SCIENCE

VOL. LXVI DECEMBER 30, 1927 No. 1722

CONTENTS

SCIENCE: A Weekly Journal devoted to the Advancement of Science, edited by J. McKeen Cattell and published every Friday by

THE SCIENCE PRESS

New York City: Grand Central Terminal.

Lancaster, Pa. Garrison, N. Y.

Annual Subscription, $6.00. Single Copies, 15 Cts.

SCIENCE is the official organ of the American Association for the Advancement of Science. Information regarding membership in the Association may be secured from the office of the permanent secretary, in the Smithsonian Institution Building, Washington, D. C.

Entered as second-class matter July 18, 1923, at the Post Office at Lancaster, Pa., under the Act of March 8, 1879.

THE NOTION OF PROBABLE ERROR IN ELEMENTARY STATISTICS[1]

WHAT I have to say to-day is not addressed to professional mathematicians or statisticians. To mathematicians and statisticians all that I shall say is already entirely familiar. There are two other classes of readers, however, to whom I hope the discussion may be of service: (1) the rapidly increasing number of laymen who, without technical mathematical training, are constantly coming upon such terms as "probable error" in their general reading, and (2) the non-mathematical research worker who is constantly tempted to embellish his numerical results by adding an imposing array of "probable errors"—obtained, alas, too often by the simple process of substituting blindly in a formula. (A formula, of course, is an essential tool; what will concern us here, however, is the underlying significance of such a formula, and the necessary limitations surrounding the proper use of it.)

What are the principles that lie behind the common use of the term "probable error"? What does it really mean when we say, for example, that a quantity x has an estimated value of 3.6 with a "probable error" of 0.2 (written $x = 3.6 \pm 0.2$)?

The conventional reply to this question will occur to all of us—namely, that "the probable error is the error that is as likely as not to be exceeded." For example, if $x = 3.6 \pm 0.2$ the conventional understanding is that the "true value" of x is as likely to lie outside the limits 3.4 and 3.8 as it is to lie between those limits.

But this conventional reply does not go very far behind the scenes—we should like to have something more fundamental. Under what circumstances can we properly speak of errors as "equally likely" to occur? What are the fundamental considerations underlying the whole range of ideas which are suggested by the term "probable error"? I believe the best modern opinion is in favor of treating the so-called "probable error" from the point of view of empirical statistics, with as little reference as possible to the technical theory of probability; and I am convinced that much misunderstanding will be avoided if we can keep as

[1] Address of the retiring vice-president and chairman of Section A (Mathematics), American Association for the Advancement of Science, Nashville, Tennessee, December 29, 1927.

far away as possible from the older language of probability.

A. ERRORS OF MEASUREMENT

The earliest use of the term "probable error" which I can discover is in a paper by Bessel in 1815. Bessel, following some then recent methods of Gauss, was discussing a problem in the adjustment of measurements of an unknown quantity. Let us begin, therefore, with a brief outline of the problem of *adjustment of measurements.* This problem is conveniently treated under two headings, first, the "probable error of a single observation" and secondly, the "probable error of the mean."

I. *The "probable error of a single observation"*

Suppose we have before us a large number of measurements of an unknown quantity; suppose next that we take the arithmetic mean of these measurements; and suppose further that we compute the deviations of the given measurements from the mean. If the number of measurements is large many of them will coincide exactly with the mean value; and among those which differ from the mean, small deviations will occur more frequently than large ones.

If now we lay off the values of the given measurements along an axis of abscissas, and at each point of this axis erect an ordinate which shows the number of times that the corresponding measurement occurs, we shall have a *frequency diagram* or *distribution diagram* for the given set of measurements. The area of the diagram (or, rather, the area divided by the smallest recognized interval along the axis) will be equal to the total number of measurements in the set. The actual form of the diagram for a given set of measurements is a matter of experience. In a large number of cases, however, the distribution is found to conform to what is known as the *normal law of error,* represented by the familiar bell-shaped curve whose equation can be found in any book on statistics. If the measurements are closely consistent with each other, most of the deviations from the mean will be small and the distribution curve will be sharply peaked; if the measurements are less consistent—that is, more scattered—the curve, though of the same area, will be flatter.

The question at once arises: how shall we secure some estimate of the consistency of the given set of measurements? One method for doing this is as follows: we may divide the area of the distribution curve into four equal parts by ordinates erected at the points $x=-r$, $x=0$, $x=+r$, where x is measured from the mean; the value r will then have the property that just half of the deviations from the mean will lie between $-r$ and $+r$. This value r is called, after Galton, the *quartile deviation* of the given set of measurements, and may obviously be taken as an indication of the consistency of the measurements; the smaller the quartile deviation the more closely packed are the measurements about their mean.

By an unfortunate use of language, for which Bessel and Gauss are chiefly responsible, this quartile deviation is commonly known as the "probable error of a single observation," for the given set of measurements. This term "probable error" is here used in a highly technical sense and does not mean at all what it would appear to mean in ordinary language. It is best interpreted as merely an obscure synonym for the clearer, almost self-explanatory, term quartile deviation. The important thing to note is that the "probable error of a single observation," in spite of its name, is not a property of any single measurement, but a property of the whole set of measurements; it enables us to say, not that any single item is more accurate than another single item, but that one whole set of measurements is more consistent with itself than another whole set of measurements. The term is used chiefly in statements describing the precision of an instrument, or the precision of some measuring process. It is not often used as the $\pm$ 0.2 that one sees annexed to numerical values.

This, then, is the first common use of the term probable error; the so-called "probable error of a single observation" means merely the quartile deviation of the given set of measurements; it serves to indicate the self-consistency of the set of measurements, or the peakedness of the distribution diagram.

II. *"The probable error of the mean"*

The second common use of the term "probable error" is in the phrase "probable error of the mean." The conventional explanation of this phrase runs somewhat as follows: suppose we have a given set of n measurements, conforming to the normal law of distribution, and having a definite mean and a definite quartile deviation. Next, let us *pretend* that we have also a large number of similar sets of measurements of the same quantity, making k sets in all, each containing n measurements; and consider the k means belonging to these k sets. These means will constitute a sort of super-set of k values which will have its own distribution diagram, its own mean and its own quartile deviation. By a subtle application of the theory of probability, the quartile deviation of this super-set is proved to be equal to the quartile deviation of the original set divided by the square root of n; and this value is what is called the "probable error of the mean," for the original set.

This conventional explanation leaves much to be desired. What is the use of pretending that we have a

"super-set" composed of "a large number of sets of measurements similar to the given set" when we have in reality only one set to work with? And why should the quartile deviation of this hypothetical super-set be of any significance in the problem of measurement?

When one examines the actual use that is made of the so-called "probable error of the mean" one finds that it is almost always associated with *the problem of combining several sets of measurements,* with a proper "weight" attached to each set. In the practical solution of this problem there is no question of a hypothetical super-set of imaginary sets of measurements; all the sets of measurements with which we are concerned are actually given. Two illustrations will make the practical method clear.

First, suppose we have two normal sets of measurements of equal consistency, one containing ten measurements, the other twenty. In combining these two sets of measurements it is natural to give the second set twice as much weight as the first, since the number of measurements in the second set is twice as great as the number of measurements in the first. The combined mean or "weighted average" of the two sets will then be the mean of the first set plus twice the mean of the second set all divided by three. The justification of this process of computing the weighted average of two such sets lies in the fact that it gives exactly the same result as if we had taken all thirty measurements as a single set of measurements and found the mean of this set in the ordinary way.

Secondly, suppose we have two sets of measurements containing the same number of items, but having unequal consistency. Suppose for example that the quartile deviation of the first set is $r_1 = 3$, and the quartile deviation of the second set is $r_2 = 4$. Before combining these two sets of measurements, we must first reduce them, so to speak, to a common denominator. To accomplish this we may make a photographic enlargement of both diagrams, until the quartile deviation of each is equal to the same number, in this case 12. This step is justified by the natural assumption that two distribution diagrams which are similar—that is, one merely an enlargement of the other—are of equal weights. Here, in the case of the first diagram we multiply the linear dimensions by 4, and therefore the area by 16; and in the case of the second diagram, we multiply the linear dimensions by 3, and therefore the area by 9; the position of the mean in each case being unchanged. The quartile deviation of each diagram is now equal to 12, so that the two revised sets of measurements are of equal consistency and can be combined by the method just described. Remembering that the area of a distribution diagram is proportional to the number of measurements, the first set must be given a weight of 16, and the second set a weight of 9. The weighted mean will therefore be equal to 16 times the first mean, plus 9 times the second, all divided by 25.

It is easy to show that the same result would have been obtained if we had multiplied the first mean by a weight equal to $(1/r_1)^2$, and the second mean by a weight equal to $(1/r_2)^2$, and divided by the sum of these weights.

The extension of this process to the combination of the two cases: namely, to the case of several sets of measurements which differ not only in consistency but also in the number of measurements in each set, presents no difficulty. We are thus led at once to the following general rule for combining any number of sets of observations which are normally distributed: *the weight to be attached to each set is directly proportional to the number of measurements in that set and inversely proportional to the square of the quartile deviation of the set.*

I hope that this brief sketch of the practical method of combining sets of measurements will make it clear that the whole subject can be presented without reference to anything except what is immediately given by the actual measurements; it is not necessary to bring into the discussion any hypothetical super-set of imaginary sets of measurements or to make any use of the technical theory of probability. The ± 0.2 placed after the numerical statement of a mean value is commonly called the "probable error of the mean." This is a quantity obtained by dividing the quartile deviation of the given set of measurements by the square root of the number of measurements in the set; it is best regarded as merely a conventional way of indicating one step in the computation of the weight which should be attached to the given value when this value is to be combined with other values of a similar nature. It is not necessary to think of it as something mysteriously connected with the theory of probability.

It is interesting to note in passing that there is another measure of the consistency of a set of measurements, which is coming more and more into use. This is the *standard deviation,* or mean square error, introduced (under the name "mean error") by Gauss in 1821. The standard deviation is the square root of the mean of the squares of all the deviations from the mean; in the case of the normal curve it proves to be simply the abscissa of the point of inflection (measured from the mean). For this curve, as is well known, the standard deviation, σ, and the quartile deviation, r, are connected by the relation $r = 0.6745\sigma$, and the ordinary method of computing the quartile deviation is first to compute the standard deviation directly from the given measurements, and then to multiply by 0.6745. The quartile deviation (or "prob-

able error") is thus about two thirds as large as the standard deviation (or "mean square error"), in the case of the normal curve.

A pretty quarrel has arisen as to which of these two quantities is the handier one to use as an indication of the consistency of a set of measurements. Gauss himself began in 1816 with the exclusive use of the probable error. In 1821 he uses the mean square error and the probable error side by side. By 1828 he begins to speak of the probable error as the "so-called" probable error; and a few years later he is quoted as saying: "the so-called probable error, I, for my part would like to see altogether banished." In 1889, Francis Galton, the grandfather of the British school of statistics, condemns the term probable error in vigorous language. "It is astonishing," he writes, "that mathematicians, who are the most precise and perspicacious of men, have not long since revolted against this cumbrous, slip-shod, and misleading phrase." Many recent writers like R. A. Fisher agree that the fact that the use of the probable error is common "is its only recommendation." On the other hand, Professor Mansfield Merriman (1884) regards the probable error as the most natural unit of comparison and insists that it alone should be used and the mean square error be discarded. At the present time both the probable error (or quartile deviation) and the mean square error (or standard deviation) are so thoroughly established in the literature that neither of them is likely to be given up.

Let us now leave the subject of errors of measurement and pass on to another use of the term "probable error," namely, its use in connection with the subject of random sampling—a subject which is coming more and more to occupy the central position in the whole modern theory of statistics.

B. Random Sampling

In the problem of errors of measurement, the final result desired is the value of a single unknown quantity, and the distribution diagrams of sets of measurements of the unknown are merely means to an end. In the problem of random sampling, however, the final result desired is the distribution diagram itself.

For example, a shoe manufacturer wishes to know what demand he may expect for various sizes of shoes. He wishes to know, for example, what proportion of the population wears a number eight shoe. What he needs is a distribution diagram of the foot-sizes of the whole population. This distribution diagram will exhibit, of course, a certain mean value; but this mean value is not now the interesting thing; and the deviations from the mean, instead of being errors to be avoided, are now important for their own sakes. The distribution diagram itself is the thing that is wanted. Now the distributions that occur in practice are by no means always of the normal form; a frequency diagram may often be "skewed" in one direction or the other; it may be more sharply or less sharply peaked than the normal curve of the same area and same quartile deviation; or it may even be of a U-shaped form, with the large deviations from the mean more frequent than the small ones.

In order to describe a distribution diagram concisely we may state the values of four parameters, two of which we have already mentioned: (1) the mean; (2) the standard deviation, that is, the square root of $1/n^{th}$ of the sum of the squares of all the deviations from the mean; (3) the third moment, that is $1/n^{th}$ of the sum of the cubes of the deviations from the mean; and (4) the fourth moment, or $1/n^{th}$ of the sum of the fourth powers of the deviations from the mean.

The standard deviation, as we have seen, gives a measure of "dispersion" or "scatter." (If the distribution happens to be symmetrical, either the standard deviation or the quartile deviation may be used as a measure of dispersion; but in the general case the standard deviation alone is available.) The third moment leads to a measure of "skewness." The fourth moment leads to a measure of what Pearson calls "kurtosis." Any given distribution diagram is sufficiently characterized for most purposes by giving the values of these four parameters; the mean, the standard deviation, the third moment, and the fourth moment.

Let us suppose then that our shoe manufacturer desires to study the distribution of foot-sizes in the whole population of a hundred million people. He obviously will find it impracticable to measure the whole population, so that he can not obtain the parameters of the distribution directly. He therefore takes a sample of a moderate number, n, of people, chosen, as we say, at random, and determines the parameters of this sample. The question is, what conclusion can be drawn about the mean, standard deviation, etc., of the total population from a knowledge of the mean, standard deviation, etc., of a single sample?

This question is being actively discussed at the present time, and all that I can do here is to indicate briefly the nature of the answer that may be hoped for. Suppose, for example, that the parameter in which we are interested is the mean. Consider the totality of all possible samples of n which can be drawn from the population in question. The number of such samples will of course be enormously large, but can be readily computed by the theory of permutations and combinations. Each sample of n will have its own mean; and the set composed of the means

of all the samples will be distributed in a perfectly definite way, depending on the nature of the total population.

Thus, it has been proved that the mean of the set of means will coincide with the mean, m, of the total population; the standard deviation of the set of means will be equal to $\sigma/\sqrt{n}$, where σ is the standard deviation of the total population; and the other moments of the set of means can be computed in terms of the corresponding moments of the total population. That is, if we assume any hypothetical values for the parameters of the total population, we can theoretically compute the parameters of the distribution of the means. Then by a subtle analysis, we can make a comparison between the distribution of the means and the observed properties of the given sample, and thus construct a test of the validity of our assumed values.

The result of such a test is commonly recorded in this form: the required mean, m, is equal to the observed mean, a, of the measured sample, plus or minus a "probable error" r. This indicates merely that if we had the totality of the means of all possible samples of n before us, 50 per cent. of these means would lie between $a+r$ and $a-r$. This use of the term "probable error" is unsatisfactory, however, since the distributions involved in the analysis are usually not symmetrical; the "standard deviation" is the more useful concept. Moreover, there is no special sanctity attached to the arbitrary choice of "50 per cent"; other ranges are often needed.

Moreover, the formulas commonly given for computing the probable error of the various parameters are only approximations which are not valid unless the original distribution is normal, and the size of the sample is large. The serious study of this whole question, for the general case of skew distributions and small samples is a product of the last two decades—one might almost say of the last two years. Some of the names associated with this study are Karl Pearson, R. A. Fisher, Tchouproff, and especially a learned British scholar who conceals his identity behind the modest pen-name of "Student." Exciting new developments are constantly appearing in *Biometrika* and similar journals; the most modern tools that mathematics can supply as, for example, the theory of integral equations, are called into play; and the very latest results are immediately put to use by practical statisticians of the Bell Telephone System and other great industrial concerns. The work is by no means completed, and even the exact nature of the answer that may be hoped for is not yet entirely clear.[2]

A splendid field for research is opening up, the fruits of which are sure to be not only of the greatest theoretical interest but also of the highest practical utility.

EDWARD V. HUNTINGTON
HARVARD UNIVERSITY

THE GENERAL RADIATION[1]

THE impacts of electrons against atoms produce two different kinds of radiation, (a) the line spectra and (b) the general radiation, sometimes called the continuous, or white, spectrum. The general radiation usually carries a far greater amount of energy than the line spectra—hot body radiation, for instance. This is true of the X-ray region of the spectrum as well as of other regions. Although X-ray spectrum lines are often strongly marked and sharply defined, the general radiation contains more energy than the lines, for it covers a much greater range of wave-lengths. In the evolution of recent thought, however, less attention has been paid to the general radiation than to the line spectra, partly because the line spectra have important bearings on our ideas as to atomic energy levels. In this address, I wish to present to you the more important characteristics of the general radiation, as they have been discovered by about twenty men, carrying on researches in different parts of the world. Time will not permit a detailed account of the subject. These details may be found in the text-books, which contain numerous references to the original articles published by the investigators.

On account of the fact that homogeneous beams of high-speed electrons can be produced, accurately controlled and measured, and because each electron has a relatively large amount of energy, the X-ray region of the spectrum provides us with a better field for investigating general radiation than do other regions.

The curve representing the distribution of energy in the general X-radiation spectrum as a function of the wave-length resembles that for the spectrum of black body radiation. There is one important difference between the two, however, namely, the general radiation spectrum has a sharply defined short wave-length limit. The quantum theory explains this limit quantitatively and qualitatively; for the electrons striking the atoms of the X-ray tube's target can not have kinetic energies greater than the product of the electron's charge into the difference of potential through which it has fallen (namely, Ve). Therefore, the hν value of the quanta of radiation produced can not be greater than Ve. Strictly speaking, if we apply the laws of the conservation of energy and momentum to the impact of an electron against an atom, we find that the value of

[2] For further information the reader is referred to H. L. Rietz's Monograph on Mathematical Statistics (Open Court Publishing Company, 1927).

[1] Address of the vice-president and chairman of Section B (Physics), American Association for the Advancement of Science, Nashville, Tennessee, December, 1928.