field of philosophy, we find that the question as to whether or not mathematics has meaning is again being thrown open. Similarly we find a few prominent mathematicians like Brouwer within the last few years

---

1. Schiller, F.C.S., "The Value of Formal Logic", Mind, Vol. XLI, No. 161, Jan. 1932, pp. 53-71.

W. A. SHEWHART'S COLLECTION

raising the issue between intuitionism and formalism[1]. It is desirable, therefore, to keep in mind throughout the following discussion the distinction between mathematics and deductive and inductive formal logic, including certain and probable inference, considered as a game played with meaningless marks on paper, and the use of such formal techniques which can lead, upon the basis of Assumption 2, only to probable inferences or judgments of types I, II, III, IV.

I should hesitate to emphasize this point before this audience if it were not for the fact that we find so much loose talk about levels of significance, degrees of certainty, accuracy, and so on, in some of our very best treatises on probability and statistical method. For example, I read in the latest edition of what I consider to be one of the best books on modern statistical method: "It will be seen from the table that for any degree of certainty, we require higher values of t, the smaller the values of n." On the same and following pages is the discussion of the significance of the mean of a small sample in which the author concludes, upon the basis of a sample of 10 patients each of whom were treated by two drugs, that the effects of the drugs were different. He says: "For n = 9, only one value in a hundred will exceed 3.250 by chance, so that the difference between the results is clearly significant."

Is the practical man to conclude from such a discussion that a significant difference established upon the basis of a sample of 2, the smallest sample size given in the table referred to in the previous paragraph, is just as significant as one established upon the basis of a very large sample, let us say 1000 or perhaps 10,000? Suppose a doctor reads the conclusion that the difference between the effects of the two drugs <u>is clearly significant</u>. Is he to pay any attention to the fact that this judgment is based upon a sample of 10? If the 10 patients constitute a random sample from some normal and homogeneous population in respect to the effects of the two drugs "Student's" theory as here used in the cited illustration tells us something just as definite for small samples as for large. This point I shall illustrate

---

1. "Intuitionism and Formalism", <u>Bulletin of the American Mathematical Society</u>, Vol. 20, 1913-14, p. 81.

R 325

later. The trouble is that as here used, we are not sure that the assumptions apply. We need to ask ourselves, therefore, what degree of rational belief that the assumptions are true in the present case is justified under the conditions. In other words, if we have evidence sufficient to justify such an assumption before a sample is taken, we can then justify a prediction about any size of sample, but if all we have to justify such a prediction is the sample itself, then a sample of 10 certainly does not give us a very satisfactory basis.

We are brought at once to consider in what way we may use judgments. I take it that our general object of trying to acquire knowledge of one kind or another is that we may use it in predicting something about the future. For example, we seek a law or relationship between two variables in order that we may make use of this law in guiding future actions. Similarly we try to measure physical constants, such as the charge on an electron, in order that we may have the results of such measurements as a basis for guiding future action. In other words, having arrived at a judgment of any one of the four types, we proceed to act as if this judgment were true, at least until further data necessitates a modification of the judgment.

Now, it follows from what we have already said that since we can never be certain of our inferences or judgments involving pointer readings or physical measurements, we can never hope to predict with certainty what may be expected. In fact it is generally assumed that no two experimental or observable conditions are exactly the same, so that all that we can hope to do is to try to recognize physical states of affairs usually characterized by some such phrase as "the same essential conditions".

For example, we conceive that certain physical constants, like the charge on an electron, or the velocity of light, are truly constant. However, the only kind of observable constancy even here, is that wherein we try to maintain the same essential conditions of measurement. All that we can hope to do under such conditions, therefore, is to obtain some kind of technique which will enable us to predict what we may expect to get in the future in some understandable way from the viewpoint of probability.

Under such conditions it is found useful to adopt the following assumption, thus introducing the concept of statistical probability:

Assumption 4

If a sequence of events happen under the same essential conditions, where an event is characterized in terms of m quality characteristics, $X_1, X_2, .., X_i, .., X_m$, then the ratio p of the number of events in the sequence of n such events having characteristics falling within the respective ranges[1] $X_1 + dX_1, X_2 + dX_2, \ldots, X_i + dX_i, \ldots, X_m + dX_m$ to the total number n of such events, approaches a definite limit p' as the number n increases indefinitely. This limit is a statistical probability.

Formally we have

$$\underset{n \to \infty}{\text{Lim}_s} \; p = p' \tag{3}$$

where Lim_s stands for a statistical limit which is characteristically different from a mathematical limit in that we never reach a value n_0 of n such that for $n \geq n_0$, the difference $|p - p'|$ becomes and remains less than some previously assigned positive quantity ε.

Perhaps the simplest example of such a sequence is one in which each event can happen in only one of two mutually exclusive ways as in the throw of a coin, the event being the throw of a head, let us say. An observed approach of the ratio p of the number of times a head was thrown in n throws to the number n for a sequence of one thousand throws of a penny is shown in Fig. 3.

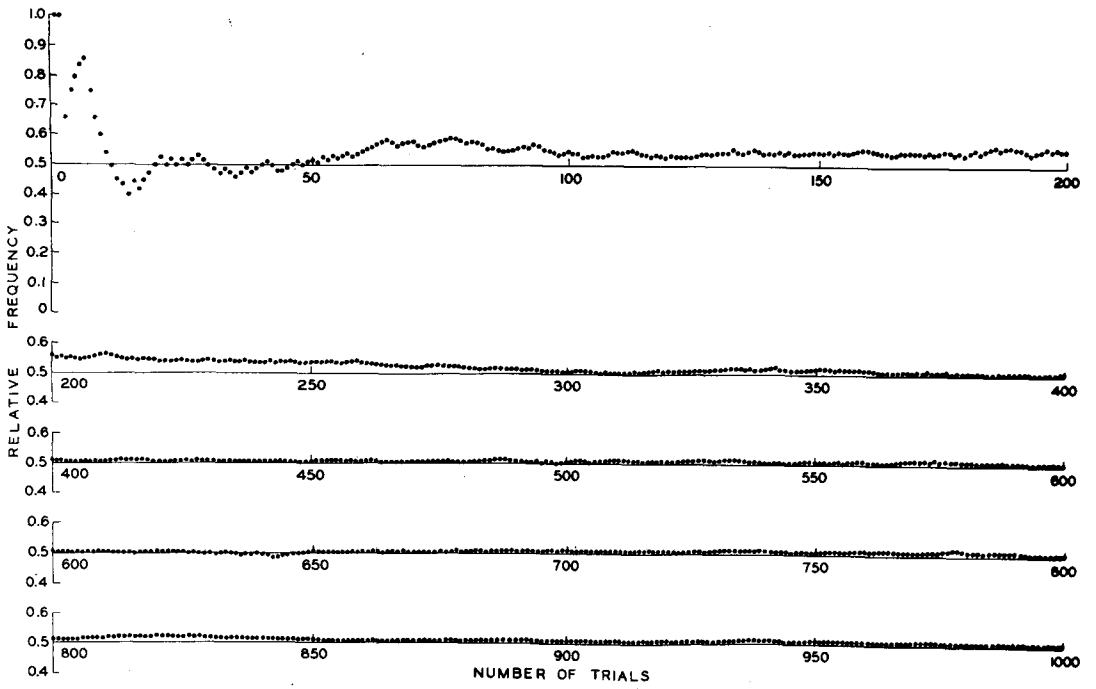


Fig. 3

1. $X_i + dX_i$, here and elsewhere is used as a short-hand expression for X_i to $X_i + dX_i$.

It is generally agreed, of course, that another experimental sequence of this kind made even with the same coin and by the same person would not be exactly the same as that indicated in Fig. 3, although it might be. Furthermore, there is no clear way of defining formally what we mean by "the same essential conditions".

It appears that all we can hope to do under such circumstances is to try to set up some apriori basis for calculating the number of times that we may expect an event to happen in a sequence of n trials made under conditions which we assume or postulate to be essentially the same. Equipped with such an hypothesis we can proceed to examine the series of events as they happen and try to satisfy ourselves as to whether or not they appear to happen in a way consistent with the belief in the assumption of the existence of the same essential conditions. Associated with this concept of limiting frequency as statistical probability, there is the mathematical framework of the calculus of probabilities and distribution theory.

In the last few paragraphs we have gotten a glimpse of what I like to think of as three kinds of probability: degree of belief probability, statistical probability, and mathematical probability. As I have indicated, there is no general agreement as to the meaning of degree of belief probability. There are those like C. S. Pierce, Ramsey, and others who see a way of interpreting this kind of probability as a kind of statistical probability. However, Ramsey at least makes very clear the inherent limitations to this interpretation although we shall not attempt to discuss the issue at this time. The concept of statistical probability is useful in the sense that it gives a basis for forming and using hypothetical estimates of the expected number of times a given event may be expected to happen in a series of n trials. In this sense it has meaning. Of course, there may be a mathematical technique associated with the use of either degree of belief or statistical probability concepts. Furthermore, in the statistical case there are generally two recognized kinds of mathematical techniques depending upon the two methods of estimating the probability, of which we shall say more shortly. In one case the mathematics treats of frequencies whereas in the other case it treats

of combinations and permutations[1].

DISCOVERY - METHODS OF ARRIVING AT PROBABILITY JUDGMENTS

In this section let us consider briefly some techniques of useful logic which may be used to guide our action in the search for truth or, more specifically, in increasing the rational degree of belief in some one or other of the four types of judgment under consideration. Specifically let us consider some ways in which these principles help us to give practical answers to the following questions: What are good data? How many times shall an observation be repeated, or how many measurements? How make efficient use of data? How lay out experiment in such a way as to minimize human effort to attain a given degree of rational belief in a judgment? At least six important techniques of drawing useful inferences call for consideration:

1. Analogy.
2. Simple multiplication of instances.
3. Elimination.
4. Principles of insufficient reason and balanced causation.
5. Consistency with body of accepted empirical knowledge or "theory".
6. Principle of maximum likelihood

I shall touch upon that part of each technique which my limited experience has shown to be significant. I fully appreciate that I do not know many of the ramifications of what is customarily characterized or discussed in the literature under these six headings, and I am sure that I have likely failed to comprehend up to the present time the significance of many of the comments of such writers as Keynes, Nicod, Eaton, and Johnson, on the points involved. Also I share what appears to be the opinion of many that these six techniques, as treated in the literature at least, are not wholly independent one of the other.

Analogy

When we make the judgment, B is of standard quality A, it is true that in order to get anywhere, we must interpret the quality of the standard in terms of physical properties that can be measured or sensed. From a

---

1. Cf. Dodd, E.L., "Probability as expressed by Asymptotic Limits of Pencils of Sequences", Bulletin of the American Mathematical Society, Vol. 36, 1930, pp. 299-305.

practical viewpoint we must still further limit ourselves to a finite number of such properties, although the usefulness of the judgment will depend to a large extent upon whether or not the likeness in respect to the specified qualities carries with it a likeness in respect to unspecified qualities of importance. What we wish to be able to infer is something like this: If A, the standard of comparison, is assumed to have m properties of importance from the viewpoint of use, of which only n may be specified in the present state of our knowledge, and if B is found to conform to the standard A in respect to these n qualities, it follows that B will probably have the remaining (m-n) properties, the probability becoming greater as (m-n) approaches zero.

The need for some method of estimating the degree of rational belief to be associated with a given inference of this kind is particularly great in the specification of a tentative standard to be used in the early stages of production of product in accord with this standard. Suppose, for example, that it has been past practice to make some piece-part from some well-known material having a set of n properties specified as a basis for a standard. Suppose that it is found that a new kind of alloy has these same n properties and can be produced much more cheaply than the kind of material previously used. Immediately the involved problem arises as to whether or not the new material, if substituted for the old, may later be found to have some as yet unknown and undesirable property or properties, such, for example, as rapid deterioration in service. In what way may we use the technique of analogy to tell us something about the degree of rational belief that we may put in the judgment that if B is of standard quality in respect to the n specified characteristics, it will also be of standard quality in respect to the (m-n) other important characteristics? In this question I, for one, strike a snag.

We must keep in mind that premises which taken together have a probability p_1' and a probable inference which would confer on its conclusion the probability p_2' if these premises were certain, will confer on its conclusion the probability[1] $p_1' p_2'$. That is, for example, starting from premises which

---

1. This is supposed true for both statistical and degree of belief probabilities.

taken together give a rational degree of belief p'_{b_1} to the judgment, B is of standard quality A, upon the basis of B and A being certainly alike in respect to n of a possible m-n quality characteristics and a set of measurements on the n characteristics which confers on the conclusion that B and A are alike the probability p'_{b_2}, then the probability of B being of standard quality A is but $p'_{b_1}\ p'_{b_2}$.

I am inclined to believe, however, that the appreciation of the existence of this snag, as it were, is an essential element in the equipment of the one who writes an engineering specification in that it forces him to appreciate as perhaps nothing else will the possible waste of engineering effort in trying to attain an uneconomically high degree of rational belief in the judgment that B is of standard quality in respect to some one or more quality characteristics where he appreciates that he cannot be certain of having specified all important characteristics.

Simple Multiplication of Instances

The particular aspect of confirmation to which I wish to direct attention is that which lies at the basis of the justification for making more than one measurement under what the experimentalist is willing to assume to be the same essential conditions. I personally feel that one of the most important characteristics of the human mind, at least insofar as the particular field we are considering here is concerned, is what appears to be its ability when properly trained to sense - and often to sense correctly in the light of future evidence - what appear to be the same essential conditions. Turn to the history of experimental science in the field of physics and I believe that we will find almost universal agreement that a trained experimentalist in a particular part of this field finally reaches a stage where he is willing to assume that the only way he can approach closer and closer to the objective value of the thing being measured is by increasing the number of observations. Under such conditions Assumption 4 constitutes the logical basis for determining how many measurements to make as soon as one has made up his mind as to how nearly he hopes to get in the statistical probability sense to the objective true value.

Take as a simple case the measurements of the charge on an electron as made by Millikan. Assuming that at the time this set of measurements was taken, Millikan was willing to assume that he had arrived at the state where the only way he could approach closer to the true objective value was by increasing the number of observations, then it follows upon the basis of Assumption 4 that by taking a series of observations $X_1, X_2, .., X_n$ there exist certain statistics or symmetric functions of the n observed values such that for any one of these statistics θ_1, let us say, there exists a statistical limit,

$$\lim_{\substack{s \\ n \to \infty}} \theta = e' , \qquad (4)$$

where e' is the objective charge on an electron. Fig. 4 shows such a statistical sequence for the successive averages of 1, 2, 3, ... 58 observed values. This postulated law of statistical approach for the case of the average has been available for use as a basis for action in discovery of this type certainly ever since the time of Laplace and Gauss. In other words, it has been generally recognized ever since then that the approach in this statistical sense is inversely proportional to the square root of the number of observations, assuming, as both Laplace and Gauss did, that the distribution of errors is normal. As a result of later studies it has been shown that such approach is characteristic for the average for all kinds of distribution of errors where the occurrence of an infinite error is assumed to be impossible, an assumption which we almost certainly have a right to make in practical work.

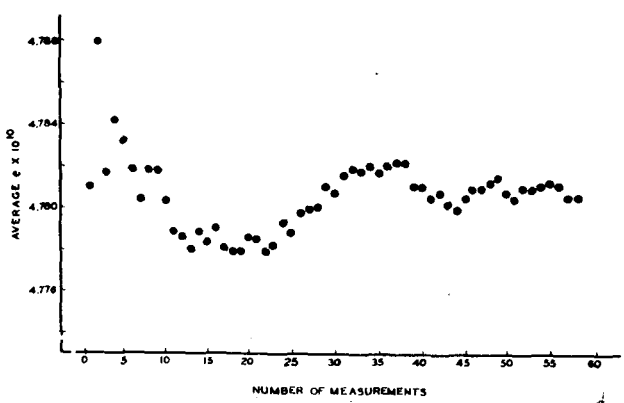


Fig. 4

The concept of statistical limit underlying the approach to truth through multiplication of instances under the same essential conditions gives us basis for determining how many measurements should be taken in the sense which I shall discuss in more detail in a later section.

Elimination

As a basis for the application of the technique of confirmation by simple enumeration, it was assumed that this technique may be applied only after a state has been attained wherein differences in observed values may be assumed to have arisen under the same essential conditions, or as one might say, under the same constant system of chance causes. Figuratively speaking, if we are on the right track to truth, a study of the technique of simple enumeration simply tells us how fast we may expect to approach truth.

In all of this, however, there is that one little word if, the same bothersome if that thwarted our uncritical acceptance of the previously mentioned conclusion that the effect of two drugs is clearly significant. In all such cases our interpretation is certain if the assumptions are true. As every physicist and engineer is well aware, it is usually a long time before one succeeds, to his own satisfaction and the satisfaction of his colleagues, in eliminating assignable or findable causes of variability. About a century ago, Mills attempted to provide a technique for discovering assignable causes of variation which he hoped would lead to certainty in respect to an induction. He gave us the well-known five methods; namely, agreement, difference, concomitant variations, joint method of agreement and difference, and the method of residues. More recent students of the subject, of course, have more or less generally agreed that the methodology provided by Mills in these canons of induction can only lead to probable inference[1].

In general, these methods considered together as a technique emphasize the fact that the process of discovering and eliminating assignable causes of variability increases our degree of belief in the resultant accepted cause of the event in a more or less orderly manner. Thus in the process of measuring a physical quantity, such as the charge on an electron, if it were possible for the experimentalist at the beginning to specify all of the possible sources of such constant error and if he could then eliminate with certainty each one of these possible sources, he would, by this process, have

---

1. One of the most interesting elementary treatments of this particular phase of the subject, from the viewpoint of probability inference, is provided by Eaton in his General Logic, New York: Scribners, 1931.

gotten himself on the right track toward truth where all that he would need to do to approach as close as he wished would be to apply the technique of confirmation by simple enumeration, subject, of course, to the statistical nature of this approach. The trouble is, however, that here again we have that word if, and we do not have what Whitehead[1] calls an ultimate ground upon which to base a rational degree of belief in the judgment that the chosen set of n possible sources of constant error constitute all of the sources.

It is usually assumed that this general process of elimination affords the best method of increasing our degree of rational belief in the resultant judgment based upon the series of observations made after we have convinced ourselves that we have succeeded in eliminating assignable causes.

The history of experimental physics justifies the assumption, I believe, that it is only through comparatively extensive researches, in general, by different men, in different laboratories, by different methods, that the degree of belief that a given method of measurement will give the true result as a statistical limit approached through simple increase in size of sample, becomes large enough to justify taking comparatively large numbers of measurements by any one method.

Now, of course, all of us realize the difficulties involved in applying the canons of Mills because of the statistical nature of many of the variables that must be handled in the practical problem, even though this point, so far as I know, was not considered by Mills himself. As a result, it becomes necessary to develop and apply statistical criteria for detecting assignable causes. I have recently discussed elsewhere[2] at some length this phase of the subject in relation to engineering problems. I wish here to emphasize again the point there made that the application of such tests involves a choice of statistical test, a choice of method of estimating the parameters entering that test, and a choice of limits to be used. The necessity for making such choices simply introduces another source of un-

---

1. Whitehead, A.N., Process and Reality, New York: Macmillan Co., 1930.
2. Shewhart, W.A. - Economic Control of Quality of Manufactured Product, New York: D. Van Nostrand Company, 1931.

certainty in addition to those long considered in the application of the principle of elimination where it is not necessary to appeal to statistical criteria.

What guide to action does probability theory give in the process of elimination, particularly when we appeal to the use of statistical criteria? In the first place, for reasons which I have given in the above reference, such criteria should never be made the sole basis for judgments as to whether or not assignable causes are present. To illustrate, let us consider the oldest criteria of this type, namely, those proposed in the literature as a basis for rejection of observations. These are usually based upon the assumption of a normal law of errors in which estimates of the expected value $\bar{X}'$ and standard deviation σ' derived from the sample are substituted. One such rule is that of Wright and Hayford quoted in many books on theory of errors and least squares. As quoted by Brunt[1] it is: "Reject each observation for which the residual exceeds five times the probable error of a single observation. Examine carefully each observation for which the residual exceeds 3.5 times the probable error, and reject it if any of the accompanying conditions are such as to produce lack of confidence". Expressed in terms of standard deviation the rejection limit is 3.373.

Personally I do not believe that an observation should ever be rejected upon the basis of application of such a criterion alone. Assume for the sake of argument that the average $\bar{X}$ and standard deviation σ of a sample of n observations are taken as estimates of $\bar{X}'$ and σ' respectively. Under these conditions we may say: If the errors are distributed normally about $\bar{X}' = \bar{X}$ with a standard deviation $\sigma' = \sigma$, then the mathematical probability of an observation falling outside the range $(\bar{X}' \pm 3.373\ \sigma')$ is .00076. Were I certain that the assumptions involved in the if were justified, I would not need to go further because I would already know the expected value $\bar{X}'$ to be equal to $\bar{X}$. So soon as we throw open the question as to the degree of rational belief in the assumption of normality and the assumptions $\bar{X}' = \bar{X}$ and

---

1. Combination of Observations, Cambridge University Press, 1923, p. 132.

$\sigma' = \sigma$, we get into deep water, as previous and later discussion clearly indicates. In other words, we cannot be <u>certain</u> - in fact most of us will agree that we are much less than certain - that the probability of getting an observation outside these limits is .00076. In addition, we must keep in mind that there is nothing apriorily sacred about the choice of 3.373 σ' or any other multiple of σ' as a basis for establishing such limits.

For such reasons I would consider using such a criterion only to assist me in picking out those observations which I should particularly scrutinize. Rejection or retension of the observation would be based upon the results of such scrutiny. My colleague, T. C. Fry, states as his rule of rejection: "Discard observations only when you are convinced they are bad; never simply because you are not convinced they are good". To this rule I heartily subscribe.

There is another difficulty with applying formally any test such as that of Wright and Hayford, although I fail to find any discussion of it in the literature. If we were to reject an observation only when it fell outside the range $\bar{X} \pm 3.373\ \sigma$, it is obvious that unless the number of observations is at least 15 no observation would ever be rejected. This follows at once from Tchebycheff's theorem that at least $(1 - \frac{1}{3.373^2})$ n of the observations lie within the range $\bar{X} \pm 3.373\ \sigma$.

Of course much the same line of argument as that concerning tests for rejection of observations applies to the results attained by application of any criterion to test for significant differences. If, for example, a sample of a new kind of alloy made by one producer, when compared with a similar sample from another producer, is found upon the basis of some accepted criterion to be significantly different, as we often say, this simply means, that the difference is such that <u>if</u> such and such assumptions made the basis of the criterion and the methods of estimating the parameters therein are true, then we have a right to expect a difference as large as that observed only once in so often.

Now, upon the basis of a given set of assumptions, it becomes possible to calculate quite rigorously the expected number of times that the

observed phenomenon will fall outside the limits of the criterion and hence be the cause of action leading to the search for trouble, even though the trouble is not there. The justification of the assumptions underlying the use of such a criterion is a matter that rests entirely, so far as I see, upon the engineer or scientist making use of the criterion, in just the same way that the justification of the methods used in eliminating assignable causes of variability in the measurements of the charge on an electron by a given method must be left to the experimentalist. In other words, put in a more popular way, such criteria enable a good scientist or good engineer to do his work efficiently and this efficiency becomes of real economic significance in such a field as the control of quality of manufactured product.

The criterion to be used in a given case should be the most efficient one to catch the kind of trouble that the engineer or scientist believes to be present in his particular case. In other words, the engineer tries to set up a criterion that will catch the kind of trouble that he thinks is present in a particular case and he will use probability and statistical theory to enable him to put such limits on the criterion as will keep him from wasting too much time looking for trouble of the kind he thinks may be there, if this particular kind of trouble is not there.

I think enough has been said to indicate the difference between the use of criteria for detection of assignable causes of variability in the sense which I have found useful, and the use of criteria quite similar mathematically as a basis for rejection of observations or, in a more general case, as a basis for deciding whether or not a population is homogeneous. In the first case the use of such a criterion is but a step in the process of arriving at a final judgment whereas, in the second, it is to a large extent made the basis of the final judgment as in the previously considered statement that the effects of the two drugs are significantly different.

Insufficient Reason and/or Balanced Causation

I shall content myself in this connection with the statement of the following assumption which forms a basis for such computations:

Assumption 5

If an event can happen in any one of several mutually exclusive ways, all of which are equally likely, and if a certain number of these be called favorable, then the ratio of the number of favorable ways to the total number of ways is equal to the probability p' that the event will turn out favorably.

It seems to me that this assumption either explicitly or implicitly underlies all of our applications of probability theory as a means of predicting the future. It is, of course, one of the basic postulates of the calculus of probabilities, and, so far as I see, it is involved in one way or another in sampling theory in some such concept as homogeneity, random condition, or same essential conditions. As Camp[1] has recently pointed out, this probability is generally used to pierce the veil of the future. Thus, as a simple illustration, assuming that the probability of throwing a head is one-half, apriori probability may be used to tell us much about a sequence of throws. In such a simple case, however, the practical man is liable to lose sight of the fact that there is a mathematical framework and at the same time a degree of rational belief that this framework applies in the particular case. Just as soon as the practical man appreciates this situation he sees that the fundamental problem involved in the application of the mathematics of probability is the establishment of a satisfactory rational degree of belief in the underlying assumption involved in some such phrase as "equally likely", "random", "homogeneous", or "same essential conditions".

The need for laying emphasis on the degree of rational belief aspect of applications of probability is of particular interest when we try to interpret or evaluate the significance of a sample. For example, in our discussion of the approach to truth by repetition, we introduced the concept of "same essential conditions", which simply means that the probability of an observed value of a variable falling within a given range is supposed to remain constant throughout the series of measurements. When this condition is maintained, we appear to approach some objective value in a statistical limit sense. Following this discussion, however, we considered briefly the very

---

1. Camp, B.H., "Definitions of Probability", American Mathematical Monthly, May, 1932.

difficult problem of elimination of assignable causes of variability so as to arrive at a condition where we could approach, in the statistical sense, some objective value.

When we come to consider the simplest case of sampling, such as taking a finite sample from an assumed homogeneous lot, we again must focus our attention upon the need for justifying our belief that the lot is homogeneous. It is easy to do this, of course, if we can do something which is equivalent to shaking the lot but, as I have pointed out elsewhere[1] at some length, in a particular case it is not often so easy to do this. Thus there are many kinds of manufactured product which we cannot thoroughly mix, as it were, and in such cases, of course, the sampling problem becomes much more involved than in the simple case of sampling from a homogeneous lot.

As all of you know and as I shall try to emphasize in a succeeding section, a sample by itself means little enough when taken under homogeneous conditions, and when not taken under homogeneous conditions, it means much less. It was the appreciation of this fact that early led some of us in the engineering field to lay emphasis on the need for finding out more about the causes of lack of homogeneity so that we could divide the product into rational sub-groups[2]. To do this, of course, amounted to laying emphasis on the problem of elimination as discussed in the previous section.

It is of interest to note that the application of criteria for detecting assignable causes of variability in applying the process of elimination leads to certain economies in the processes of production. One of the most important economic consequences of such work, however, is that it gives us the kind of positive information about a given product which we need to have so that we can set up specifications for economic standards and thus attain a product whose component piece-parts, including raw material, may be divided into homogeneous groups, thus making possible efficient design in terms

---

1. Shewhart, W.A., "Random Sampling", Bell Telephone Laboratories Monograph B-576.
2. The term rational is introduced here to emphasize that such division is not based solely upon numerical criteria but involves the element of rational judgment at every step.

of the inherent variabilities in the homogeneous groups.

Consistency With Body of Accepted Empirical Knowledge or "Theory"

There seems to be quite general agreement among logicians that one of the most important ways of increasing the degree of belief in an empirical judgment is to show that it is consistent with the general body of knowledge in that particular field. Consider, for example, the measurement of Planck's constant, the velocity of light, and the charge on an electron, whose objective values we shall denote by h', c', and e', respectively. There is, of course, the following theoretical relationship between these quantities:

$$\frac{h'c'}{2\pi e'^2} = 137. \qquad (5)$$

Sir Arthur Eddington, for example, in his latest discussion of this relationship states that he believes that the value 137 is there obtained by pure deduction employing only hypotheses already accepted as fundamental in wave mechanics[1]. If we accept the hypotheses of wave mechanics and Eddington's conclusions, Eq. 5 becomes a condition which the measured values of the constants must satisfy.

Let us consider a little further some of the problems involved in trying to estimate the degree of rational belief contributed by tests of consistency with previously obtained empirical knowledge. To do this we shall take the problem of measuring the charge on an electron. As a result of Millikan's measurements up to 1917 he gives as his estimate

$$e' = (4.774 \pm 0.005) \times 10^{10} \text{ abs. es. units.} \qquad (6)$$

For the time being, we shall assume that (6) may be taken as a symbolic statement of the judgment that the objective charge e' on the electron lies within the range there indicated. Now in accord with Assumption 2, there is a degree p'_{b_1} of rational belief in this judgment. We do not know the magnitude of this degree of belief but we may represent it schematically on a scale of belief as in Fig. 5.

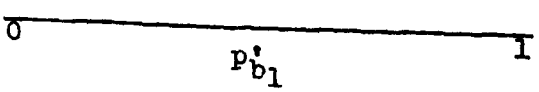


Fig. 5

1. "Theory of Electronic Charge", Proceedings of the Royal Society, Series A, Vol. 138, No. A-834, pp.17-41.

Now, as a result of considering the data available in 1929, Birge, as already noted, gives a corresponding judgment

$$e' = (4.770 \pm 0.005) \times 10^{10} \text{ abs. es. units,} \qquad (1)$$

where (1) is to be interpreted in the same way as (6). Similarly, in 1932, considering the data then available, Birge gives a judgment

$$e' = (4.7688 \pm 0.0040) \times 10^{10} \text{ abs. es. units.} \qquad (7)$$

Thus we have before us for consideration three empirical judgments. Judgment (6) is based upon Millikan's data in 1917, judgment (1) upon Millikan's data of 1917 together with similar measurements up to 1929, and judgment (7) upon all of the data up to 1932. In other words, we have three judgments and three quantities of information, let us say Q_6, Q_1, and Q_7, respectively.

Let us consider first the question: How is our degree of rational belief in judgment (6) based upon information Q_6 modified through consideration of information Q_1 and Q_7 respectively. In accord with Assumption 2 there is a degree of rational belief in judgment (6) for each of these three quantities of information. Let us call these p'_{b_1}, p'_{b_2}, and p'_{b_3} respectively. For example, Milikan[1] himself in 1930, making use of information Q_1, again arrives at judgment (6). What would be the relative positions of p'_{b_1}, p'_{b_2}, and p'_{b_3} if indicated schematically in Fig. 5? I am still looking for an answer to this question.

Now, there is another question, the answer to which is of practical significance: What are the relative magnitudes of the degrees of belief corresponding to the three judgments (1), (6) and (7)? I am still looking for an answer to this question.

If we assume that Eq. 5 is a condition in respect to the general body of knowledge which must be satisfied by the three constants, h', c', and e', what effect does this piece of information have upon each of the rational degrees of belief which we have just been considering, because it surely must have some effect in accord with our assumptions? I am still looking for an answer to this question.

---

1. "Most Probable 1930 Values of the Electron and Related Constants", Physical Review, May 15, 1930, pp. 1231-37.

The questions we have just considered in respect to the relative magnitudes of rational belief based upon available techniques of induction are, I believe, about the simplest kind of practical questions of this character which we may raise. When we come to consider some of the more complicated problems involved in trying to determine our degree of rational belief in a law of relationship the problem becomes far more complicated. This is particularly true when one attempts to weigh the significance of the assumptions underlying the mathematical techniques involved in reaching judgments (6), (1) and (7), including the choice of functional relationship.

In this connection, as most of you know, there always has been and still remains a real issue among scientists who have looked into this question, as to how much significance can be attached to the difference between such techniques. Thus Birge makes use of the method of least squares. Millikan, on the other hand, arrives at his result by a process involving graphical determination. Harold Jeffreys in an article which has just come to my desk suggests a method of modifying the theory of errors to account for small sample sizes and to take into account apriori hypotheses as to the distribution of the standard deviation[1]. Now, the application of the methods of Birge, Millikan, and Jeffreys to the same set of data such, for example, as Millikan's 1917 data, would lead to three judgments of type (6) upon the basis of the same set of data, that is we would have three ranges based upon the same set of data. The question that bothers me is: Which one of these methods should I use, admitting, of course, that there is some importance to be attached to the fact that an analytic one always leads to the same results in the hands of different individuals, while the graphical one does not. In other words, associated with the three judgments of the type (6) obtained through the application of these methods to the same set of data, in accord with Assumption 2, there are three different degrees of rational belief, but the question as to the relative magnitudes of these three seems to be far more involved than the corresponding questions which we have already considered. Whenever I ponder over such questions I remember Lord Rayleigh's remark that the theory of errors was a

---

1. "Theory of Errors and Least Squares", Proceedings of the Royal Society, Series A, Vol. 138, No. A-834, pp. 17-41.

good thing to read up on and then forget, and I can sympathize with the lay engineer or scientist attempting to arrive at judgments of the types (6), (1), and (7), when he is faced with the problem of trying to choose from among such techniques as even the simpler ones we have just considered in connection with the names of Millikan, Birge, and Jeffreys.

In the light of such considerations, there is a need for revision in the statements found in many practical treatises on the theory of errors and the theory of statistics. I cannot take time here to give more than one illustration although it would be easy to extend this list. I shall choose my illustration from the last paper which has come to my desk dealing with the application of the theory in industrial science. In the conclusion of this article after discussing some of the recent work of "Student", R. A. Fisher and others, the following statements are made: "Statistical theory provides him with a measure of the accuracy of his result. When only a small number of determinations have been made the chemist finds he cannot be so sure of his final result. He finds, however, a range of values within which the true result most probably lies. ....." I must confess in the light of the discussion immediately preceding that I am not able to state the range of values within which a true result, such as the charge e' on an electron most probably lies. For example, I fail to see how statistical theory will provide me with a measure of the accuracy of Millikan's result. In brief, I believe that such a claim takes in too much territory and in so doing is harmful because it puts before the practical man an objective which sooner or later, upon the basis of our present knowledge of statistical theory, must be rejected.

The Principle of Maximum Likelihood

Anyone who has seen an accident, listened to a group of witnesses tell the judge how they think it all came about, observed the judge weigh the evidence and give the final decision, has witnessed the application of the principle of maximum likelihood. In other words, given an observed event, there may be several alternative ways of accounting for it. We often try to choose that explanation or theory upon the basis of which the given event is most likely, as we say. Of course, if one considers critically the signifi-

cance of the terms "most likely" or "most probable", etc., used under such conditions, he will find that they do not all mean exactly the same thing. I do not, however, wish to consider this particular phase of the subject here but rather to turn for a moment to the significance of a statistical estimate of a parameter determined by the method of maximum likelihood as defined by R. A. Fisher[1] in the following terms:

"If in any distribution involving unknown parameters λ_1', λ_2', λ_3' ..., the chance of an observation falling in the range dX be represented by

$$f(X, \lambda_1' \sigma, \lambda_2' \sigma, \lambda_3' \sigma \ldots)dX$$

then the chance that in a sample of n, n_1 fall in the range dX_1, n_2 in the range dX_2, and so on, will be

$$\frac{N!}{(n!)_p} \left(f(X_p, \lambda_1' \sigma, \lambda_2' \sigma, \lambda_3' \sigma \ldots)dX_p \right)^{n_p}.$$

The method of maximum likelihood consists simply in choosing that set of values for the parameters which makes this quantity a maximum, and since in this expression the parameters are involved only in the function f, we have to make $\Sigma(\log f)$ a maximum for variatons of $\lambda_1' \sigma$, $\lambda_2' \sigma$, $\lambda_3' \sigma$ ..."
The method was introduced and used much earlier by Gauss but has been studied at length by Fisher in several recent papers.

Let us consider a very simple problem. The following sample of four:

1.7
.2
1.4
.5

is the first of a series[2] of 1000 such samples drawn from an experimentally normal universe. Subject to minor corrections allowing for the fact that no experimental distribution can be rigorously continuous, the distribution in the bowl from which this sample was drawn is fixed by the two parameters $\bar{X}'$ and σ' in this particular case where we know the distribution to be normal.

---

1. "On the Mathematical Foundations of Theoretical Statistics", Phil. Trans. Roy. Soc. of Lond., Series A, Volume 222, pp. 309-368, 1922.

2. Shewhart, W.A., Op. Cit. p. 442.

Let us consider the problem of estimating $\bar{X}'$ and σ' from the sample.

The a priori probability of getting a sample of n values $X_1, X_2, \ldots X_n$ under the assumed conditions is

$$\pi = \left(\frac{1}{\sqrt{2\pi}\sigma'}\right)^n e^{-\frac{(X_1-\bar{X}')^2+(X_2-\bar{X}')^2+\ldots+(X_n-\bar{X}')^2}{2\sigma'}} . \quad (8)$$

Whittaker and Robinson[1], using the method of maximum likelihood, take as estimates those values which make (8) a maximum and in this way get from the two equations,

$$\frac{\partial \pi}{\partial \bar{X}'} = 0 \text{ and } \frac{\partial \pi}{\partial \sigma'} = 0 \quad (9)$$

the estimates,

$$\bar{X}' = \bar{X} \text{ and } \sigma' = \sigma \quad (10)$$

where $\bar{X}$ is the average and σ is the standard deviation of the sample.

If instead of applying the method of maximum likelihood to (8) we apply it, as Irwin recently suggested, to the distribution $f(\sigma)$, we get

$$\sigma' = \sqrt{\frac{n}{n-1}}\,\sigma . \quad (11)$$

Substituting the observed value of σ in (10) and (11), we get respectively:

$$\sigma' = \sigma = .618 \text{ and } \sigma' = 1.155\sigma = .714.$$

Granting for the sake of argument that the mathematical basis for (11) is better than for (10) as some maintain, the practicing statistician is still faced with the problem of deciding how good such an estimate really is or perhaps better, what it means? Before considering this question, however let us throw into the ring certain other proposed estimates.

For example, some have suggested as an estimate the solution of

$$\frac{\partial f(\sigma)}{\partial \sigma} = 0,$$

from which we get

$$\sigma' = \sqrt{\frac{n}{n-2}}\,\sigma, \quad (12)$$

where $f(\sigma)$ is the distribution function of the observed standard deviation.

---

1. Calculus of Observations, Blackie and Son, London, 1924, pp. 186-187.

To these estimates we might add those expressed in terms of the mean deviation of the sample as these are often recommended in treatises on theory of errors and in engineering handbooks. There is also a group of estimates based upon inter-quartile and other ranges.

To get on with our argument, let us assume, however, that we could eliminate estimates not based upon some multiple of the observed standard deviation on the score of efficiency. We have then left for consideration various estimates of σ' all of which can be written in the form

$$\sigma_{\bar{x}} = c\sigma'$$

where c is a constant depending upon the particular choice of the basis for our estimate.

Of course, we know the a priori distribution function for the standard deviation of samples of n drawn from a normal distribution. For n = 4, this distribution expressed in terms of σ' is shown in Fig. 6.

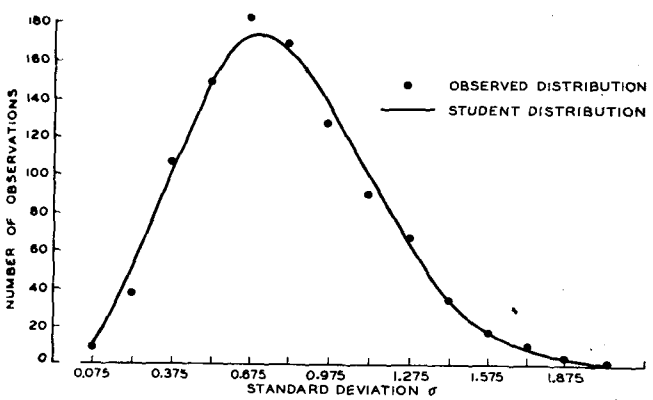


Fig. 6 *8890

The dots following quite closely the contour of this curve represent the experimental distribution of the standard deviations of a thousand samples of four. With this curve before us, let us consider the significance of any estimate based upon the observed standard deviation of the previously mentioned sample. All that we can really say is that the observed standard deviation .618 is one from the distribution shown in Fig. 6, lying roughly within the range zero to $2\sigma'$. Since we can say no more about the observed σ itself in its relation to the set of possible σ's, it seems to me that we can say no more than that about a multiple of σ. In other words, we are forced to face the fact that σ from a small sample is just σ from a small sample.

It may be helpful to others as it has been to myself to consider the significance of the different estimates in the very simple case where M samples of four have been drawn from presumably as many different universes. Consider-

ing only estimates (10), (11) and (12), we would have the following three series:

σ_1	$1.155\sigma_1$	$1.4142\sigma_1$
σ_2	$1.155\sigma_2$	$1.4142\sigma_2$
.	.	.
.	.	.
.	.	.
σ_M	$1.155\sigma_M$	$1.4142\sigma_M$

Assume now that it so happens that the M universes are the same, that is, they are all normal with the same parameters $\bar{X}'$ and σ', although this fact is not known to the investigator. Which of the three sets of M estimates of σ' would have given us the greatest percentage of estimates within a narrow range $\sigma' \pm \Delta\sigma'$? Obviously the answer is set (12) and not the maximum likelihood sets. If, however, $\Delta\sigma'$ is increased sufficiently, the answer becomes set (10).

Now, of course, we may approach this problem by making use of Bayes' theorem as recently set forth in an interesting article by Molina and Wilkinson.[1] Making use of their method and assuming that one may choose an apriori distribution function satisfactory to the given case, one may calculate the probability distribution of the unknown mean. Even here, however, we must at some place or other choose estimates of $\bar{X}'$ and σ'. In the case of estimates of σ' these are usually if not always expressed as multiples of the observed standard deviation σ. Here again, however, we must recognize the fact presented in Fig. 6 that for samples of 4, an observed σ expressed in units of σ' may lie anywhere within a range of approximately 0 to $2\sigma'$.

To all such methods discussed at length in the literature we must add a host of others less well defined. Thus in the case of the sample of four I might argue as follows:

I have never known of an experimental normal universe that was not chosen symmetrical about zero. In the second place it is obviously much more simple to construct such a one than any other. It is reasonable to believe that an experimentalist will try to have as many cells as possible and yet not have an excessive number of chips. Furthermore, because of simplicity, there

---

1. "Frequency Distributions of the Unknown Mean of a Sampled Universe", Bell System Technical Journal, October 1929, pp. 632-645.

are advantages in choosing the standard deviation σ' as unity and marking the abscissae in terms of σ'. All of this is based upon previous experience.

Looking at the sample, we find that the four values are consistent with the hypothesis that cell intervals are in tenths. If they represent $.1\sigma'$, then to cover the range $\bar{X}' \pm 3\sigma'$ would require 61 intervals, and such a distribution, as can easily be shown, would require about 1000 chips. All of this seems reasonable. If the intervals were, say, in $.01\sigma'$, the number of chips required would be several thousand and hence difficult to use experimentally.

Now, assuming that $\sigma' = 1$ and $\bar{X}' = 0$, the probability of getting a sample whose average is not greater in absolute value than the observed average .95 is .66. The observed average is not unreasonable on this hypothesis. Furthermore, it is unreasonable to expect that the average $\bar{X}'$ is identical with the observed average .95. Hence we might conclude that $\bar{X}' = 0$, $\sigma' = 1$, whereas $\bar{X} = .950$ and $\sigma = .618$.

Although I have for the most part confined my discussion to the problem of estimating σ' of the normal distribution in the bowl, somewhat similar remarks could be made about the methods of estimating the average $\bar{X}'$. Even in this very simple case, where one is given a priori that the universe is normal, I think every student of the subject appreciates the difficulties involved in trying to justify a given choice.

Under these conditions does sampling theory serve in any way as a guide to action? My answer is "yes" in that, first of all, it leads us to be cautious about what we say based upon the evidence given by a sample alone, even though it be drawn under such ideal conditions as here considered, unless the sample size is large enough. But how large is large enough? The answer to this question depends upon the circumstances in hand. Fig. 7 shows the approach of the expected value of observed standard deviation σ in a sample of n to the objective σ' of the universe, together with limits expressed in terms of σ' within which approximately 99% of observed σ's of samples of size n drawn from a normal universe may be expected to lie. I have found this to be one of the most helpful pictures to keep in mind when trying to determine

how many measurements to take in a given case. It is interesting to see how rapidly these limits approach the true σ' as a function of sample size n.

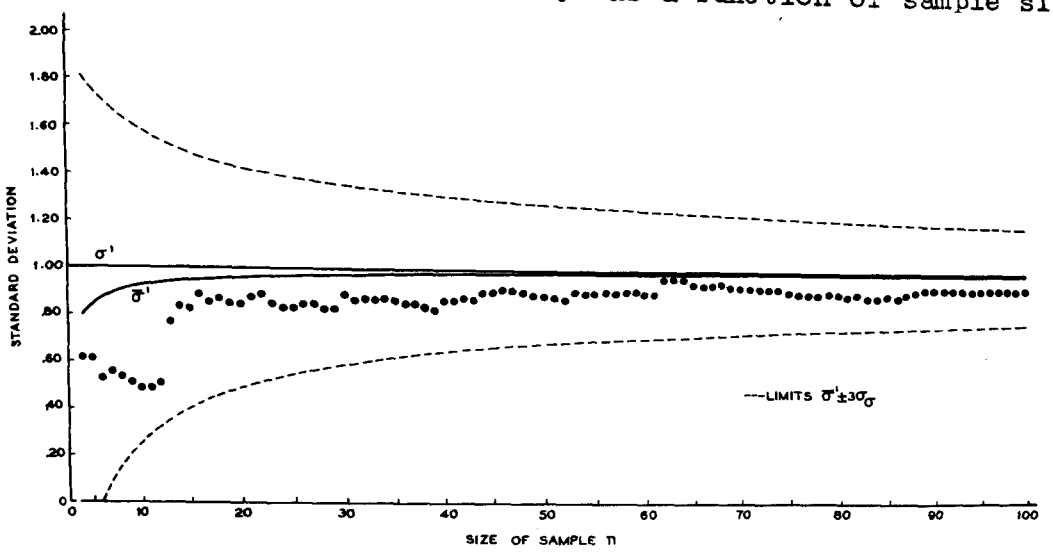


Fig. 7

In a practical case of sampling, we seldom, if ever, know in the sense of the previous simple experiment that the universe is normal. Furthermore, the thing under measurement is usually some objective value and there is always the question as to the existence of assignable causes of variability such as constant errors and erratic fluctuations in the measurements. As an illustration let us assume that instead of drawing a sample out of a bowl, we are making a series of measurements of the charge on an electron. In this case, as previously noted, we must first try to establish a reasonable degree of rational belief in the assumption that assignable causes of variability have been eliminated; that the errors of measurement are distributed normally about the objective value and that the measurements have been taken under random conditions. It usually works out in such instances that before a practical man has come to the place where he is willing to believe in the assumption that a given method will lead to the objective value in a statistical sense simply through repetition of results, he has already been led to make some sort of judgment that the objective value being sought for lies within some agreed-upon range and to experience what he considers to be a rational degree of belief in this judgment which is not much increased by taking more than 5, 10, or 15 readings by a single method, as anyone can justify for himself from dis-

cussions such as those given by Birge and others on the measurement of physical quantities.

At this point we should consider briefly the meaning of probable error as used in the literature. Thus the estimate of a physical quantity λ' is often expressed in the form of

$$\lambda' = \lambda + \varepsilon, \tag{13}$$

where ε is an estimate of $.6745 \frac{\sigma'}{\sqrt{n}}$, σ' being the objective value of the standard deviation of the error of measurement. Confining our attention to the simplest kind of problem, let us consider the meaning of such an estimate based upon the previously mentioned sample of 4 drawn from a normal universe.

"Student" in 1908 derived the distribution of

$$z = \frac{\bar{X}' - \bar{X}}{\sigma}$$

as a function of sample size. Upon the basis of this work it follows that before any samples of a given size n are drawn, we may say something about a sequence of such samples. In other words, Student succeeded in calculating the probability that the mean $\bar{X}$ of a sample of n, drawn at random from the normal population in the bowl, will not exceed in the algebraic sense the mean $\bar{X}'$ in the bowl by more than z times the standard deviation in the sample. Thus taking the case $n = 4$, we may say apriori before a sample is drawn, that the probability is .50 that the difference $|\bar{X}' - \bar{X}|$ for a sample of 4 will not be greater in absolute value than $.44\sigma$.

Stated in another way we may say before any samples of 4 are drawn that if we draw a sequence of M such samples and if after the M samples are drawn we set up the M ranges $\bar{X}_1 \pm .44\sigma_1$, $\bar{X}_2 \pm .44\sigma_2$; ... ; $\bar{X}_M \pm .44\sigma_M$ we may expect 50% of these ranges to include $\bar{X}'$. Fig. 8 is such a series of ranges constructed for the first 100 samples of four drawn from the same normal universe from which the previously mentioned sample (the first sample in this series) was drawn. Now the $\bar{X}'$ in the bowl is 0.0 and we see that 51 of the 100 ranges include this value - a close check on theory.

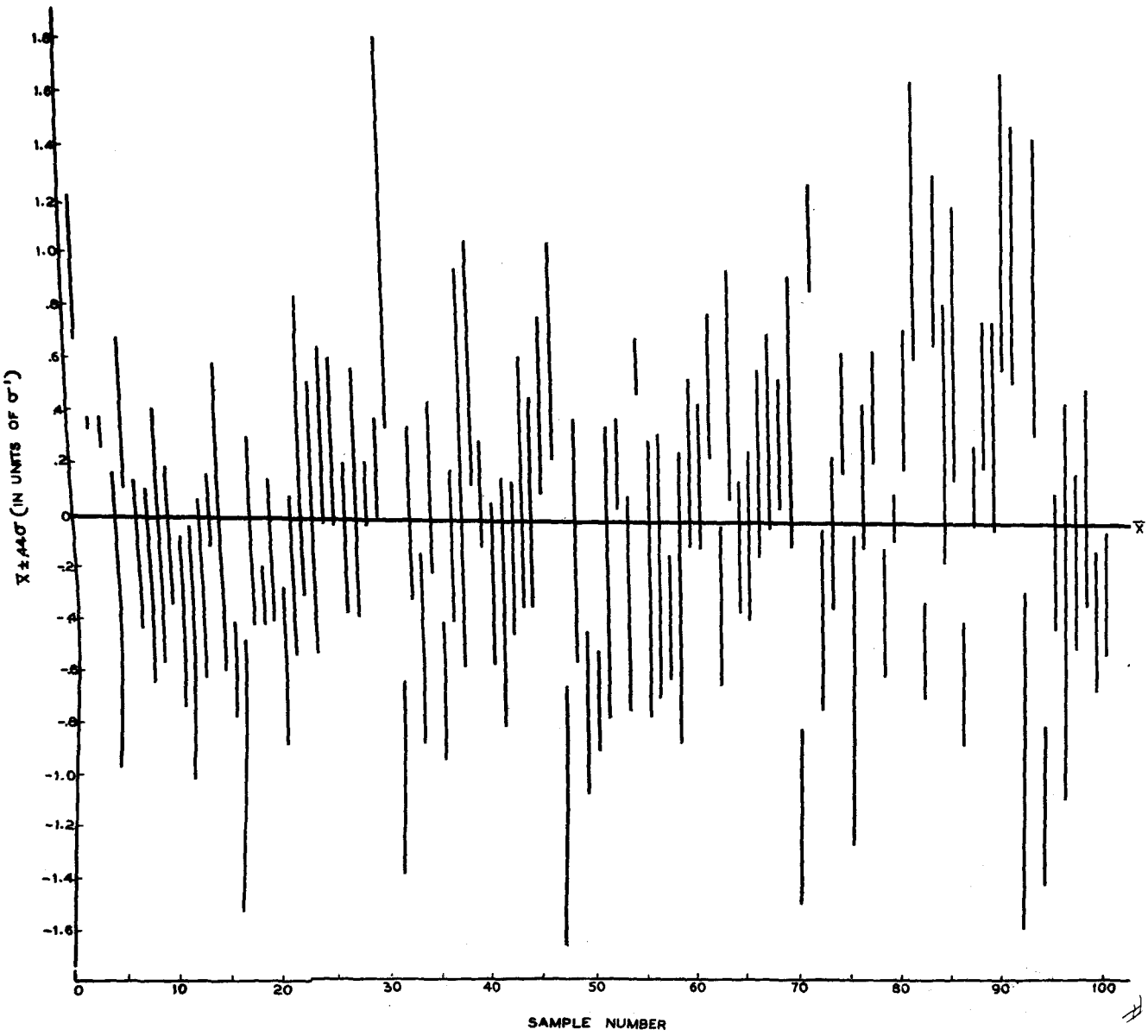


Fig. 8

In the simple case where we have M samples known to be drawn from M different normal universes "Student's" theory enables us to set up a range upon the basis of each sample such that the expected number of universes whose true average values lie within the associated ranges could be made any fraction of M that we please. It should be noted that "Student's" theory, however, has to do only with a sequence of samples.

Enough has been said to indicate the meaning of estimate (11) when used in "Student's" theory but this use must be kept separate in our minds from the use of an estimate to tell something about the universe from which the sample is taken.

Incidentally it is of interest to note that this same distinction exists independent of sample size. Thus if we take samples of 1000 from the previously considered normal universe and set up a range

$$\bar{X}_i \pm .6745 \frac{\sigma_i}{\sqrt{n-1}}$$

for the ith sample, where i takes the values one to m, the number of such samples, then we may expect, before we have taken any such samples, that 50% of the ranges thus established will include the point $\bar{X}' = 0$. Making use of available data, Fig. 9 shows such ranges for four samples of 1000 each. For sake of comparison this is drawn to the same scale as in the case of ranges established upon the basis of "Student's" theory. In the first place as is to be expected the width of the ranges is much less but it is interesting to note that only two of the ranges include the true value $\bar{X}' = 0$. Dame Fortune happened to be good to me in this particular case making a close check between theory and practice!

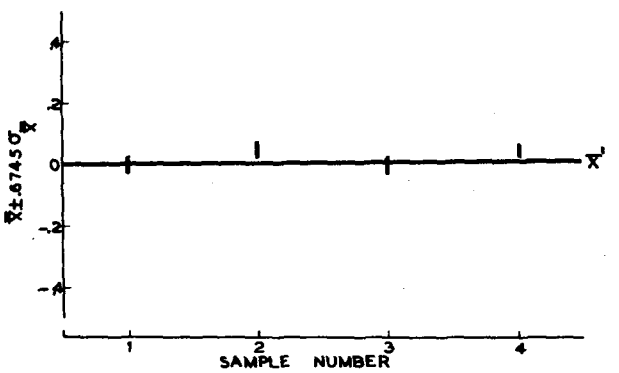


Fig. 9

In other words, if we know that we are sampling from a normal universe, apriori distribution theory enables us to set up ranges, such as I have just been considering, that may be expected to include the true value in the bowl any given fraction like .50 of the number of times that such samples are drawn, independent of the size of sample. Upon the basis of theory and experimental evidence, such as we have just considered, it seems reasonable that we may entertain the same degree of belief in such a prediction in the case of small samples as we can in the case of large samples. It is very important though to note the significance of the size of sample in closing up, as it were, the ranges. For this reason it seems to me that such a range should always carry with it a statement as to the size n of sample to which it applies.

When, however, I focus my attention upon the problem of estimation

and ask myself, after a sample of a given size has been taken, what is the value of the two parameters of the population in the bowl, I immediately confess that I do not know the answer. If I were faced with the necessity of having to make a decision as to the estimate of the parameters in the bowl, upon the basis of information that I now possess about experimental sampling, and upon the basis of the information given by the previously considered sample of four, I should take as estimates $\bar{X}'$ equals zero and σ' equals unity, for the reasons which I set forth earlier in this section. In all cases in trying to answer a question of this type I am in agreement with the conclusion of Molina and Wilkinson, expressed in their article previously referred to, that the data given by the sample should be considered together with all other available information. Of course, this cannot always be done in the comparatively simple way involving the use of Bayes' theorem.

In the last few paragraphs, of course, we have limited our discussion purposely to the simplest kind of problem where it is known a priori that the distribution is normal and that the series of observations are free from assignable causes of variability. When we step over into the practical field where we must take these factors as assumptions, the problem naturally becomes much more difficult and our estimate of the degree of belief that we may entertain in the underlying assumptions must become the controlling factor in our analysis and interpretation of a given set of data.

CONCLUSIONS

I have tried to indicate the significance of keeping in mind always the difference between certain and probable inference. Furthermore I have stressed the significance of distinguishing between the problem of estimating the degree of rational belief in the assumptions underlying the prediction based on probability theory and the problem of predicting if we were certain that the assumptions were satisfied in a given case.

In a very general way, such a survey of the contribution of probability theory as a guide to action indicates the need for giving attention first to the problem of eliminating or segregating assignable causes of variability until we have arrived at a state where we may rationally justify the use of

the calculus of probabilities and distribution theory in predicting what may be expected to happen. The field of industrial science, and particularly that having to do with the production of finished goods to satisfy human wants, offers an exceptional opportunity for the development of important applications of probability theory. Here we set out to do a thing again and again within limits that are economical. A consideration of probability theory serves as a guide in establishing efficient ways of eliminating assignable causes of variability and of specifying economic standards and control methods. In doing this, use is made of many important mathematical contributions in the calculus of probability and statistical theory. Although there are several workers in this field in America, England, and Germany including such men as "Student", E. S. Pearson, Tippett, Plaut, Daeves and Von Mises on the other side of the water, it is my firm conviction that we have only just scratched the surface of what appears to be a veritable mine of important applications of probability theory in securing efficient human action in engineering and manufacturing.

W. A. SHEWHART'S COLLECTION