STATISTICAL METHODS AND CONTROL IN BACTERIOLOGY[1]

CHURCHILL EISENHART AND PERRY W. WILSON

College of Agriculture, University of Wisconsin, Madison, Wis.

CONTENTS

It has been alleged that certain people use statistics as a drunk does a lamp-post—more for support than illumination. Such a criticism of the misuse of statistical methods is unfortunately too often justified and probably is the basis of the somewhat passive but nevertheless widespread opposition encountered when these methods are first introduced into any biological discipline. That this first phase of more or less passive resistance is about completed in bacteriology is evidenced by the ever-increasing references to statistical treatment of the data in its journals. Not so very long ago statistics touched bacteriology primarily in just one field—bacterial enumeration—but current literature provides statistical analysis of data from diverse experiments. As examples the following are cited: virulence of streptococci for mice (10); relation between growth of bacteria and the heat stability of their enzymes (25); deterioration of cellulose fibers by fungi (43); reliability of soil counts (54, 55); disinfection of trout eggs (39); test of fungicides on mold spores (111); probability of isolating a pure culture in a Petri dish (68). These applications, with many others which will be described in greater detail in the text, suggest

[1] Much of the experimental work by Wilson and his associates selected for illustrative purposes in this paper was aided by a grant from the Rockefeller Foundation. This material was used primarily because of convenience,—the complete original data necessary for the detailed calculations were readily available.

that the time is appropriate for a review of the use of statistical tests and control in bacteriological work.

This paper is somewhat different from that usually published in the BACTERIOLOGICAL REVIEWS in that the literature serves primarily as a source of more or less familiar examples useful for illustrating the statistical principles discussed. Thus, the intrinsic significance or insignificance of the data selected for purposes of illustration is beside the point. The extension to data which differ in content but not in the principles involved should not be too difficult and "is an exercise left for the reader."

Part I deals with the distributions of variables; application of the knowledge of different types of distribution is illustrated by examples of statistical control of laboratory procedures. The important point in this section is an appreciation of the principles rather than mastery of the details of the calculations. In contrast, the arithmetic of the statistical tests of significance discussed in Part II is given in some detail since this facilitates an understanding of their application to actual data. Although, at first sight, some of the problems solved in this part may appear rather complicated and the calculations formidable, close examination will demonstrate that they involve easily-followed procedures. In passing, it should be noted that the titles of these major sections refer to the primary emphasis in that portion of the paper, but the subject matter is not entirely restricted to that implied by the section heading. Thus, some tests of significance are necessarily used in Part I in order to illustrate certain aspects of the distributions of variables. Likewise, the various statistics discussed in Part II have other functions, *e.g.*, for description of data, as important as is their use for the statistical tests. In both Parts I and II it has been necessary to introduce certain technical terms which are precisely defined only in mathematical terminology. To maintain continuity of both style and content in the text such terms are used without comment and are defined and discussed in the *Appendix*.

Finally, it is emphasized that this paper is not intended as a course in statistics. The mathematical formulation has been kept to a reasonable minimum and emphasis placed on mathematical assumptions and principles of statistical theory as they relate to statistical interpretation of experimental results. Without special knowledge, it is often difficult to say whether these assumptions obtain in a given body of experimental data. Blind application of the formulae may therefore lead to error. This pitfall is best avoided by seeking the aid of a qualified statistician. Except for a few simple applications, such as determining *means* or *slopes* of lines, the biologist should ordinarily no more attempt the statistical analysis of his complicated experimental findings without consulting the trained specialist than he should try to analyze his cultures for an isotope without the advice of a physicist. It is natural for the biologist to ask: "Then why should I know anything about the subject?" Primarily, to recognize problems in his research which might benefit through statistical interpretation or control. Of almost equal importance, to be sufficiently informed so that he can present his material intelligently and concisely to the statistical consultant.

Statistical Control. In the foregoing the term, *statistical control*, has been used frequently and a more detailed exposition of what is implied in this phrase may be of value. The importance of statistical control of laboratory and field techniques needs greater emphasis in biological literature. Laboratories using statistics have employed almost exclusively *tests of significance* and have neglected the opportunity afforded for day-to-day check on the reliability of a routine analytical method. An experimental procedure is said to be in a state of statistical control when the observations to which it gives rise under what are assumed to be 'essentially the same conditions' fluctuate in a random manner and are free from trends and non-random shifts in magnitude. Unless a sampling procedure—and in a sense all processes leading to observations *are* sampling procedures—is in a state of statistical control, it is not possible to make valid inferences about the 'population' which the observations are supposed to represent. In the important paper by Fisher, Thornton and Mackenzie (34) this is stated: "Any significant departure from the theoretical distribution is a sign that the mean may be wholly unreliable."

In science, industry, and commerce where decisions must be made on the basis of results of some series of measurements, the reliability of the methods must be known. To know that the methods provide good checks or even that two operators obtain similar results is not enough. It is generally recognized that in arguing from the particular to the general the wrong decision will occasionally be recommended by the observations obtained, and we should know how frequently these errors are apt to occur. If too often (in industry, the economic consequences furnish a valuable measure of how frequent is 'too often'), then the procedure must be altered so as to reduce the expectancy of false decisions. When a procedure is statistically controlled, the expectancy of a false decision is a minimum. The Western Electric Company (equipment manufacturer for the Bell Telephone System) has led the way in applying statistical control to manufacturing processes, and other large industrial organizations (*e.g.*, General Electric Company, United States Steel Corporation) have found it profitable to follow suit.[2] The statistical staff of the Rothamsted Experimental Station in England has pioneered in showing biologists how to check their sampling techniques and has stressed the importance of doing so.

More evidence regarding statistical control might well be included when publishing research, since its absence may modify conclusions profoundly. When a series of observations exhibits properties widely divergent from those characteristic of random samples of a hypothetical population that is strongly suggested by intuition, the first inference is that the sampling technique (which includes laboratory procedures, *etc.*) is not statistically controlled and that greater pains must be taken, *e.g.*, in mixing solutions before withdrawing samples. Experience in biology and in industry over a period of nearly two

[2] See Shewhart (81, 82) and Simon (83). Dr. Shewhart is the father of statistical control techniques in industry. The results achieved by Colonel Simon through the use of statistical control at the Picattiny Arsenal, Aberdeen Proving Grounds, have won him wide recognition and have been a major factor in convincing the Ordnance Department of the value of quality control.

decades shows that the foregoing is frequently the correct conclusion, and that a state of statistical control can be attained by persistent efforts directed toward improvement in technique. In some instances, however, the divergence is due to the choice of a hypothetical population which is unsuited to the phenomenon in question. For example, in studying the occurrence of larvae on a field, the distributions observed differ widely from those anticipated under the hypothesis that larvae are distributed on the field *independently* and *at random*. Biological considerations, *e.g.*, the fact that the eggs are laid in 'masses,' suggest that the presence of a larva in a given neighborhood increases the probability of there being others nearby. Accordingly, hypothetical 'contagious' distributions have been devised (69) with which the experimental facts seem to be in full agreement. In bacteriology such difficulties can usually be avoided by shaking suspensions well before taking counts, but as discussed in the following sections, other details of the technique may interfere with the realization of the hypothetical population.

PART I. DISTRIBUTIONS OF VARIABLES

A problem common to many branches of science is to determine whether values taken under one set of conditions differ significantly, in the statistical sense, from other values taken under other circumstances. The problem arises because individual values for any measurable quantity are rarely identical but show degrees of variation. Among the numerous causes may be cited: (a) errors in measurements because of lack of precision in the measuring instruments or ineptitude of the measurer; (b) variations among the individuals comprising the population—all men are not created equal. To decide if one set of data differs significantly from a second set, the statistician endeavors to define the characteristics of the populations from which the two sets were obtained and then to determine whether the two populations are identical in one or more respects. Our first problem, therefore, is to consider various types of populations and the methods by which the measurements of a variable can be used to calculate the significant parameters of each type.

THE BINOMIAL DISTRIBUTION

If the probability of an event occurring in *any* single trial is p, then the probabilities of it occurring exactly $0, 1, \cdots, x, \cdots, n$ times in n *independent* trials are given by the successive terms of the binomial expansion of $(q + p)^n$, where $q = 1 - p$ is the probability the event will not occur in any single trial.[3] The terms so generated form the *binomial distribution*, one of the most important hypothetical populations in biological research. It is sometimes called the *point binomial* since a variable so distributed can assume only integer values from 0 to n, and in consequence the probabilities are concentrated at these points. It is also referred to as the *Bernoulli Distribution* after its discoverer, James Bernoulli (1654–1705).

[3] For details see any text book on college algebra; particular the topics *Binomial theorem* and *Probability*.

A count, x, distributed in random samples in accordance with the above point binomial has a mean np and a standard deviation $\sqrt{npq}$, so that the observed proportion, $p' = x/n$, has a mean of p and a standard deviation $\sqrt{pq/n}$. When n is large, x is approximately normally distributed about its mean with the indicated standard deviation (23), as is p' also.[4] It is often convenient to employ $\theta = \arcsin \sqrt{p'}$ in statistical analyses instead of p', since for large values of n, the variance of θ is independent of the value of p, which is generally unknown. Tables are available (5, 6, 35) to facilitate this transformation. The point binomial possesses a reproductive property: the sum of N *independent* counts $x_1, x_2, \cdots, x_N$ based on samples of sizes $n_1, n_2, \cdots, n_N$ from the *same* population, *i.e.*, p the same in each case, is binomially distributed with $n = n_1 + n_2 + n_j \cdots + n_N$. In consequence, a composite sample obtained by combining several independent samples may be regarded as a single large sample. The observed proportion in the composite sample, $\hat{p} = (x_1 + x_2 + \cdots + x_N)/n$, provides an unbiased estimate of p which contains all the information about p available in the data. For purposes of checking on *statistical control* it is advisable, however, to keep a record of the size (n_j) and count (x_j) for each of the respective samples.

Applications of the binomial distribution are numerous in genetics where Mendelian theory specifies the value of p. It is possible, however, to test whether a series of counts has properties characteristic of samples from a binomial distribution without knowing the value of p. Agreement with the binomial distribution is taken as evidence of the *independence* of whatever operations constitute 'trials' and of the *constancy* of p from trial to trial, which properties jointly comprise one form of *statistical control* often known as *simple sampling*.[5]

In sampling biological populations it is often desirable to test for agreement with the binomial to ascertain whether the sampling technique employed is statistically controlled. Likewise, when *random* samples are taken from each of several parts of a large body of material, or at different times from an ever-changing population, a test of whether the several samples may be regarded as samples from a *single* binomial, constitutes a test of whether p, the proportion possessing the characteristic under investigation, is the same throughout. If not, the population sampled is heterogeneous in respect to that characteristic, and heterogeneous material cannot, for purposes of inference, be treated statistically as though it comprised a single population.

Example: Table 1 (unpublished results of L. C. Ferguson and M. R. Irwin) gives data on the relative frequency of monocytes in the blood cells of a certain cow. Samples of 100 blood cells were counted at weekly intervals over a period of approximately two years; as is shown in the second column of the table, in the 113 samples, 19 contained exactly 4 monocytes, 2 contained exactly 12 monocytes, *etc.* Of the 11,300 cells counted 673 were monocytes, therefore $\hat{p} = 673/11{,}300 = 0.059558$. The expected frequencies, given by the successive

[4] See the discussion of the normal distribution, p. 96 ff.

[5] For an excellent discussion of *simple sampling* and of the various types of departures from it, see Yule (123).

terms of

$$113\ (0.940442 + 0.059558)^{100},$$

were computed with the aid of seven-place logarithms. A table of logarithms of $n!$ greatly facilitates the calculation, *e.g.*, table 49 of ref. (77). These calculated values are given in the third column of table 1; comparison with the observed frequency distribution indicates that the latter is more widely spread about the mean (5.9558) than would be expected in binomial sampling.[6] Various explanations of this apparent discrepancy are suggested: (a) the selection of the samples from the bloodstream was non-random; (b) the bloodstream was not homogeneous in the proportion of monocytes present, *i.e.*, the true proportion of monocytes varied from week to week; (c) the monocytes tended to occur in small clusters instead of being distributed at random within the bloodstream.

TABLE 1

Frequency of monocytes in blood of a cow

NUMBER OF MONOCYTES PER 100 BLOOD CELLS (X)	OBSERVED FREQUENCY F_X		EXPECTED FREQUENCY f'_X	
0	0		0.2	
1	3	8	1.5	6.5
2	5		4.8	
3	13		10.0	
4	19		15.3	
5	13		18.7	
6	15		18.7	
7	12		15.9	
8	10		11.7	
9	11		7.6	
10	7		4.4	
11	3	12	2.3	8.6
12	2		1.1	
Over 12	0		0.8	
	113		113.0	

It is not possible to infer from the data as arranged in table 1 which of these explanations is the proper one. From an examination of the original data sheets and from other evidence, however, it appears that (b) is the correct explanation.

THE POISSON SERIES

A distribution which is frequently of value in the description of biological material is the *Poisson Series*, the so-called law of small probabilities. Specifically, it defines phenomena whose occurrence is governed by the following

[6] A more precise method for comparing the two distributions is described in the section on *Testing for Agreement between Observed and Expected Frequencies.*

conditions: (a) the probability of occurrence, p, is very small, *i.e.*, order of 0.01 or less; (b) the number of individuals exposed to the 'risk' is extremely large so that the mean number of successes, or occurrences, np, is some small number; (c) the frequency of occurrence is represented by *small whole numbers*. Poisson showed that under these conditions, the probability of obtaining various frequencies is given by the series:

$$[1] \qquad e^{-m}, e^{-m}\frac{m}{1!}, e^{-m}\frac{m^2}{2!}, e^{-m}\frac{m^3}{3!}, \cdots \frac{e^{-m}m^x}{x!}, \cdots,$$

probability of 0, 1, 2, 3, $\cdots$ x, $\cdots$

in which m is the average number of occurrences per sample.

A few properties of this distribution should be carefully noted:

(i) Its mean is m and its standard deviation is $\sqrt{m}$; hence an estimate of m, provides an estimate of its own error.

(ii) If in N independent samples from the same population an event occurs $x_1, x_2, \cdots, x_N$ times respectively, then the observed mean $\bar{x} = (x_1 + x_2 + \cdots + x_N)/N$, provides an unbiased estimate of m, the expected frequency per sample, and furthermore, $\bar{x}$ contains all the information about m available in the data (30a, 34).

(iii) $\bar{x}$ is approximately normally distributed about m with standard deviation, $\sqrt{m/N}$, for any m if N is sufficiently large, and for any N (*e.g.*, for $N = 1$ so that $\bar{x} = x_1$) if m is sufficiently large. It is sometimes convenient to utilize the fact that the variable $y = \sqrt{\bar{x}}$ is approximately normally distributed about $\sqrt{m}$ with standard deviation $\sqrt{1/4N}$, which is independent of m, under these conditions (1, 2).

(iv) If T is the sum of N components which are independently distributed in a Poisson series of parameters $m_1, m_2, \cdots, m_N$ respectively, then T itself is distributed in a Poisson series of parameter $m = m_1 + m_2 + \cdots + m_N$, so that T may be regarded as the frequency in a single sample from a Poisson series whose mean is estimated as T with estimated standard error $\sqrt{T}$.

(v) For small values of m the stability of the occurrence of events is very high—*e.g.*, when $m = 1$, the probability of no occurrence is $1/e = 0.368-$, which is also the probability of a single occurrence; the probability of 2 occurrences is $1/2e = 0.184-$, and the probability of more than 2 occurrences is therefore $1 - (5/2e) = 0.080$. There is considerable skewness with the 'tail' to the right; as m increases, this skewness diminishes although symmetry is attained only in the limit as $m \to \infty$.

It is of interest that the first experimental tests of the series were concerned with biological events. Bortkiewicz (8) showed that the number of men killed from the kicks of horses in each of 14 Prussian army corps for 20 successive years followed the Poisson law of small numbers. In 1907 "Student" (89), the famous chemist-statistician at a Dublin brewery, demonstrated that under somewhat idealized laboratory conditions the distribution of yeast cells on the squares of a hemocytometer conformed to a Poisson distribution; he also independently derived the law from considerations of how the yeast cells should distribute themselves in the squares of the counting chamber. Greenwood and White (44) investigated from the point of view of the Poisson distribution the ingestion of tubercle bacilli by phagocytes. Bortkiewicz's treatment contained the elements of statistical control of experiment, an application for which the series has been most useful. He discarded the records of 4 corps in which the

deaths were considerably higher than the others, allegedly, because the men in charge mistreated the animals so that they were more vicious than the horses in the other corps.

Following these publications an increasing number of diverse phenomena have been compared with the distributions predicted by Poisson series, including: emission of alpha particles from polonium; number of noxious weed seeds in a sample of timothy seed; number of umbrellas left on buses (statistical control, eliminate rainy days); death notices for men over 85 in the obituary column of London *Times*; wrong number connections in a telephone exchange, number of fires in New York City during a year (statistical control, eliminate July 4th and Election day); defects in a manufactured article; calls for a reference book in a University library. (See Thorndike (99) for an interesting discussion of many of these.)

Statistical control of bacterial counts by chamber method. Consideration of the conditions under which a count of yeast or bacteria is made in the various types of counting chambers leads to the conclusion that the distribution of organisms per square should follow a Poisson series since: (a) the probability that a given organism will be found in a given square is extremely small, but very large numbers of organisms are exposed to this small 'risk'; (b) the count per square will be some small whole number. "Student's" experiments with yeast were more for the purpose of verifying the law than for testing the methods of counting, but Wilson and Kullmann (114) definitely used the distribution for statistical control of a laboratory technique. They estimated numbers of the root nodule bacteria (*Rhizobium trifolii*) in a Petroff-Hausser counting chamber; because this organism produces gum, it clumps readily which frequently interferes with the reliability of results. Various refinements in technique were developed to overcome clumping, and the method as finally adopted was tested by counting the distribution of cells in the 400 squares of the chamber.

Figure 1 illustrates the results of four trials using the method of Thorndike (99) for testing agreement with the proper series. She has shown that if the relative frequency of obtaining at least c occurrences in data from a Poisson series is plotted on a special graph paper,[7] the points should follow a straight line drawn from the number at the base which corresponds to the mean number of occurrences (m). When the value of m is unknown—the usual case—the observed mean, $\bar{x}$ = (total number of organisms counted)/(number of squares examined) = T/N, may be taken as its estimate. The fit of the points to the theoretical lines in figure 1 is satisfactory in all four cases, especially in the center where the data are more reliable.

If the plotted points show a negative slope (*i.e.*, the points are to the left of the vertical in the upper portion and to the right of the vertical in the lower portion of the graph) the explanation of the non-conformance with Poisson

[7] Arithmetic probability paper. A scale proportioned to the normal probability curve is used for the ordinate, a linear scale for the abscissa. Logarithmic probability paper, with the abscissa scale in logarithmic units, permits the simultaneous portrayal of series with widely differing values of m.

sampling is often found in some restraint on large frequencies. Thorndike gives data on the number of calls in five-minute intervals from a pair of pay telephones which exhibit such a departure from Poisson expectations, "because of the fact that the number of calls which could possibly be made in five minutes from a group of two telephones is certainly finite and probably rather small." She gives also a sample of Perrin's data on particles in Brownian movement which show a similar departure and advances as an explanation that "it is

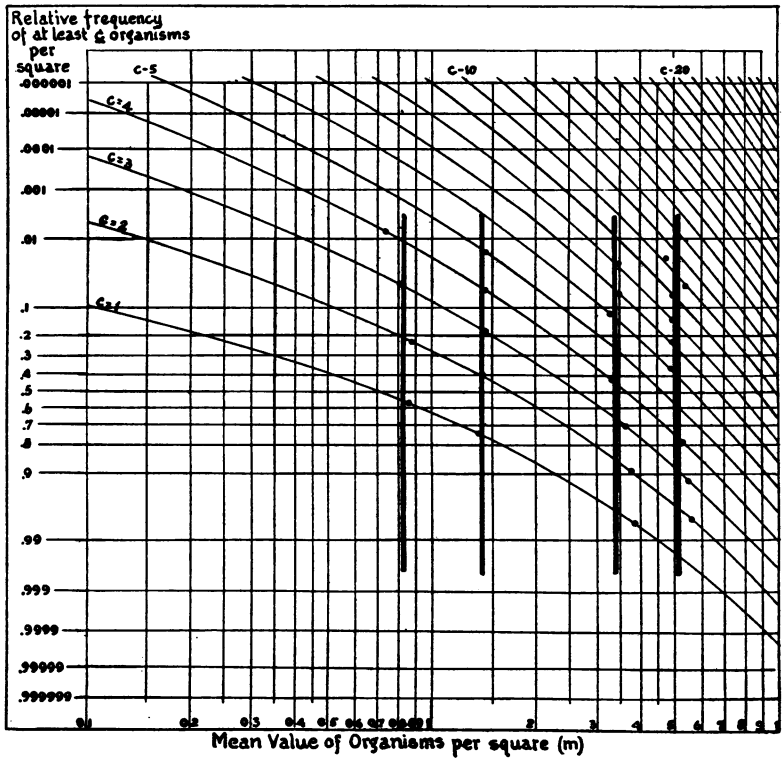

FIG. 1. COMPARISON OF OBSERVED DISTRIBUTIONS OF RHIZOBIUM TRIFOLII WITH THEORETICAL GIVEN BY POISSON'S EXPONENTIAL SUMMATION

The chart illustrates the use of a probability paper for testing whether data follow Poisson's law. A straight line is drawn from the point on the abscissa which corresponds to the mean number of bacteria per square; the experimental points represent the relative number of squares showing at least 1, 2, 3 . . . organisms. From Wilson and Kullman (114).

difficult to judge by the eye the number of particles visible simultaneously if that number is more than three or four." In bacterial counts such a departure from expectation on Poisson theory might arise from a tendency to underestimate the number of bacteria in crowded squares, or from a real restraint on large frequencies occasioned by competition among organisms. In either case greater dilution is a remedy. If the plotted points depart from the vertical with a positive slope, clumping or heterogeneity of material sampled (*i.e.*, m not constant throughout) are generally the explanations, although such a

departure could arise from a tendency to overestimate the numbers of bacteria in crowded squares.

Like that of most graphical methods of analysis, the principal advantage of this method of testing conformance to Poisson sampling is its rapidity, and its chief defect is its failure to provide an objective criterion for judging whether the discrepancies observed are meaningful or merely fortuitous. Nevertheless, with experience it can become a valuable test of experimental technique, and, when used in conjunction with a method giving a probability measure of the discrepancies, it provides a convenient portrayal of the diagnosis.

TABLE 2

Comparison of the theoretical distribution with that observed when counting Rhizobium trifolii in Petroff-Hausser counter

Mean = 2.50

NUMBER PER SQUARE	THEORETICAL f_t	OBSERVED f_o	$f_o - f_t$	$\frac{(f_o - f_t)^2}{f_t}$
0	32.83	34	+1.17	0.04
1	82.08	68	−14.08	2.41
2	102.61	112	+9.39	0.86
3	85.51	94	+8.49	0.84
4	53.44	55	+1.56	0.05
5	26.72	21	−5.72	1.22
6	11.13	12	+0.87	0.07
>6	5.67	4	−1.67	0.49
Totals.....	400	400		

$X^2 = 5.98$ $P = 0.43$ D.F. = 6

A more exact but slower method for testing the distribution is to compare the observed frequencies with the theoretical values obtained from the terms of the expansion

$$[2] \qquad Ne^{-m}\left(1 + \frac{m}{1} + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots + \frac{m^x}{x!} \cdots\right)$$

in which m is the true mean number of organisms per square, and N is the number of squares examined. When m is unknown, the observed mean number per square, $\bar{x}$, provides the appropriate estimate of m. Tables are available, which facilitate the determination of the theoretical frequencies, *e.g.*, those of Soper (86), which provide the values of $e^{-m}m^x/x!$ to six decimals for $m = 0.1$ to $m = 15.0$. These tables have been reprinted by Pearson (Table 51, ref. 77).

In the example given in table 2, a total of 1000 organisms were counted in the 400 squares, hence $\bar{x}$ equals 2.50. A comparison of the observed frequencies with those predicted by equation 2 ($m = 2.50$) is afforded by the first three columns of table 2, and indicates a reasonably close agreement; an exact evaluation of the agreement will be presented later in connection with the chi-square test of *goodness-of-fit*. The results of this and other similar trials indicate that

counts made with a Petroff-Hausser chamber under laboratory conditions follow the theoretical distribution when sufficient care is taken to break up clumps by thorough mixing of the suspension, thus confirming the reliability of the chamber count method when carefully executed.

How many observations should be taken? Property (iii) states that the mean number of organisms per square, $\bar{x}$, obtained from an examination of N squares has a standard deviation of $\sqrt{m/N}$, where m is the true mean number of organisms per square. The practical significance of this property is better appreciated when it is noted that the standard deviation of $\bar{x}$ is $100/\sqrt{Nm}$ per cent of the true mean, m, and is readily estimated as $100/\sqrt{T}$, from property (iv), T being the total number of organisms counted. Thus, for the data of table 2, $T = 1000$ gives an estimated standard deviation of 3.16 per cent for the mean (2.50) there shown. It follows, from property (iii), that the probability is approximately 0.95 that the mean observed here is precise to within $1.96 \times 3.16\% = 6.0$ per cent, that is, the probability is approximately 0.95 that the interval 2.50 ± 0.15 includes the true mean number of organisms per square for the dilution here employed.[8] Alternatively, the precision wanted can be decided beforehand and sufficient squares examined to provide that degree of precision. Thus, if it is desired to have a probability of 0.95 that the observed mean will be within 10 per cent of the true mean, the total number of organisms counted will have to be at least 400 since $1.96 \times 100/\sqrt{400} = 10$. As it is the duty of every scientist to make the right kind of observations, it is also his duty to make a sufficient number. In the words of Shewhart (82)

"The applied scientist in order to be 'successful' cannot afford to make too many mistakes even though they be small, and in no case can he afford to make a mistake that is large enough to cause serious trouble. He does not consider his job simply that of doing the best he can with the available data; it is his job to *get enough data before making his estimates.*"

A modified method of counting. Counting of the 400 odd organisms necessary to give reasonable assurance that the observed mean number per square will be precise to within 10 per cent can be accomplished in practice by either: (a) employing a dilute suspension and examining a large number of squares, or (b) employing a dense suspension and examining only a few squares. Of these two alternatives the former is preferable for at least two reasons. First, with low cell concentration, a true Poisson distribution of the organisms is more likely to be realized—with heavy suspensions clumping, competition between the organisms, *etc.*, frequently distort their distribution, in consequence of which no confidence can be placed in the observed mean number of organisms per square as an estimate of the true concentration. Second, with high concentrations, mistakes in counting the organisms arise from difficulties in discerning the individuals and from mistaken estimates of their number.

To reduce mistakes in counting, Tippett (101) has proposed a modified method which may prove to be of considerable practical value in bacteriological work.

[8] See discussion of *confidence intervals* in Appendix.

It consists of recording as data merely the numbers of squares containing, 0, 1, 2, $\cdots$, t, and 'more than t' organisms, where t is some small number, say 3 or less. When $t = 0$, so that N_0 and N, the numbers of squares containing no organisms and the total number of squares examined, respectively, constitute the 'data', the maximum likelihood[9] estimate of m and its standard error are

$$\hat{m} = 2.303 \log (N/N_0) \text{ and } \sigma_{\hat{m}} = \sqrt{(e^m - 1)/N}.$$

Thus, for the data of table 2, $N_0 = 34$ and $N = 400$, giving $\hat{m} = 2.46$ with an estimated standard error of 6.65 per cent in contrast to the estimated standard error of 3.16 per cent corresponding to the complete enumeration of the 400 squares. Otherwise stated, when $m = 2.50$, a 'present-absent' enumeration of 400 squares is equivalent to a complete enumeration of 90 squares.

As one might expect, an optimum density exists for each value of t. For $t = 0$ ('present-absent' enumeration) this optimum is $m = 1.6$; the standard error of m from 400 squares in this case being 6.19 per cent in contrast to 3.95 per cent for a complete enumeration of all 400 squares. Alternatively stated, when $m = 1.6$ a complete enumeration of 160 squares is slightly less accurate than a 'present-absent' analysis of 400 squares. When t is greater than zero, the equations determining the maximum likelihood estimate, $\hat{m}$, of m cannot be solved directly, and solutions must be obtained by iteration. However, Tippett gives charts for $t = 1$, 2, and 3 from which the value of $\hat{m}$ is readily obtained. He gives also a graph from which the standard error of $\hat{m}$ can be estimated. Thus, for $t = 3$, the relevant 'data' of table 2 are the total number of squares examined and the number with none, one, two and three organisms respectively. They yield $\hat{m} = 2.51$ with an estimated standard error of 3.35 per cent, which compares favorably with the result obtained by a complete enumeration of 400 squares, *viz.*, 2.50 with an estimated standard error of 3.16 per cent. For $t = 1$ the optimum density happens to be $m = 2.5$ (from graph) and with $N = 400$ the standard error of the appropriate $\hat{m}$ is 4.3 per cent (from graph) so that when m is approximately 2.5 and a 'none—one—more-than-one' analysis of 400 squares is carried out, the probability is 0.95 that $\hat{m}$ is accurate to within 10 per cent, which is quite adequate for most purposes. Otherwise stated, a 'none—one—more-than-one' analysis of 400 squares is as accurate when $m = 2.5$ as a complete enumeration of 216 squares, and, without doubt, a great deal more rapid if Tippett's charts are at hand to facilitate the calculation of $\hat{m}$.

Statistical control of bacterial counts by plate method. Although many bacteriologists may never use the chamber method for counting organisms, the same can hardly be said about the plate method. Obviously then, of much more general application would be a procedure for statistical control of this technique. Since plate counts constitute samples from Poisson series, theoretically, the same type of test could be used as was described for the counting chamber, but considerations of time, labor, apparatus, and expense would render such a course highly

[9] An explicit account of the properties of maximum likelihood estimates is given in the Appendix. In this paper log refers to logarithms to base 10, ln, to base e.

impractical. If, however, as is often the case, series of counts are to be made on some material at certain intervals (daily, weekly, *etc.*) a statistical control on the precision of the plating technique is possible even though only 4 or 5 plates are used for each determination. Fisher (30, 34) has shown that if an index of dispersion,[10]

$$[3] \qquad D^2 = \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}} = \frac{N\Sigma x_i^2 - (\Sigma x_i)^2}{\Sigma x_i}$$

(where Σ denotes summation over i from 1 to N) is calculated from the counts $x_1, x_2, \cdots, x_N$ provided by a set of N parallel plates, then in a sequence of such sets D^2 will be distributed according to the X^2 distribution for $N - 1$ degrees of freedom *when the plating technique is in statistical control.* The expression at the extreme right of equation 3 is generally the more convenient for purposes of calculation.

Many investigators have used this valuable contribution of Fisher's for checking the accuracy and reliability of plate counts made on various materials—often with surprising and revealing results. In the case of pure cultures grown on specially developed media by means of carefully standardized techniques, the observed distributions of D^2 have on the whole agreed quite satisfactorily with the theoretical. These studies include data of: 3-plate counts of *Escherichia coli* in milk (data of Breed and Stocking discussed by Fisher, Thornton and Mackenzie (34)); 3- and 4-plate counts of *Rhizobium trifolii* on yeast-extract agar (114); 4-plate counts of *Bacterium globiforme* and *Pseudomonas fluorescens* on nutrient agar (92). Using special selective media, a number of workers have shown that more heterogeneous populations likewise give a reasonable distribution of D^2 if the technique is carefully controlled; these populations include protein- and starch-splitting organisms and actinomyces in soil (52), and actinomyces and fungi in soil (54).

Whenever a very complex population such as that found in the soil is studied, however, departure of the observed distributions of D^2 from the theoretical distribution is almost always noted. The departure usually consists of a great excess of large values of D^2, but occasionally an excess of subnormal variation is also found. In either case the use of the data for drawing any profound conclusions is highly questionable. Instead, steps should be taken to locate the origin of the abnormal variation and, if possible, to eliminate it. In the studies to date this has not always been successful, but it has been of assistance on several occasions and has definitely led to the uncovering of unsuspected information in the data or of defects in the technique.

Using Cutler's data on the number of organisms found in daily counts of the

[10] Fisher denotes this index of dispersion by X^2, and most writers have followed him in this usage, which has an excellent mathematical basis. We have made the change to D^2 in order to distinguish this index of dispersion from the X^2 *goodness of fit* criterion also discussed in the text (see p. 122 ff.). These two criteria are intimately related. It is our hope that, by using D^2 for the above and similar indexes of dispersion, and reserving X^2 for instances in which a frequency table (such as table 2) or a contingency table is concerned explicitly, the confusion which has arisen in some quarters may be lessened.

soil at the Rothamsted Experimental Station, Fisher, Thornton and Mackenzie (34) showed that in 156 sets of 4-plate counts and in 156 sets of 5-plate counts, an excess of both extremely low and extremely high values of D^2 occurred. If these were eliminated from the comparison, the remaining values agreed quite well with the theoretical. Searching for an explanation of the abnormal variances, they found that the excessively high values occurred in 'epidemics' during certain periods of the year. Although the origin of these epidemics could not be traced with certainty, evidences from other experiments suggested that they might be associated with the presence in the soil of certain species, usually of the spreading type, whose development inhibited the growth of other microörganisms. This not only led to an abnormally high variance (reflected in high values for D^2), but also seriously disturbed the reliability of the indicated mean.

The cause of the subnormal variance was even more obscure, but there was a suggestion that an apparently minor alteration in the preparation of the medium may have been a factor. These authors emphasize that an excess of low values for D^2 is just as much of a danger sign as excessively high values. Although no one is inclined to take too seriously results which show high variability, replicates in which the variation is abnormally low, far from exciting suspicion, are frequently exhibited as evidence of unusually reliable data. Fisher, *et al.* (34), cite, as an example, bacterial counts on cane sugar products in which the conditions which lead to the realization of the theoretical Poisson series were apparently operative in only about 45 per cent of the cases. An equal proportion was definitely subnormal with respect to variance, while 10 per cent were abnormally high. That some factor was concerned which disturbed random sampling was evident from the several sets in which the counts were practically identical on all six plates—a highly improbable result.

Harmsen and Verweel (52) likewise encountered an excess of high D^2 values from series of 10 plates used for counting bacteria in the soil of the Zuider Zee reclamation area in Holland. When soil or yeast-extract was added to the semi-synthetic medium used, the excessive variability diminished but did not completely disappear.

Probably the most extensive and thorough exploration of methods for estimation of microörganisms in the soil by standard plating methods in which the D^2 criterion was used for statistical control is provided by the studies of James and Sutherland (54–57) at the University of Manitoba, Canada. As has been already mentioned, control experiments (4 plates) with *P. fluorescens*, *B. globiforme* as well as mixtures of these pure cultures plus sterile soil led to distributions of D^2 which agreed most satisfactorily with the theoretical values. An example from their studies is shown in figure 2; they concluded that their laboratory technique introduced no significant source of variation, and that difficulties with the counting must be ascribed to other factors. When the technique was used for counts on soil (493 sets of 4-plate data in 1937, 468 sets in 1938), an excess of high values of D^2 was obtained. Investigation revealed that time of plating after taking the sample definitely affected the variability encountered as is illustrated in figure 3. Seeking an explanation of this rather unusual source of variance, James and Sutherland investigated a large number

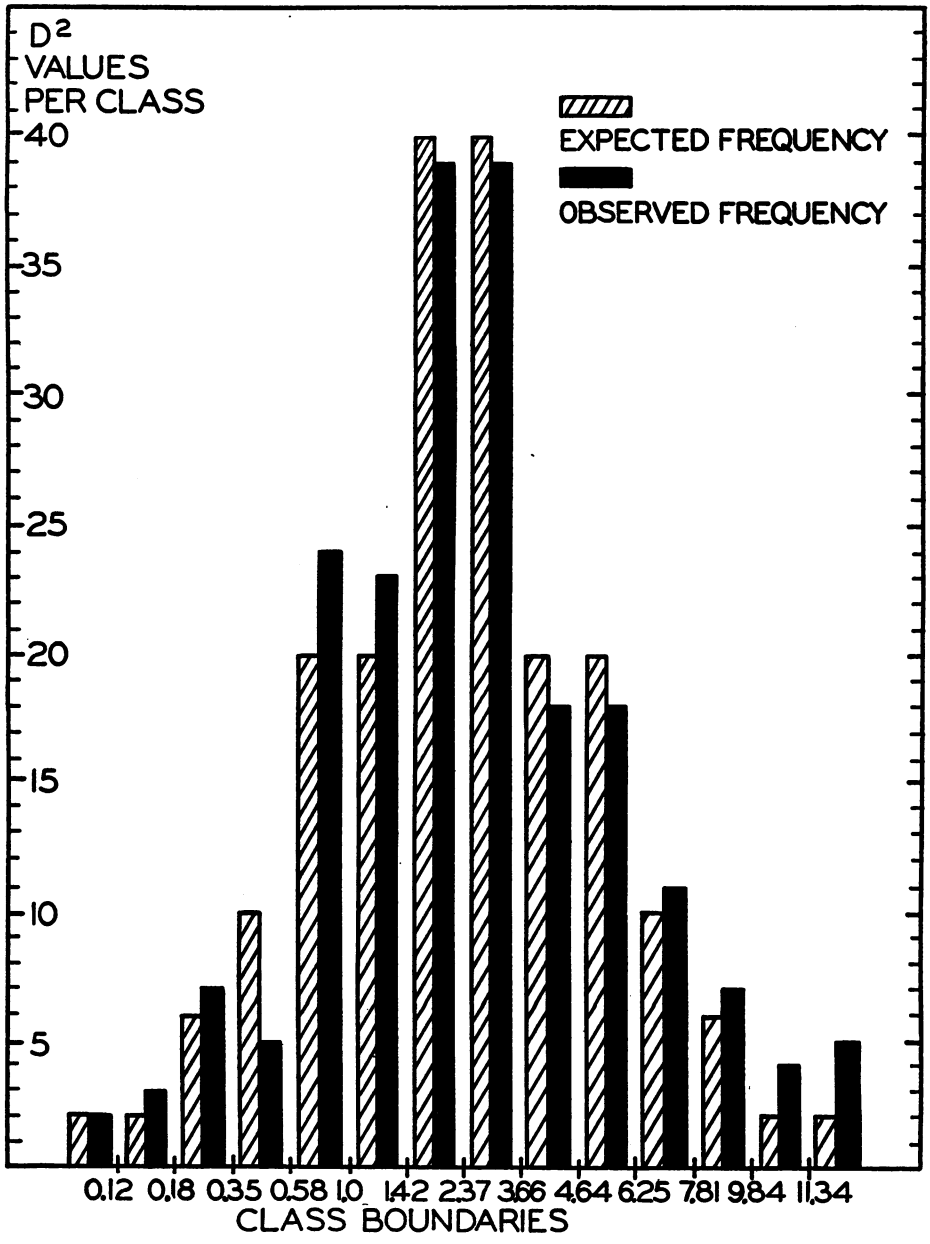

FIG. 2. HISTOGRAM OF D^2 DISTRIBUTION OF 200 SAMPLES OF PSEUDOMONAS FLUORESCENS PLUS STERILE SOIL

The value of D^2 for each set of 4 plates was calculated according to equation 3. The 200 values so obtained were classified according to the indicated class boundaries. The expected frequencies were calculated from an appropriate table of X^2 distribution with $n = 3$. Comparison of the *observed* with *expected* by the Chi Square goodness-of-fit test (see Part II) led to a value of $X^2 = 11.13$ corresponding to a probability of 0.6 that a worse fit might have arisen by chance. Both figures 2 and 3 are from the papers of James and Sutherland (54, 92). We thank these authors and the publishers of the *Canadian Journal of Research* for permission to reproduce these data.

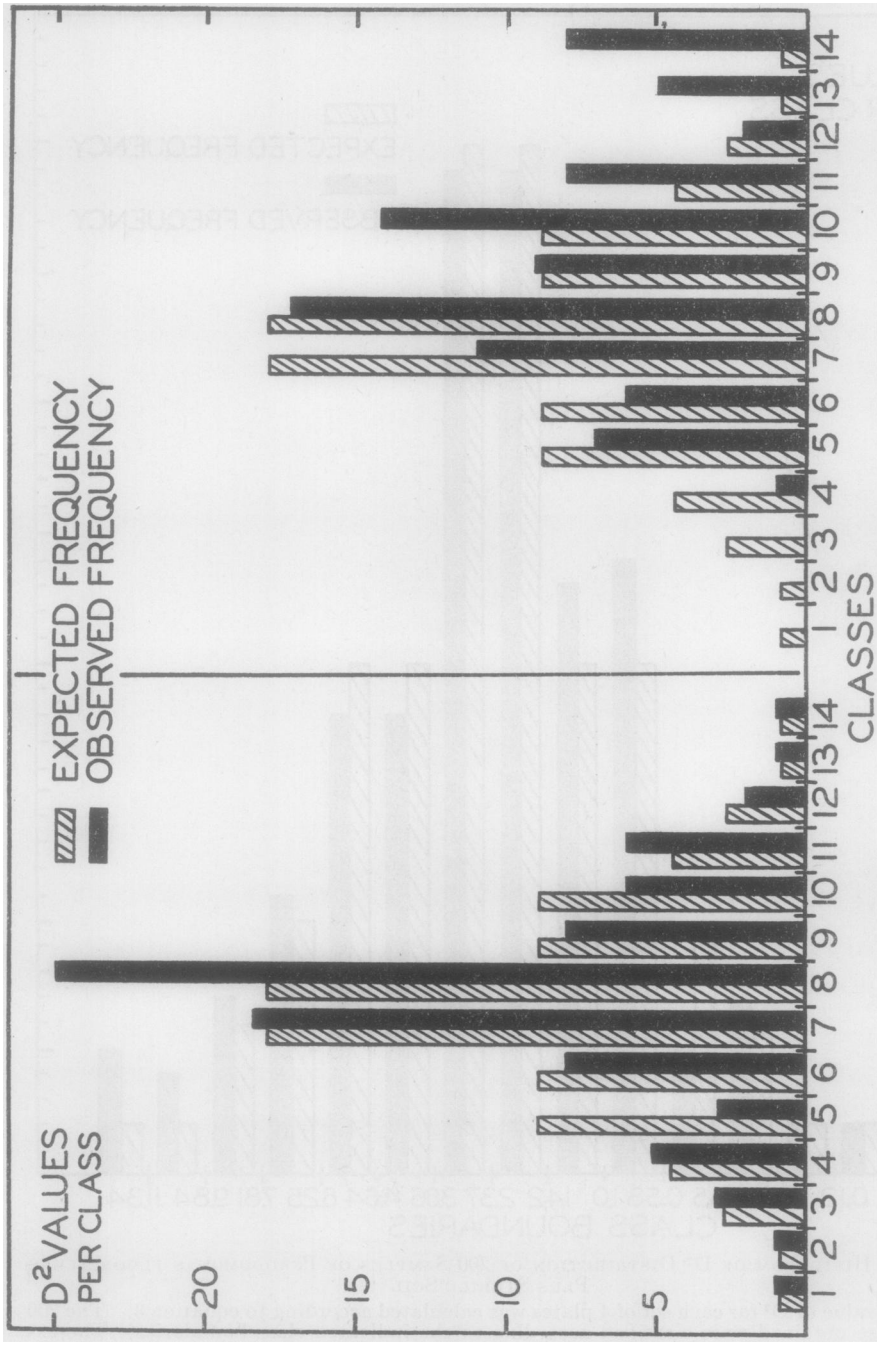

FIG. 3. HISTOGRAM OF D^2 VALUES ON SAMPLES OF SOIL

Left: Six replicates, 3 to 6 hours after sampling; $X^2 = 9.62$, $P = 0.73$. Right: four replicates, one day after sampling; $X^2 = 93.86$, $P < 0.01$. In this experiment further holding resulted in additional increases in X^2. Arbitrary numbers are used for class boundaries.

of possible factors and eliminated source and moisture content of the soil, technique, and medium. They did find, however, that associated with the abnormally large values of D^2 was the appearance on one or more plates of large numbers of pin-point colonies with or without large spreading colonies of the *Mucorales*. The presence of species of *Fusarium* or *Alternaria* as well as other fungi, which appeared only rarely, had no demonstrable effect on the value of D^2.

Control chart applied to plate counts. The foregoing method of appraising statistical control of plate counts from a sequence of sets of N parallel plates has three principal weaknesses. First, it cannot be applied until a large number of D^2 values has been obtained, by which time much of the data have become historic, and supplementary information which might throw light on the discrepancies is lost forever. Second, in forming a histogram (figures 2, 3) of the observed values of D^2 for comparison with the histogram expected on the supposition of statistical control, the order in which these values were obtained is disregarded, thereby discarding all characteristics of the *sequence* which are

TABLE 3

*Probability levels of D^2**

P \ N	2	3	4	5	6	7	8	9	10
0.995	0.000	0.010	0.072	0.207	0.412	0.676	0.989	1.344	1.735
0.975	0.001	0.051	0.216	0.484	0.831	1.237	1.690	2.180	2.700
0.500	0.455	1.386	2.366	3.357	4.351	5.348	6.346	7.344	8.343
0.025	5.024	7.378	9.348	11.143	12.833	14.449	16.013	17.535	19.023
0.005	7.879	10.597	12.838	14.860	16.750	18.548	20.278	21.955	23.589

N stands for the number of plates in the set.

P denotes the probability of a value of D^2 exceeding the value given in the body of the table when a state of statistical control prevails.

* Taken from table calculated by Thompson (98).

intimately associated with order. Third, in a laboratory where replicate plate counts are made at regular (or irregular) intervals, it does not provide a basis for action (acceptance or rejection) with respect to current determinations. A statistical control technique which does not suffer from these weaknesses is the *control chart method* developed at the Bell Telephone Laboratories by Dr. Walter A. Shewhart and now employed in various industries and by the United States Army Ordnance Department.[11]

The application of the control chart procedure to a sequence of values of D^2 is simple because the distribution of D^2, when a state of statistical control prevails, depends only on the number of plates involved. Table 3 gives probability

[11] The best general introduction to the control chart method is provided by the *American War Standards* published and sold by the American Standards Association (29 West 39th Street, New York City):

Z1.1 (1941) *Guide for Quality Control*

Z1.2 (1941) *Control Chart Method of Analyzing Data*

Z1.3 (1942) *Control Chart Method of Controlling Quality During Production.*

levels useful in constructing control charts for D^2 values. Figure 4 shows a control chart for some of A. R. Colmer's plate counts of total bacteria in 1:50,000

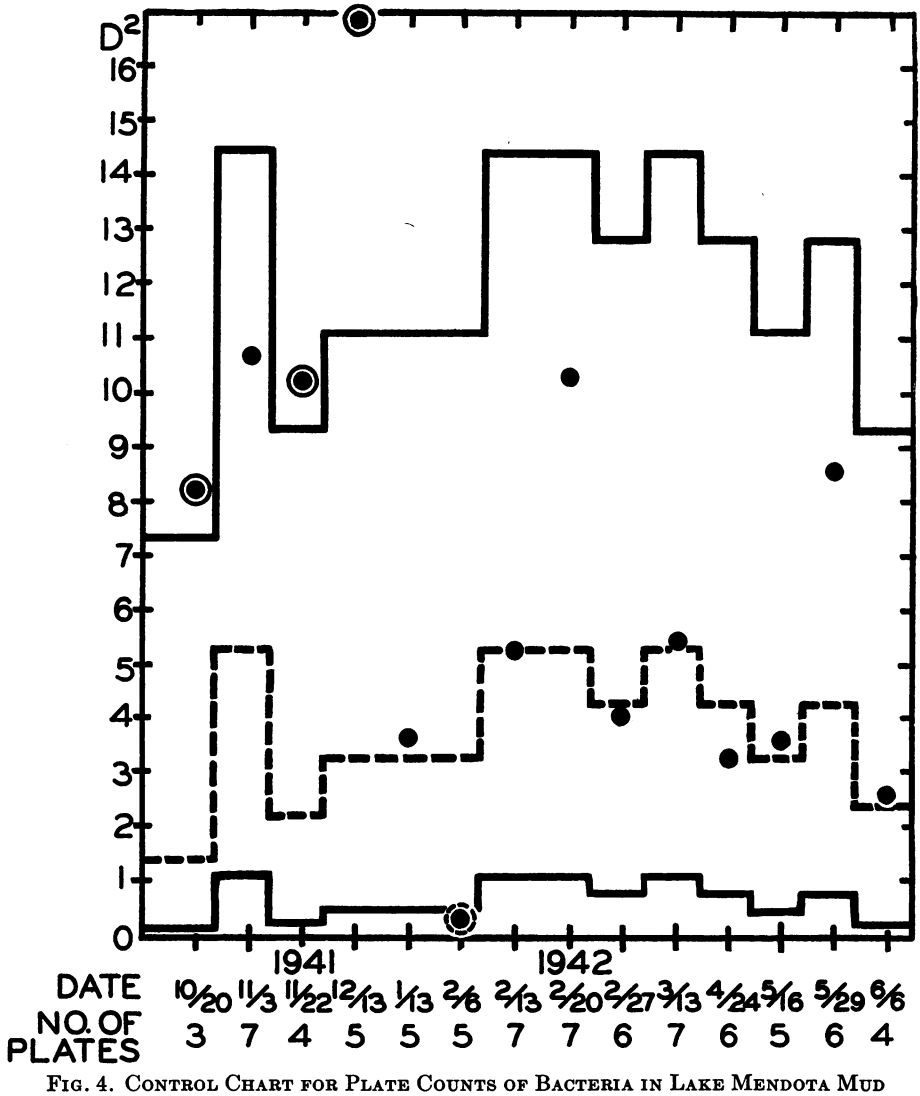

FIG. 4. CONTROL CHART FOR PLATE COUNTS OF BACTERIA IN LAKE MENDOTA MUD

Upper and lower heavy lines represent control limits outside of which experimental points should fall no oftener than 1 in 40 on average. These limits change because the number of plates counted vary, since one or more of the samples plated at each date had to be discarded because of spreading colonies. The dotted line represents the value of D^2 for P = 0.5; hence one-half of points should be above, one-half below.

dilutions of Lake Mendota bottom mud. *When a state of statistical control prevails,* the upper heavy line is the control limit which should be exceeded by only 1 point in 40 on the average, and the lower heavy line is the control limit

below which only 1 point in 40 should fall on the average. The control limits wag up and down in a manner dependent only on the number of plates of colonies counted. This varies since, although 8 plates were always prepared, film and other factors caused the rejection in every instance of one or more plates before the actual counts were made. Three of the first four sets of plates represented here gave D^2 values above the upper control limit, indicating excessive variation between the counts on the parallel plates. Reference to notes made at the time of counting, revealed that *Bacillus mycoides* was recorded as a disturbing factor in the plates rejected before counting, and the plates retained may have been affected to some extent by this factor. It is concluded that the mean counts corresponding to these points are not trustworthy. The sixth point is just below the lower control limit. No remarks were on record in the laboratory notes which might explain this extremely close agreement among the plates, and it was decided to regard the result as fortuitous. Beginning with the fifth, the points indicate a state of statistical control. As a further check on control, a central line corresponding to $P = 0.50$ has been added to the chart, and it may be noted that, beginning with the fifth point, six points are above and four points below this line—an excellent agreement with expectation.

The above choice of control limits may be expected to lead us to look for trouble once in 20 times when statistical control prevails. Therefore, if in a long plating program, experience shows that roughly 1 point in 20 lies outside these control limits, *and* that the great majority of these are 'false alarms', *i.e.*, no assignable cause for the discrepancy is discovered, then it will be desirable to use the upper and lower 0.005 limits so that 'false alarms' will arise only 1 time in 100 so long as statistical control prevails. Whatever limits are used, the median ($P = 0.50$) line should be drawn, and the occurrence of a statistically significant excess of points above (or below) this line, or long runs of points above (or below) this line, provides as much evidence of lack of statistical control as does the falling of points outside of the limits.[12]

The use of D^2 as a check on statistical control is not limited, of course, to plate counts. When direct counts by the microscopic or the chamber method are made in duplicate (or more), D^2 can be used to check the statistical control of the technique. When only two counts are involved,

$$D^2 = (x_1 - x_2)^2/(x_1 + x_2), \tag{3a}$$

and the control technique is readily applied.

Dilution count

Estimation of organisms by noting growth in successive dilutions was introduced early in bacteriology, but except for the important test for *Escherichia coli* in water, milk, and other products, its potential usefulness has been appre-

[12] The expected number of runs of length r above the theoretical median line (or below the median line) in a succession of m points is $(m - r + 3)/2^{r+2}$ for $1 \leq r \leq m - 1$ and $1/2^r$ for $r = m$. If the runs of length r on both sides of the theoretical median line are counted, the expected number is twice that given above. See W. G. Cochran (20).

ciated only recently. The dilutions are usually made in units of 10, and for many years the interpretation was simple and erroneous. If growth was obtained in a dilution of 10^{-2} but not in 10^{-3}, the count was said to be somewhere between 100 and 1000, which was probably true. If, however, a skip occurred (*e.g.*, growth in 10^{-2}, but not in 10^{-3}, growth in 10^{-4}, and no growth in the higher dilutions), it required an official ukase to obtain agreement. Officially the count was 1000 in this instance, since in case of skips the decree was that the result to be taken was the reciprocal of the dilution next higher than the smallest one giving a positive test. Although this solution may have worked satisfactorily in practice, it gave no greater assurance of accuracy than did the proposal to a state legislature to make the 'legal' value of π exactly 3.

The correct solution of the problem has occupied the statistician for many years, and judging by the recent output his interest remains undiminished. It is debatable whether bacteriologists have shared this concern, undoubtedly because few use the dilution method. Many may even question the appropriateness of including in this review an analysis of the rather extensive literature on this subject, arguing that the problem has intrigued the mathematician out of proportion to its practical value to the bacteriologist. Ample justification for doing so, however, exists. First, although no other aspect of bacteriology has been so thoroughly examined from the point of view of statistics, many important contributions have been published in journals seldom consulted by workers in this field. Second, important decisions affecting the health of all citizens are made in sanitary water analysis based on results of the test; correct interpretation of these is essential. Moreover, extension of the method to enumeration of organisms other than *E. coli* may provide a useful tool hitherto neglected. It appears to be superior to the plate count in certain cases of mixed populations for which selective media are used and may also prove useful in estimating organisms in unusual types of industrial products, *e.g.*, pickle brine in packing houses (124). Finally, it is emphasized that, although the principles involved have been developed from the point of view of estimating numbers of viable bacteria, extension of the reasoning and mathematics to other problems in bacteriology is possible, for example, number of bacteriophage particles in a suspension (17), direct count on bacterial smears (103), infestation of an animal by insects (48), and securing a pure culture of an organism by dilution (30). Because of these and other possible applications of the general theory on which the dilution count is based, the statistical literature on the method will be examined critically and in some detail.

Underlying probability theory. Before a satisfactory evaluation of the statisticians' contribution to this problem is possible, we should consider briefly what might be termed the philosophy of the various attempts to answer the following type question. Given a certain result, what can we say about its 'cause'? For example, if growth is obtained in a number of tubes of a medium when inoculated with a known dilution of bacteria and no growth in others, what is the 'best' estimate of the number of organisms in the original suspension?

First, let us consider the inverse probability approach. In calculating the 'most probable density' which gave rise to an observed result and its probability limits, it is assumed

that the sample is from a long (strictly infinite) sequence of samples in which all densities within certain limits occur with definite relative frequency. Suppose we obtain the result.[13] '5/10 in 1 cc'. From the appropriate equation to be developed in the next section, the relative frequency with which each density per 100 cc, x, will 'produce' the result '5/10 in 1 cc' may be calculated, and using these values as the ordinates corresponding to the proper x, a graph such as is shown in figure 5 is constructed. The curves shown in this figure correspond to the case where all admissible bacterial densities are *a priori* equally probable. Then, if in a given type of research, all densities within some range do occur with equal relative frequency in the long run, it follows from Bayes' Theorem[14] that when '5/10 in 1 cc' is observed it will have resulted from $x = 69$ in more cases in the long run than for any other single value of x. Therefore, if under these circumstances it is stated that $x = 69$ whenever '5/10 in 1 cc' is observed, this statement will be correct more frequently in the long run than if it were stated that x was some other number, such as 50, and it is in this sense that 69 is the 'most probable value' of x. Likewise under these circumstances it can be shown that 99 per cent of the results '5/10 in 1 cc' will be 'produced' in the long run by values of x less than 189, and it is in this sense that in a single such instance it is permissible to say that the probability is 0.99 that x is less than 189.

Unfortunately this correspondence with long-run experience in a sequence of actual assays depends upon the validity of assuming that in the long run all admissible values of x will occur with equal relative frequency. If certain values of x occur more frequently than others, so that the *a priori* distribution of x is not a constant within some range of x and zero elsewhere, then the 'most probable value' and the 'probability limits' for x will differ in general from those found by the above procedure, but can be found with the aid of Bayes' Theorem *when the a priori distribution of x is known.* When the *a priori* distribution is unknown, then information essential for the application of Bayes' Theorem is lacking.

Lack of factual information regarding the range of x, and of the relative frequencies with which values of x occur in a particular kind of research certainly do not constitute sufficient reason for assuming that all values of x within some range occur with equal frequencies. It is the merit of Bayes' Theorem, not its weakness, that the inherent probabilities of the admissible values of an unknown quantity are taken into account, and, when from experience the *a priori* probabilities are known to a fair degree of approximation, better estimates can be obtained by utilizing this information than by ignoring it. "Bayes' Theorem is just as sound logically as any other part of the Theory of Probability, and may be trusted to give reliable results *when we can get a grip on it.* The trouble is that we so seldom can" (36, p. 128).

Until quite recently, in cases where the *a priori* information needed for the application of Bayes' Theorem was lacking, there appeared to be no alternative other than assuming such *a priori* distributions as seemed reasonable or convenient, and then proceeding with Bayes' Theorem undaunted. Two papers by R. A. Fisher (28, 30a) give impetus to a new way of looking at the problem of estimation. In these papers Fisher showed that, in repeated sampling from the same population, *maximum likelihood estimates* (see Appendix) based on a large number of observations would hover at least as closely about the true value of a parameter as estimates obtained by any other procedure from the same number of observations, and that in many instances this property extended to maximum likelihood estimates based on only a few observations. The search for 'most probable values' of a parameter was abandoned, therefore, and maximum likelihood estimates accepted as 'good' estimates, since maximum likelihood estimates would *generally* be 'close' to the true

[13] '5/10 in 1 cc' means that from 100 cc of the suspension under investigation 10 subsamples of 1 cc were used to inoculate 10 tubes of which 5 showed growth. Since cc was used in the original publications, we use this symbol in these sections instead of the preferred ml used in the remainder of the paper.

[14] This important theorem of probabilty is concerned with the probability of causes. An excellent discussion of it is given by Fry (36, chap. V, sec. 95).

value of the parameter, *i.e.*, they will be 'close' to the true value except when the observations comprise an unusual sample from the population in question. Table 4 shows how this principle works and throws some light on the meaning of 'close' in the above context. The first column shows the possible outcomes with 10 tubes, the second gives the maximum likelihood estimates of x corresponding to these outcomes, and the third and fourth give the percentage of cases in which these outcomes (and hence these estimates) will be obtained when $x = 43$ and when $x = 138$, respectively. Thus, when $x = 43$, in about 50 per cent of the cases the result 3/10 *or* the result 4/10 will occur, leading x to be estimated as 36 or 51 respectively; only infrequently would an estimate of 0 or 10 be chosen; rarely would an estimate of 160 or more arise. Similarly, when $x = 138$, either 120 *or* 160 would be chosen about 53 per cent of the time, and an estimate of 36 or less would practically never be chosen. It should be noted, furthermore, that while the maximum likelihood estimate will generally be 'close', it cannot hit the nail on the head except in those cases where x happens to be one of the numbers which is a maximum likelihood estimate corresponding to a possible outcome of the experiment.

TABLE 4

Maximum likelihood estimates of, and upper 0.99 confidence limits for, bacterial density per 100 cc corresponding to all possible results 'in 1 cc,' and relative frequencies of occurrence when density is 43 and 138

RESULT IN '1 CC'	MAXIMUM LIKELIHOOD ESTIMATE	PERCENT OF CASES WHEN $x = 43$	PER CENT OF CASES WHEN $x = 138$	0.99 UPPER CONFIDENCE LIMIT
0/10	0	1.35	0.00	47
1/10	10	7.25	0.00	70
2/10	22	17.56	0.04	94
3/10	36	25.22	0.31	121
4/10	51	23.77	1.62	152
5/10	69	15.36	5.84	189
6/10	91	6.89	14.60	237
7/10	120	2.12	25.03	305
8/10	160	0.43	28.16	412
9/10	229	0.04	18.77	687
10/10	∞	0.01	5.63	∞

The final breaking away from the shackles of Bayes' Theorem took place about 1930 with the development of the concepts of *fiducial limits* by R. A. Fisher and of *confidence intervals* by J. Neyman (see Appendix). The construction for the result '5/10 in 1 cc' of a *confidence interval* of the form $x < M$ corresponding to a *confidence coefficient* of 0.99 will illustrate the procedure: Using the terminology customary in connection with tests of significance, an observed proportion, r'/n, will be 'significantly *less*' than a theoretical proportion, p, at the 0.01 level of significance if $P\{r/n \leq r'/n \mid p\} \leq 0.01$, where $P\{r/n \leq r'/n \mid p\}$ denotes the probability of observing a proportion as small as or smaller than r'/n when the true proportion is p. Now

$$[4] \qquad P\{r/n \leq r'/n \mid p\} = \sum_{r=0}^{r'} \frac{n!}{r!(n-r)!} p^r(1-p)^{n-r}$$

and from tables of this summation (12, 24) it is found that $P\{r/10 \leq 5/10 \mid p\} \leq 0.01$ for $p \geq 0.849$, so that 5/10 is significantly *less* than any proportion ≥ 0.849 at the 0.01 level of significance. It can be further shown that $p \geq 0.849$ implies $x \geq 189$ (63); hence $x < 189$ is the desired '0.99 confidence limit'.[15]

[15] The value of x corresponding to a probability of 0.99 determined from the '5/10 in 1 cc' curve of figure 5 (see section on *Accuracy of Estimate*) also is 189. In this particular case

In the last column of table 4, '0.99 upper confidence limits' are given for x corresponding to each possible outcome 'in 1 cc'. The 0.99 confidence property of these intervals can be seen as follows. Suppose repeated sampling is being done from a supply for which $x = 138$, then a *false* statement about x will be made *whenever* 0/10, 1/10, 2/10, or 3/10 occurs, since in these instances the inferences made will be $x < 47$, $x < 70$, $x < 94$, and $x < 121$, respectively. In these cases only will a false statement about x be made. But the probability of some one of these events occuring when $x = 138$ is only 0.0004 (by adding the probabilities of these events), so that the probability of some one of the other events (*i.e.*, 4/10 through 10/10), each of which lead to a correct statement, is 0.9996. Therefore, if whatever ratio arises, the corresponding interval is used, then the probability of a correct statement is ≥ 0.99 when $x = 138$. By virtue of the way in which these intervals were constructed they will have this same property for any value of x.[16]

The 'Best' Estimate. So far as the authors have been able to determine, McCrady (63) first approached the problem of estimating bacterial concentrations from dilution data with the aid of the theory of probability. He considered the selection of a single value to use as *the* estimate and the equally important question of the accuracy of the estimation. Acknowledgment was made of mathematical assistance received from Wm. D. Cairns, at that time Associate Professor of Mathematics at Oberlin College. Four cases were considered: (a) one dilution, one tube; (b) one dilution, several tubes; (c) several dilutions, one tube at each; and (d) several dilutions, several tubes at each dilution. No attempt was made in case (a) to obtain a single estimate of the number of bacteria. In cases (b) to (d), McCrady selected as the 'most probable number' the number which assigns the greatest probability to the event actually observed. Thus, the events '5/10 in 1 cc', '4/5 in 1 cc' and '9/10 in 1 cc' lead to the estimates 69, 160, and 229 bacteria per 100 cc, respectively, and not to 50, 80, and

---

the values of x corresponding to a given probability level are identical, independent of whether inverse probability or maximum likelihood statistics are used. This is largely a coincidence and in general does not obtain. It should be noted that no claim is made in confidence interval theory that a *single* 0.99 confidence interval such as $x < 189$ will include the true value in 99 per cent of the cases in which it is employed, *i.e.*, in which '5/10 in 1 cc' occurs. The confidence coefficient 0.99 applies to the *entire set* of confidence intervals which as an aggregate constitute an estimation procedure. If, as the different possible events occur, the *corresponding* intervals are employed, then in the long run 99 percent of the inferences regarding x made with this set of intervals can be expected to be correct irrespective of whether x varies from case to case or remains the same. This distinction between the two forms of inference does not appear to be adequately appreciated among research workers in spite of the fact that Fisher, Neyman and others have stressed it for over a decade. By a coincidence, 0.99 confidence limits for x, determined from the outcome of several tubes at a single dilution, are identical with 0.99 probability limits for x based on the assumption that all admissible values of x are *a priori* equally probable. Similarly, for other levels of confidence. Therefore, it is true that, *if all values of x do occur with equal relative frequency*, a single 0.99 confidence interval such as $x < 189$ may be expected to include the true value in 99 per cent of the cases in which it is employed; if the relative frequency of the values of x is otherwise, this expectation does not obtain.

[16] Owing to the discontinuous nature of the variable r/n it is not possible to construct intervals such that the probability is exactly 0.99 that a correct statement will be made. This can be done when the variable observed varies continuously. Note that a correct inference will *always* result in the present situation when $x < 47$.

90 per 100 cc as might be inferred. As developed by McCrady from the inverse probability point of view, his 'most probable number' depends on the validity of regarding all admissible concentrations of bacteria as being equally probable before the event. The 'most probable numbers' turn out to be 'maximum likelihood estimates' as well, but as pointed out in the preceding section, from the point of view of maximum likelihood the justification of the choice does not depend in any way upon the relative frequencies with which admissible values of the unknown quantity occur. Such estimates are not regarded as 'most probable values' but are chosen on the basis of the manner in which they are distributed about the true value in repeated trials.

It will be instructive to consider in detail an example of McCrady's estimation process. Case (c)—one dilution, several tubes—lends itself especially well to such consideration. He assumes that the x bacteria in the V units of volume comprising the sample under investigation are distributed randomly and independently throughout this sample. It follows that the probability of a single unit of volume containing *no* bacteria is $[(V-1)/V]^x$. Thus, when V is 100 cc and 1 cc is taken, the probability of *no* bacteria in the 1 cc is $(0.99)^x$, and the probability of some (*i.e.*, at least one) bacteria is $1-(.99)^x$. If n samples of V volume units each were drawn at random from the solution under investigation and a subsample of 1 volume unit taken from each, then the probability of exactly r of the subsamples containing bacteria is

$$[5] \qquad \frac{n!}{r!(n-r)!}\left[1-\left(\frac{V-1}{V}\right)^x\right]^r\left[\left(\frac{V-1}{V}\right)^x\right]^{n-r}$$

McCrady regarded equation 5, written in slightly different notation, as a sufficiently close approximation to the probability that exactly r out of n subsamples will contain bacteria when all n subsamples are taken from the *same* sample of V volume units. This is not strictly true since the reduction in number of bacteria is not necessarily proportional to the reduction in volume of fluid arising from withdrawal of the successive subsamples. As McCrady (63) notes, however, so long as n is small compared with V, the discrepancy will not be great. In figure 5, graphs of equation 5 for $V=100$ are given when $n=2$ and $r=1$, *i.e.*, for '1/2 in 1 cc', and when $n=10$ and $r=5$, *i.e.*, for '5/10 in 1 cc', with x as abscissa and the probability of the event concerned as ordinate. Both curves attain their maximum values for $x=69$, which is McCrady's estimate of the number of bacteria per 100 cc in either case. The maximum of equation 5 in general occurs at the value of x which is the solution of the equation

$$[6] \qquad 1-\left(\frac{V-1}{V}\right)^x=\frac{r}{n}$$

that is, by

$$[7] \qquad x=\left(\log\frac{n-r}{n}\right)\Big/\left(\log\frac{V-1}{V}\right).$$

McCrady gives equation 6 in slightly different notation, but not 7, and notes that at a given dilution the results r_1/n_1 and r_2/n_2 yield the same x whenever these fractions are equal. The estimations will not be of equal accuracy, however.

While McCrady's 1915 paper might be said to have 'completely solved' the case of one or more tubes at a single dilution, as much cannot be said of his treatment of the case of one or more tubes at each of several dilutions. He

showed how to develop equations from which can be calculated the 'most probable number' of bacteria per cc corresponding to the various possible outcomes of inoculating several tubes at each of several dilutions *when all admissible numbers of bacteria per cc occur with equal frequency in the long run.* The equation being somewhat difficult to solve, he gave a table of the 'most probable numbers' for "most of the practically possible" results which may occur from the systems: (a) Two tubes 'at 10 cc', ten tubes 'at 1 cc', and (b) two tubes 'at 10 cc', ten tubes 'at 1 cc', and ten tubes 'at 0.1 cc'. In a subsequent table (64) he gave the 'most probable numbers' corresponding to all possible combinations for several special cases including those where 5 and 10 tubes are used at each dilution. As already noted, *these* 'most probable numbers' are also the corresponding maximum likelihood estimates. Wolman and Weaver (119), by making a few minor approximations, rendered McCrady's equations easier to solve, but their contribution lost its importance once tables of the solutions were available. Continuing with the type of reasoning he employed in the case of one or more tubes at a single dilution, McCrady indicated how, with the aid of Bayes' Theorem, probability limits for the number of bacteria per cc corresponding to the various outcomes of several tubes at each of several dilutions might be obtained. He did not attempt to derive any formulae, however, and thus left unsolved the matter of accuracy of the estimates he tabulated.

Others were studying the interpretation of dilution data at about the same time as McCrady, and shortly after the publication of his first paper these researches began to appear in print. W. F. Wells, in a series of papers (106–109) and in a joint paper with P. V. Wells (110), considered various ways of handling and interpreting dilution data. Objections to these methods have been raised by various writers, among them Cairns (15), who, as noted above, assisted McCrady with the mathematical portions of his analysis.

For the case of several tubes at a single dilution, Stein (87) proposed the use of what amounts to the maximum likelihood estimates of the bacterial density (*e.g.*, number of organisms per cc) from the observed proportion of *negative* tubes. A table of these estimates is given but the values shown are not accurate. Using the formula for the standard deviation of an observed proportion under simple sampling,[17] he presented in tabular form calculations of the number of tubes necessary to make the standard deviation of the estimated bacterial density equal to 10 per cent or to 5 per cent of the true bacterial density. He found that, for bacterial densities between 1.058 and 1.900 per cc, the number of tubes needed to reduce the standard deviation of estimated density to 10 per cent of the true values is between 155 and 165, and that outside of this density range the number of tubes needed mounts rapidly. Fisher (30) has given the minimum number as "about 155" at a density of 1.6. In a second paper, Stein (88) furnishes a chart from which the estimated density corresponding to an observed proportion of positive tubes can be read. The mathe-

[17] Stein refers to the *standard deviation* as the "mean error", "expected error", and "probable error". These latter expressions generally have a different meaning in statistical papers.

matical discussion here is somewhat fuller, and two different approaches are given.[18]

Apparently unaware of McCrady's work, Greenwood and Yule (45) approached the problem of interpreting dilution data essentially as McCrady had. They introduced one important simplification (employed independently by Stein), which has been utilized by most subsequent writers. Instead of considering the preparation of the tubes at one or more dilutions as *subsampling* from a sample of volume V, which McCrady takes to be 100 cc, and attempting to estimate the number of bacteria x, in *this* sample, Greenwood and Yule regard the preparation of the tubes as constituting sampling from the supply itself, which is considered as having a practically infinite volume and in which the density of bacteria per cc is λ. In this way, the difficulties arising from changes in the volume as successive tubes are prepared is avoided, and the formulae are also somewhat simpler.[19] A table is given of the maximum likelihood estimates—interpreted as 'most probable values' as in McCrady—for all results involving at least one positive and at least one negative tube, corresponding to the use of 10 or less tubes at a single dilution. Where comparisons are possible, these estimates of Greenwood and Yule agree with Fisher's values (30) except for occasional difference of unity in the last digit.

On the general question of what series of dilutions to use and with what numbers of tubes at each dilution, Greenwood and Yule remark: "One obvious condition, strangely overlooked, is that the size of any one sample should be greater than the sum of the sizes of the smaller samples. Otherwise the observer is simply asking for 'inconsistencies' in his results.[20] A geometrical series fulfills the required condition ... [and] ... seems also a natural one to use as the chance of an inconsistence is the same at every point of the series [when an equal number of tubes is used at each dilution]. ...[With a single tube at each dilution] r being the (ascending) ratio of the series, the chance of an inconsistence between any adjacent pair of samples ... is $1/(r + 1)$."

Basing his analysis on the results of Greenwood and Yule (45), which were obtained with the aid of Bayes' Theorem and the assumption that all admissible

[18] We have found two errors in connection with the second approach. First, Stein's relation between a, the number of bacteria per cc in the supply, and Q, the expected proportion of negatives in N tubes when n cc are introduced into each tube, is incorrect. The correct formula is $a = -(1/n) \ln Q$. Formula XI (87, p. 254), expressing the relation in terms of Q for the simple sampling deviation of the number of bacteria per cc in a random sample of Nn cc, is also incorrect because the incorrect relation between a and Q was used.

[19] Since the limit of $[(V - v)/V]^x$ is $e^{-v\lambda}$ as V and x both increase indefinitely with x/V tending to λ as a limit, the latter (with N for v) appears in the formulae of Greenwood and Yule where the former occurs in McCrady's formulae. Accordingly, by means of this relation, one can pass from McCrady's results in terms of V, x, and v to Greenwood and Yule's results in terms of λ and N, and *vice versa*.

[20] When a larger proportion of tubes give positive results at a certain dilution than at a *lesser* dilution, the result is said to be 'inconsistent' or 'anomalous'. Thus, with one tube at each of three increasing dilutions, the results $++-$ and $+--$ are 'consistent', whereas the results $+-+$ and $--+$ are 'inconsistent'. (In this quotation the expressions in square brackets have been inserted by the present writers.)

densities were equally probable *a priori*, Reed (78a) made a detailed study of the case where the estimation is based on a set of five tubes containing 100 cc, 10 cc, 1 cc, 0.1 cc, and 0.01 cc, respectively, of the solution under investigation. He notices that the most probable densities (which are maximum likelihood estimates also) corresponding to the various possible outcomes form a "yardstick" with "very coarse divisions," and for this reason this set-up "is suitable for grading waters that vary widely in the extent of pollution." He notes, and illustrates graphically, that for the so-called consistent outcomes, the estimated density and its accuracy (from the inverse probability viewpoint, at least) are almost entirely determined by the two tubes where the results change from + to −. In the so-called inconsistent cases, however, both changes of sign play a part in determining the estimated density and its estimated accuracy, and "it would be better to regard them as further subdivisions of the yardstick, having their own probabilities, than to treat them as inconsistencies." He states that the use of five tubes at a single dilution will be more accurate than five tubes in geometric dilution series, at least when accuracy is evaluated by the Bayes' Theorem approach, provided the five identical tubes are run at the most suitable dilution. Except for cases where the customary neighborhood of the bacterial density is known, the difficulty lies in picking the most suitable dilution beforehand.

Fisher (28) considered the case of several tubes at each of several dilutions from the viewpoint of the method of maximum likelihood, and gave in condensed notation the general equation determining the maximum likelihood estimate of the bacterial density. Except for notation, this equation is identical with that obtained by Greenwood and Yule for the 'most probable density' on the assumption that all admissible densities are equally probable *a priori*, and with the equation obtained by Halvorson and Ziegler (49) for the 'most probable value' under the same assumption. The last mentioned writers claim that their equation is more general than the equation of Greenwood and Yule. There appears little justification for this claim, as noted by Swaroop (93), who also shows how, by a slight rearrangement of the equation, its solution may be obtained more readily. Swaroop prepared a table of estimates for the case where the dilutions are in the ratio 1/2, 1/10, 1/100 with a single tube at the first dilution and 5 tubes at the other two dilutions, and for the case where the dilutions are in the ratio 1/10, 1/100, 1/1000 with 5 tubes used at each dilution. He remarks that two mistakes in McCrady's tables have been uncovered, and that McCrady's approximate values (when the estimates exceed 20) have been replaced by values correct in the units place. Halvorson and Ziegler (48, 49) have tabulated the estimates corresponding to code numbers[21] likely to be encountered in practice for the case where the (ascending) dilution factor is 10, and 10 tubes are used at each dilution.

[21] If 10 tubes are used at each of five consecutive dilutions, an overall code of 10-10-10-8-3 might be observed, meaning that all 10 tubes show growth at the first three dilutions, only 8 of the 10 in the next, and 3 of the 10 in the final. The critical code for entering the table is 10-8-3, and multiplication of the tabulated estimate by the appropriate power of 10 would give the density per cc, per 100 cc, *etc.* as desired.

As already stated, the 'most probable density' derived on the assumption that all admissible densities are equally probable *a priori* advocated by McCrady, Greenwood and Yule, Stein, Reed, Halvorson and Ziegler, and others is identical with the estimate indicated by the method of maximum likelihood advocated by Fisher and Swaroop. Therefore, so far as choice of a *single* estimate of λ is concerned, these writers are unanimous, although the reasons for their choice differ. Gordon (40–42) takes issue with all of the foregoing writers—although he singles out the work of Halvorson and Ziegler for specific criticism—and proposes an estimate, which, though based on Bayes' Theorem and the assumption that all admissible densities are equally probable *a priori*, differs from the 'most probable density' corresponding to this assumption. *If* this assumption is valid and simple sampling prevails, then, in our opinion, a table of the 'most probable densities' being available, it is curious to advocate any other estimate, since the 'most probable density' can be expected under these conditions to hit the nail on the head more often than any other. While Gordon does not question the use of Bayes' Theorem with uniform *a priori* distribution for λ, he does question the validity of the simple sampling assumption, which underlies the work of the aforementioned writers, that the individual bacteria are distributed in the fluid *independently at random*. It is his thesis that bacteria "exert a certain mutual uniformizing influence on one another" so that "the numbers of bacteria caught up in [a series of samples] are somewhat more closely clustered about the true average density, than we should compute them to be on the basis of the above assumption of complete randomness." Gordon apparently claims (41, p. 169) that his method is less affected by bias from this source, but it is difficult to see how he reaches this conclusion since his analysis utilizes a formula (equation 3, p. 170) which is based on simple sampling and which differs only in notation from the fundamental formulae employed by the other writers.

Pearson (74) has given a very careful evaluation of Gordon's paper, including a discussion of the relative merits of the inverse (Bayes' Theorem) and direct (maximum likelihood, confidence interval) methods of approach. Noting that the 0.95 confidence intervals are quite broad, (see section on *Remarks and Suggestions*, especially table 6), Pearson remarks that "having regard to this, the differences between the single value estimates are of little importance." In reply to Gordon's implications that the maximum likelihood estimates are biased, Pearson states "if a method of obtaining accurate fiducial limits were available, the bias in the single-valued estimate would be of no importance." An accurate chart of confidence intervals would provide such a method, and would also make possible a check on agreement of actual samples with expectancy under simple sampling, *i.e.*, a check on *statistical control* of laboratory technique.

Accuracy of the 'Best' Estimate. In the preceding section we have focused our attention primarily on the efforts of different contributors to determine the 'best' single value from a given set of laboratory results. Although knowledge of the 'most probable' number of organisms is of undeniable importance, it has little significance unless something about its accuracy is also known. We shall

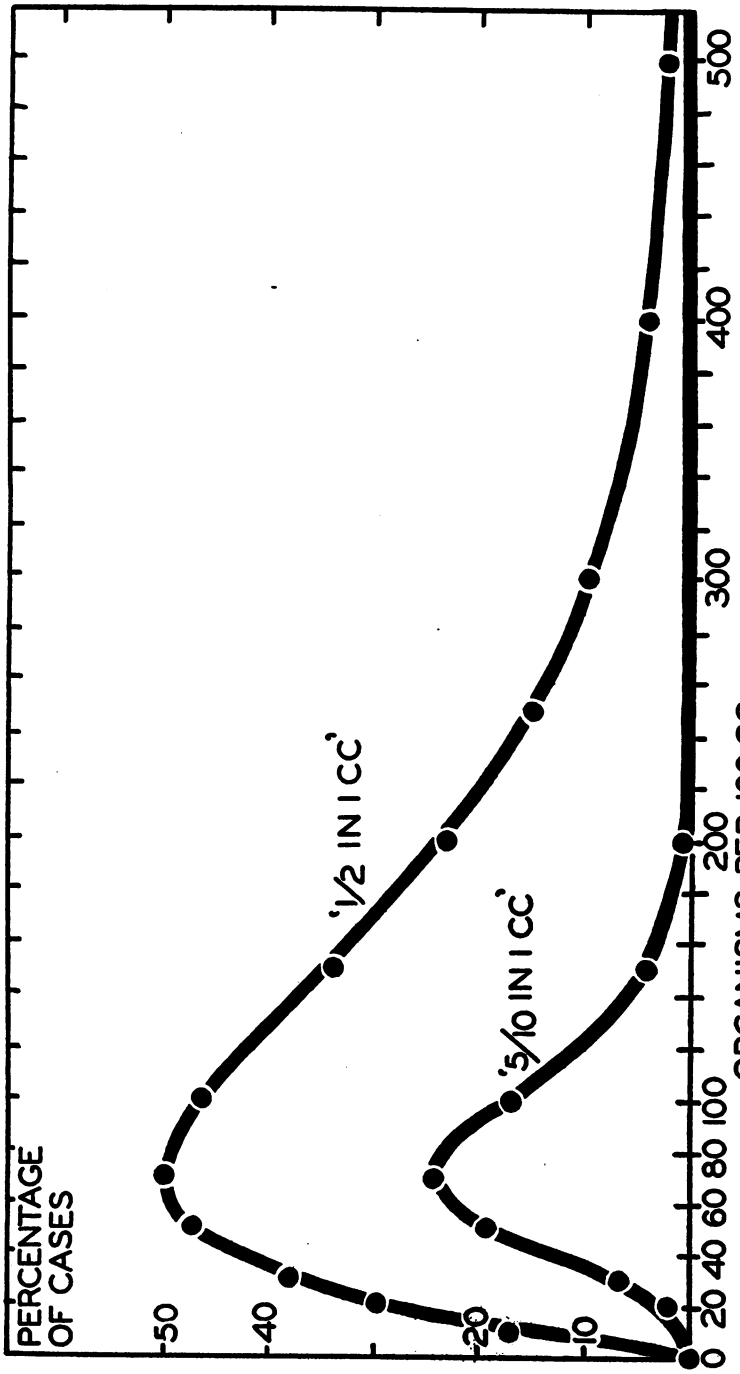

FIG. 5. EXPECTED FREQUENCIES OF TWO POSSIBLE RESULTS FOR VARIOUS TRUE DENSITIES OF BACTERIA

When the true number of organisms is that given in the abscissa, the observed result, 'one tube in two inoculated show growth' (upper curve), will on the average be observed in the indicated percentage of cases. The lower curve represents the result, 'five tubes of ten inoculated show growth.' Note that 'most probable number' (maximum in curve) is the same for both curves since in both instances 50% of tubes show growth. The 'range' (for example, the number of organisms which will include 95% of the cases) however, is much smaller with the larger number of tubes. Chart constructed after McCrady (63).

now consider how various workers have sought to determine this property of their estimates. McCrady's (63) approach to the problem of accuracy is best summarized in his comments regarding the curve for '5/10 in 1 cc' (figure 5): "Inspection of the corresponding curve . . . shows that the practically possible numbers of bacteria have a range from about 15 to about 200 per 100 cc." He gives no range for the '1/2 in 1 cc' case but, presumably, he would consider the "practically possible" range from 1 to 600 per 100 cc. To explain his reasoning, he points out that by using Bayes' Theorem with equal *a priori* probabilities, "the ordinates of the curve give the relative probabilities that the corresponding abscissae were responsible for the result," so that "the general shape of the curve indicates roughly the degree of confidence which may be assigned the inclusion of x within certain limits." As an elaboration of this procedure, he notes that, by summing the ordinates of the curve from $x = 0$ to $x = k - 1$ and dividing by the sum of the ordinates from 0 to ∞, the probability that the samples contained less than k bacteria per 100 cc can be obtained. Formulae are given (63, p. 197) to facilitate these summations. As an example he shows that, all admissible values of x being assumed equally probable *a priori*, the occurrence of the result '1/2 in 1 cc' implies that the probability is $89.937/99.503 = 0.90386$ that x is less than 300. With the same assumption, it may be shown that the results '1/2 in 1 cc' and '5/10 in 1 cc' imply that the probability is .99 that x is less than 527 and less than 189, respectively. Since both of these results yield 69 as the 'most probable value', the greater accuracy of the result corresponding to 10 tubes is evident.

In his 1919 paper, Stein (87) furnishes a curve showing the estimated density $\pm$ its standard deviation for 30, 100 and 360 tubes. He appears to have used the standard deviation of a_N, the number of bacteria per cc in a *random* sample of Nn cc from a supply in which the actual density is a per cc. Since a_N cannot be observed, its standard deviation is not of great practical value. What can be observed is the proportion, say q, of negative tubes of N inoculated with n cc of the supply; from q an estimate, say $\hat{a}$, of a can be calculated: $\hat{a} = -(1/n) \ln q$. To a first approximation, the standard deviation of $\hat{a}$ is $\sqrt{(e^{na} - 1)/N}$, of which the estimate from the data is $\sqrt{p/qN}$, where $p = 1 - q$ is the observed proportion of positive tubes. This latter formula was given by Greenwood and Yule in different notation. The standard deviation of $\hat{a}$ will always be larger than that of a_N since $\hat{a}$ depends merely on the presence or absence of bacteria in the Nn cc withdrawn and takes no account of the number of bacteria in the positive tubes. Thus for $\hat{a} = 0.7$, corresponding to $q = 0.50$, the correct limits are about 20 per cent wider than indicated in Stein's curves. However this does alter the fact, noted by Stein, that above densities of about 1.9 per cc (expected proportion of positive tubes, above 0.85), these curves become so flat that small changes in the observed proportion cause large changes in the estimated density.

To obtain limits which are today interpreted as confidence limits, Stein proposed the use of Tschebycheff's inequality—in the first paper he appears to suggest that this be applied with the standard error of the observed proportion of positive (or negative) tubes to obtain limits for the expected proportion, from

which limits those for the bacterial density could then be calculated from the equation relating these two quantities. In the second paper he inadvertently drops the inequality sign (and the inequality is generally great) and apparently proposes its use with the standard error of the estimated density to get limits for the actual density. These two procedures are not equivalent; thus, with 100 tubes and an observed proportion close to 0.5, the former would assign a confidence coefficient of *at least* 0.96 to the density range of 0.29 to 1.40 bacteria per cc, and the latter to a range of 0.35 to 1.05 bacteria per cc. Actually the limits obtained by either of these processes are too wide, since Tschebycheff's inequality is extremely loose—a correct range for the above confidence coefficient would be approximately 0.51 to 0.92 bacteria per cc.

Fisher (28), employing a method available only for maximum likelihood estimates, obtained the variance of ln $\hat{\lambda}$, where $\hat{\lambda}$ (which he denotes by n) is the maximum likelihood estimate of the density, for the case where s, the number of tubes used at each dilution is large. When s is small, Fisher's expression will give the variance of ln $\hat{\lambda}$ to a first approximation. Only minor alterations are necessary to extend Fisher's expression to cover the case where unequal numbers of tubes are used at the respective dilutions, and, although Fisher's derivation is for the case where the dilutions are in geometric progression, the validity of his result can be extended without difficulty to cover other arrangements.

Using Fisher's method and notation, Swaroop (93) derived an expression for the first approximation to the standard error of $\hat{\lambda}$, the maximum likelihood estimate of the bacterial density, λ. Substituting $\hat{\lambda}$ for λ in this formula, he tabulated (together with the values of $\hat{\lambda}$) for certain combinations of tubes of dilutions the corresponding estimates of the standard deviation of $\hat{\lambda}$. In two subsequent papers (94, 96), he studied the effect on the standard error of $\hat{\lambda}$ and on the coefficient of variation of $\hat{\lambda}$ of varying: (a) the number of tubes used at each dilution, (b) the true bacterial density, and (c) the number and type of dilutions used. He found that all three factors must be taken into account in determining the accuracy of $\hat{\lambda}$. For low densities, *e.g.*, 20 organisms per 100 cc, he points out that an equal number of tubes at the dilutions 1/2, 1/10, 1/100 provides more accurate results than the same number of tubes at each of the dilutions 1/10, 1/100, 1/1000. To facilitate the use of the former dilution system he tabulated (95) the estimates $\hat{\lambda}$ and their estimated standard errors when 2, 3, 5, or 10 tubes are employed at each dilution.

In connection with Swaroop's results two points need to be kept in mind. First, the formula employed for the standard error of $\hat{\lambda}$ is strictly valid only when a large number of tubes are employed, giving merely a first approximation of *unknown accuracy* when only a few tubes are involved. Second, the standard error of a quantity is a good measure of its sampling variability only when the distribution of the quantity is approximately normal (Gaussian). To what extent these may impose limitations in actual practice can be inferred from direct calculations carried out by Halvorson and Ziegler (50, 51). For the case of ten tubes at each of three dilutions in geometric progression with (ascending) factor of 10, Halvorson and Ziegler (51) calculated, from the terms of an appro-

priate multinomial expansion, the exact probabilities of observing the various possible codes when the true density was 0.15, 0.25, 0.50, and 1.50 per cc, and then, with the aid of their tables showing the correspondence between codes and estimates, they obtained the probability distribution of the maximum likelihood estimator, $\hat{\lambda}$, when the true density was each of the four preceding values. They found: (a) that the distributions of $\hat{\lambda}$ in these cases were moderately skewed with

TABLE 5

Comparison of arithmetic and logarithmic estimation

NOTE	FORMULATION	ACTUAL BACTERIAL DENSITY PER CC (λ)			
		0.15	0.25	0.50	1.50
(i)	$M(\hat{\lambda})$	0.164	0.284	0.558	1.648
(ii)	$\sigma(\hat{\lambda})$	0.0659	0.1171	0.2263	0.6894
(iii)	$\sigma(\hat{\lambda})/\lambda$	0.440	0.468	0.453	0.460
(iv)	Lim $\sigma(\hat{\lambda})/\lambda$	0.354	0.353	0.407	0.357
(v)	$[M(\hat{\lambda}) - \lambda]/\sigma(\hat{\lambda})$	0.212	0.290	0.256	0.215
(vi)	$\log \lambda$	−0.824	−0.602	−0.301	+0.176
(vii)	$M(\log \hat{\lambda})$	−0.816	−0.578	−0.285	+0.184
(viii)	$\sigma(\log \hat{\lambda})$	0.163	0.164	0.163	0.168
(ix)	Lim $\sigma(\log \hat{\lambda})$	0.1535	0.1532	0.1768	0.1550
(x)	$\dfrac{M(\log \hat{\lambda}) - \log \lambda}{\sigma(\log \hat{\lambda})}$	0.049	0.146	0.098	0.048

Notes:

(i) The arithmetic mean of the distribution of the maximum likelihood estimator, $\hat{\lambda}$, taken directly from table 1 of Halvorson and Ziegler (51).

(ii) The standard deviation of $\hat{\lambda}$, *i.e.*, the root mean square deviation from $M(\hat{\lambda})$, taken from table 1 of Halvorson and Ziegler (51).

(iii) Entries obtained by dividing those in (ii) by λ, and *not* equal to Halvorson and Ziegler's coefficient of variation from mean, since their divisor is the corresponding entry in (i).

(iv) The limiting value of $\sigma(\hat{\lambda})/\lambda$ as the number of tubes at each dilution increases indefinitely, taken from table 2 of Swaroop (94).

(v) Obtained from the preceding rows of the present table.

(vi) The logarithm (to base 10) of the actual bacterial density per cc.

(vii) The arithmetic mean of $\log \hat{\lambda}$, taken from table 1 of Pearson (74) who calculated them from table 1 of Halvorson and Ziegler.

(viii) The standard deviations of $\log \hat{\lambda}$ taken from Pearson's table 1, who calculated them from table 1 of Halvorson and Ziegler.

(ix) The limiting value of $\sigma(\log \hat{\lambda})$ as the number of tubes at each dilution increases indefinitely, taken from Pearson's table 1 who calculated them from the formula of Fisher (28).

(x) Calculated from the preceding rows of the present table.

the long tail toward the large values; (b) that the degree of skewness was practically independent of the true density, λ, for the range of λ considered; (c) that the standard deviation of $\hat{\lambda}$, $\sigma(\hat{\lambda})$, increased with λ; and (d) that the coefficient of variation, $\sigma(\hat{\lambda})/\lambda$, was practically constant for the range of λ considered. Some of their results are given in table 5, together with additional results derived from their exact distributions of $\hat{\lambda}$ and from Fisher's formula.

From this table it is evident *for the density range considered*: (a) that the standard deviation of $\hat{\lambda}$ varies with λ, being to a good approximation equal to 0.45λ; (b) that the standard deviation of $\log \hat{\lambda}$ is nearly constant; (c) that $\log \hat{\lambda}$ and $\hat{\lambda}$ both have a positive bias, and in standard deviation units the bias of the former is somewhat less; and (d) that the relative error in using the limiting standard deviation in place of the exact value is less with $\log \hat{\lambda}$ than with $\hat{\lambda}$. Pearson (74) gives a table of values for the limiting standard deviation of $\log \hat{\lambda}$ for λ from 0.10 to 4.00, and from a graphical portrayal of these values it appears that the limiting standard deviation of $\log \hat{\lambda}$ ranges from 0.151 to 0.180, with a median value of approximately 0.166. Since the distributions of $\log \hat{\lambda}$ are more nearly symmetrical than the distributions of $\hat{\lambda}$, it is of interest to see how well the distributions of $\log \hat{\lambda}$ can be approximated by assuming that it is normally distributed with a standard deviation of 0.166 about $\log \lambda$. Using this approximation, the probability should be 0.95 that $\hat{\lambda}$ will lie in the intervals 0.07–0.32, 0.118–0.723, 0.236–1.058, 0.671–3.158 (end points included) when $\lambda = 0.15$, 0.25, 0.50, and 1.5, respectively. From Halvorson and Ziegler's table 1 the exact probabilities appear to be 0.949, 0.986, 0.948, 0.937, respectively. The agreement is remarkably good, especially in the first and third cases, in view of the fact that $\hat{\lambda}$ in each instance can take only *certain* discrete values, the probabilities of which do *not* increase monotonically but exhibit many vicissitudes as $\hat{\lambda}$ approaches the true value, λ, from either side.

Remarks and Suggestions. To remind a bacteriologist of the many statistical procedures which have been advocated for solving the type problem represented by the dilution count may not strike him as entirely a favor. He may, indeed, find himself somewhat distracted by the variety of aid offered, his position not unlike that of Joel Chandler Harris's Br'er Fox hesitating to decide which barbecue to attend. The bacteriologist of this joint undertaking, therefore, put these two questions to the statistician:

1. Which of the proposed methods should I employ in a given situation at the present time, *i.e.*, with only existing tables and charts to facilitate their application?

2. If funds were available, what further research should be undertaken to advance statistical methodology relating to the dilution count?

It is realized that the following answer of the statistician represent "one man's opinion" and is to be accepted or rejected as such.

In order to answer the first question, a more specific definition of the situation is desirable. If a routine check on a controlled bacterial population is concerned, as in a routine analysis of the water in a city reservoir, so that the customary whereabouts of the bacterial density is known, the most effective 'control' will be obtained by running all tubes at a single dilution, the dilution and number of tubes being chosen in such a way that the probability of obtaining a proportion of positive tubes which will result in condemnation will be very small when the bacterial density is within the customary neighborhood, and large when the bacterial density exceeds the permissible limit. The standard proposed in the Report to which Reed's (78a) analysis is an appendix is a step in this direction.

To improve the chances of a routine analysis detecting the entrance of trouble before it leads to condemnation, the routine should include the use of a control chart in terms of the proportion of positive tubes, as outlined in American War Standard Z1.3.[11]

If on the other hand, an isolated analysis or a routine analysis of an uncontrolled bacterial population is concerned, as in an analysis of stream water prior to treatment at the Water Works, the general whereabouts of the bacterial density being unknown in advance, then 10 tubes should be run at each of a series of dilutions in geometric progression with (ascending) ratio of 10. A single estimate of the bacterial density can readily be obtained from the table of Halvorson and Ziegler (48, 49); its confidence limits can be estimated from a chart[22] prepared by Miss Supińska (91). In the absence of Miss Supińska's chart, confidence intervals computed from $\log \hat{\lambda} \pm (1.96)\,(0.166)$ can be used, the associated confidence coefficient being close to 0.95.

Table 6 gives, for three hypothetical cases discussed by Gordon (41), the maximum likelihood estimate, $\hat{\lambda}$, tabulated by Halvorson and Ziegler; the estimate recommended by Gordon, $\bar{\lambda}$; and the upper and lower limits to 0.95

TABLE 6

Estimation by interval and by single value

CODE	$\hat{\lambda}$	$\bar{\lambda}$	0.95 CONFIDENCE INTERVALS					
			Supińska		Normal Approximation		Matuszewski, *et al.*	
			λ min.	λ max.	λ min.	λ max.	λ min.	λ max.
10–7–3	1.53	1.43	0.67	3.00	0.72	3.24	0.75	4.28
8–5–1	0.267	0.291	0.125	0.525	0.126	0.563	0.179	1.153
4–2–1	0.080	0.086	0.029	0.165	0.041	0.182	0.034	0.203

confidence intervals obtained (a) by Supińska's method, (b) by the normal approximation discussed above, and (c) by a method described by Matuszewski, *et al.* (67) which depends only upon the sum of the three components of the code. The intervals obtained by normal approximation agree moderately well with

[22] This chart has been reproduced by Matuszewski, Neyman, and Supińska (67). It was not based on an exact mathematical solution, but was obtained by graduating a series of experimental sampling results as described on p. 76 of their paper, *viz.*,

"The method followed by Miss J. Supińska consisted in a complex sampling experiment, using Tippett's random sampling numbers. The experiment produced a series of values of the variates x_0, x_1, and x_2 [*i.e.*, codes x_0–x_1–x_2] following the sampling distribution which they would follow in our hypothetical conditions of the experiment. For each series of x_0, x_1, and x_2, it was possible to read up from the table of Halvorson and Ziegler an estimate, say λ', of the concentration λ. The estimates λ' have been then tabulated and an empirical frequency distribution of λ' corresponding to several fixed values of λ has been determined. Following the method described by J. Neyman, these empirical frequency distributions were then used to construct confidence intervals as if they were the accurate ones. As the random variation could not fail to affect the limits of the intervals it was felt necessary to correct them by fitting two parabolae, one marking the lower and the other the upper limits of the confidence intervals."

Miss Supińska's intervals, and in the absence of her chart will probably be close enough for most practical purposes. It will be noted that they are slightly wider and displaced somewhat to the right. This displacement is due in part to the fact, noted above, that log $\hat{\lambda}$ tends to overestimate log λ slightly—*e.g.*, when $\lambda = 1.50$, the mean of log $\hat{\lambda}$ is 0.184, the antilogarithm of which is 1.53—and a correction for this bias might be devised.

Of these three sets of intervals, those taken from the table of Matuszewski, *et al.*—whose paper provides an excellent introduction to the construction and interpretation of confidence intervals—are the only ones which are based on an exact mathematical solution. The mathematical approach which they have adopted, is, essentially, an extension of a short-cut proposed by Fisher (28) whereby the estimation is based solely upon the sum of the components of the code. Thus the codes 10–10–0, 10–9–1, $\cdots$, 10–7–3, $\cdots$, 10–5–5, 9–9–2, $\cdots$ yield the same single estimate by Fisher's short-cut method, and lead to the same confidence interval by the corresponding method of Matuszewski, *et al.*, whereas they lead to quite different values of $\hat{\lambda}$ in the table of Halvorson and Ziegler and thence to quite different confidence intervals from Miss Supińska's chart. The confidence intervals of Matuszewski, *et al.* are mathematically rigorous but they utilize only a portion of the information in the data, and hence are not to be recommended for accurate work. In the absence of Halvorson and Ziegler's table or Miss Supińska's chart, they will provide broad interval estimates of λ which will be correct in the long run, in at least 95 per cent of the instances in which they are used.

In two recent papers Halvorson and his associates (47, 80) considered the use of too frequent occurrence of rare codes for diagnosing the character of departures of laboratory technique from statistical control. Stein (88) gave a graphical means of checking the consistency of the results obtained at each of two dilutions differing by a factor of 10 when an equal number of tubes were run at each dilution, and his method can be extended to an equal number of tubes at each of three dilutions. Perhaps these two approaches can be combined to give a rapid graphical test of the statistical control of the experimental technique in terms of the consistency of the results of successive dilutions.

With regard to further research, it seems highly desirable to construct mathematically exact charts giving 0.95 and 0.99 confidence intervals for the bacterial density for the case of several tubes at a single dilution, and for the case of one or more tubes at each of the successive dilutions. For several tubes at one dilution, such charts can easily be constructed from Clopper and Pearson's (18) chart of confidence intervals for the binomial distribution, or from Fisher and Yates' table of fiducial limits for the binomial distribution.[23] When a sufficient number of exact frequency distributions of $\hat{\lambda}$, such as are given in Table 1 of Halvorson and Ziegler (51) have been computed, exact confidence charts patterned after Miss Supińska's empirical chart can be constructed for the case of 10 tubes at each of three successive dilutions in geometric progression with

[23] We understand that the second edition of Fisher and Yates (35) scheduled for 1942 publication includes such a table.

(ascending) factor of 10. Charts of this kind not only provide confidence limits for λ, given $\hat{\lambda}$, but for any particular (*e.g.*, maximum permissible) value of λ the charts can be used also to obtain control limits for $\hat{\lambda}$ such that, if $\hat{\lambda}$'s lying outside the limits are interpreted as danger signals, there will be a small known probability of an unwarranted "Wolf! Wolf!" While the limits corresponding to 10 tubes at each of the (three) successive dilutions with (ascending) dilution factor of 10 would probably be most widely used, the limits for 5 tubes at each dilution might be included on the same chart to show the effect of reducing the number of tubes on accuracy of estimation. Similar charts could be constructed for the McCrady-Swaroop system. A comparison of these charts would bring out at once some of the relative advantages and disadvantages of the two systems.

THE NORMAL DISTRIBUTION

It is evident that the graph of a binomial distribution consists of a series of bars of appropriate lengths erected at the points $x = 0, 1, 2, \cdots, n$, that they are always spaced at intervals of 1 unit whatever the value of n, and that the number of these bars increases as n increases. When n becomes only moderately large, say 25, evaluation of successive terms of the point binomial becomes laborious, and increasingly so as n increases. Furthermore, since the sum of the terms of $(q + p)^n$ is always unity for $0 < p < 1$ and $q = 1 - p$, the values of the respective terms, and hence the lengths of the corresponding bars, must decrease toward zero as n increases. In other words, whatever the value of p, the probability of the event occurring in any single trial, the probability of its occurring *exactly* x *times* in n independent trials will be practically zero when n is very large. Therefore, what is generally sought in practical work is *not* the probability that an event will occur exactly 600 times in 1000 independent trials, but the probability that it will occur between 525 and 650 times, or the probability that it will occur *at least* 600 times. In short, for large values of n, what is needed is a convenient method for summing the appropriate terms of a point binomial.

The normal probability curve, whose equation can be written

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad [8]$$

where m and σ are the parameters of the distribution, was developed by Abraham De Moivre (23) in 1733 as a means of finding easily the sum of consecutive terms of a binomial distribution when n is large. He showed that for large values of n the term involving p^x in the expansion of $(q + p)^n$, *i.e.*, the term which gives the probability of the event occurring *exactly* x times in n independent trials, is well *approximated* by equation 8 with $m = np$ and $\sigma^2 = npq$; and that the sum of the terms corresponding to values of x from X_1 to X_2 *inclusive* where $0 \leq X_1 \leq X_2 \leq n$ is given *approximately* by the area under the curve between $x = X_1 - 1/2$ to $x = X_2 + 1/2$ when $m = np$ and $\sigma = \sqrt{npq}$. For the approximation to be good, n should be large and $p - q$ should be numerically small

compared with $\sqrt{npq}$. Nevertheless, the approximation is fairly good even when n is as low as 10 and $p = 1/3$, as is shown in Table 7. The fit for $p = 1/2$ is much better, of course.

Until quite recently (76), this early work of De Moivre (23) had been overlooked, and the development of the normal distribution was often attributed to Gauss, who proposed its use as the distribution of errors of observation arising in astronomical work. Gauss discussed in detail the adjustment of measurements for errors of observation and laid the foundations of the method of least squares so thoroughly that many of the techniques employed today are essentially those of Gauss. Throughout his work the 'errors' were regarded as normally distributed random variables; equation 8 written in slightly different

TABLE 7

Comparison of binomial expansions with corresponding normal distribution curves

NUMBER OF "SUCCESSES"	$1000\ (\frac{1}{2}+\frac{1}{2})^{10}$		$1000\ (\frac{2}{3}+\frac{1}{3})^{10}$	
	Binomial expansion	Normal curve	Binomial expansion	Normal curve
0	1.0	2.2	17.3	28.7
1	9.7	11.2	86.7	80.6
2	44.0	43.5	195.1	178.8
3	117.2	114.4	260.1	256.5
4	205.1	204.7	227.6	238.6
5	246.0	248.0	136.6	143.8
6	205.1	204.7	56.9	56.2
7	117.2	114.4	16.3	14.3
8	44.0	43.5	3.1	2.3
9	9.7	11.2	0.3	0.2
10	1.0	2.2	0.0	0.0
	1,000.0	1,000.0	1,000.0	1,000.0

$1000\ (\frac{1}{2}+\frac{1}{2})^{10}$ might represent the distribution of the number of heads obtained per throw if 10 coins are tossed 1000 times; the "binomial expansion" gives the theoretical values obtained by expanding the binomial; the "normal curve" represents those obtained from the normal curve corresponding to this binomial, *i.e.*, where $m = np = 5$ and $\sigma = \sqrt{npq} = \sqrt{10/4}$.

Similarly $1000\ (\frac{2}{3}+\frac{1}{3})^{10}$ could represent the distribution of 'successes' when 10 dice are thrown 1000 times and a 'success' consists of *either ace or six* appearing on a die.

form, was referred to as the *law of errors* or the *error function*, names which it still retains in the physical sciences. Because of its widespread use in the theory of errors, the normal distribution is often called the *Gaussian Distribution*. Laplace developed the normal distribution as an approximation to the binomial distribution when n is large, apparently unaware of De Moivre's earlier development, and applied it to a variety of phenomena, especially to games of chance and vital statistics. In his analytical treatise on probability, Laplace (59) laid the foundations of modern mathematical statistics.

Adolphe Quetelet (1796–1874) led the way in applying the normal curve to biological and social phenomena, and Francis Galton (1822–1911) applied it to biological variables of every sort. Both were impressed by the way their data seemed to conform to this curve. There appears to have been a belief among many biologists of this period that this curve was an ideal to which

most biological distributions ought to conform, and that some explanation was needed when they did not. In consequence, the expressions "normal law" and "normal distribution" took root as substitutes for "law of error", "Gaussian distribution", *etc.* However, as more and more data were studied, and better methods of comparison were developed, it became evident, principally through the work of Karl Pearson (1857–1936), that the normal distribution is not a universal law of nature.

The practical importance of the distribution has not declined, however, on this account, but has actually increased in importance in recent years. The principal reasons for this are:

(a) Even though a population departs radically from normality, a secondary population formed of the arithmetic means of sufficiently large random samples drawn from it can be regarded without sensible error as normally distributed about the mean of the parent population with a variance equal to that of the parent population divided by the size of the sample. Hypothetical populations can be devised for which the preceding statement is false, but it may be considered true for biological populations, since biological variables take on finite values only. How large the samples need to be depends largely on the asymmetry of the population; for symmetrical and only moderately asymmetrical *continuous* distributions, samples of 4 are often large enough and samples of 10 are quite adequate, but with very skewed distributions somewhat larger samples may be needed.

(b) The normal distribution has many mathematical properties which make it particularly attractive in the development of statistical theory and techniques; hence most new techniques are developed on the supposition that the underlying variables are normally distributed.

(c) The assumption of an underlying normal distribution does not lead to serious error, since many such techniques are based on statistics and test criteria whose sampling distributions are relatively stable for moderate departure of the true underlying distribution from normality, especially if the sample is large.

Since different normal distributions are obtained by assigning different values of m and σ in equation 8, it is advantageous for purposes of tabulating properties (*e.g.*, the values of y corresponding to a given value of x, the area under the curve between two points, *etc.*) to standardize this family of curves by reducing all to a single form. This is accomplished by substituting for x a new variable defined by

$$u = \frac{x - m}{\sigma} \tag{9}$$

which reduces equation 8 to

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{10}$$

The area under the curve of equation 8 between $x = X'$ and $x = X''$ is equal to the area under the curve of equation 10 between $u = u'$ and $u = u''$, where

$u' = (X' - m)/\sigma$ and $u'' = (X'' - m)/\sigma$. When x is normally distributed, u is also normally distributed and has a mean of zero and a variance of one; a random variable having these properties is said to be a *normal deviate with unit variance*. Tests of significance based on the normal distribution are discussed and illustrated in Part II.

PART II. TESTS OF SIGNIFICANCE

As an introduction to the general topic of tests of significance let us consider a simple example: In cases of disputed parentage where each of the alleged parents has clear blue eyes, the eye-color of the child can form the basis of a test of the hypothesis that the claimants really are the child's natural parents. Since blue eye-color is a recessive trait, the mating of two individuals each with clear blue eyes cannot result in a brown-eyed child except through a mutation of the relevant gene, the probability of which is less than 1/10,000. If, therefore, whenever each of the 'parents' has clear blue eyes, their claim is rejected for a brown-eyed child, and accepted tentatively for a blue-eyed one, then this procedure constitutes a test of the hypothesis that these claimants actually are the child's natural parents. Several features of this test of significance should be noted: (a) the probability of falsely rejecting the hypothesis tested, which Neyman and Pearson (71, 72) term committing "an error of the first kind", is low, here less than 1/10,000; (b) in a region where brown eyes are not uncommon, the probability of a brown-eyed child from a random mating—we may assume that eye-color is not generally influential in mating—will be considerably larger than 1/10,000, so that the probability of rejecting the hypothesis by the above procedure will be considerably larger when it is false than when it is true, obviously a desirable property; (c) tentative acceptance of the hypothesis when it is false, which Neyman and Pearson term committing "an error of the second kind", will occur, however, whenever the child *and* each of the 'parents' has clear blue eyes although the child was not from the mating of these 'parents'—the probability of such an occurrence would not be negligible in a community in which blue eyes were not uncommon, and would be large in an immoral community in which blue eyes were quite common. On account of the sizable risk of an error of the second kind, one would not rely solely on this test.

These features are common to tests of significance in general. The probability of false rejection of the hypothesis, which is termed the *level of significance* of the test, is usually chosen in advance by the research worker in testing statistical hypotheses, thereby determining which outcomes will lead to rejection of the hypothesis tested and which will lead to tentative acceptance. The probability of the test rejecting the hypothesis tested when some particular alternative is true is termed the *power* of the test relative to this alternative (73). For the test to be critical with respect to an alternative to the hypothesis tested, the power with respect to this alternative must exceed the level of significance, the greater the excess the better the test relative to this alternative. A discussion of the power of the customary tests of significance is outside the scope of this paper, but it may be stated categorically for the tests to be discussed herein-

after that the power of these tests relative to a fixed alternative increases rapidly with the number of observations employed.

DATA FROM NORMAL DISTRIBUTION

To illustrate the use of the normal distribution in tests of significance let us consider the following data which represent milligrams of yeast produced in two media.

	Medium A	*Medium B*
	100	85
	110	96
	85	74
	90	80
	93	83
Total	478	418
Average	95.6	83.6

Example 1: Suppose past experience has shown that an individual estimate of yeast growth by the method used has a standard deviation of 10 milligrams and that on the average a yield of 90 milligrams is obtained in a certain standard medium. We ask then: Are the observed values consistent with the hypothesis that medium A is equivalent to the standard medium? That is, are the observed values in medium A independently and normally distributed about an expected value (m) of 90 with a standard deviation (σ) of 10. The admissible alternative is that the observed values are independently and normally distributed with a σ of 10 about a mean different from 90.

To test this hypothesis the normal deviate

$$[11] \qquad x = (\bar{X} - m)/(\sigma/\sqrt{N}) = (95.6 - 90.0)/(10/\sqrt{5}) = 1.25$$

is calculated from the data, N and $\bar{X}$ denoting the number of and average value of the observations, respectively. The deviate found in the present instance being numerically less than the 5 per cent significance level,[24] 1.96, the hypothesis may be accepted tentatively, that is, we tentatively conclude that medium A is equivalent to the standard medium for propagation of yeast.

Since in this example the number of observations involved is very small, the test has little discriminating power with respect to nearby values of m. It may be shown, for instance, that for $N = 5$ the above test conducted at the 5 per cent level of significance has slightly better than a 0.50 chance of rejecting $m = 90$ when in fact $m = 80$ or $m = 100$.[25] On the other hand, for $N = 16$ and

[24] See, for instance, Fisher (30) table 1.

[25] It may be noted here that, if the test were conducted at the 1 per cent level of significance, the probability of rejecting $m = 90$ when $m = 80$ or $m = 100$ would be only 0.27, so that the probability of an error of the second kind would be 0.73 in these instances. This illustrates a general principle: reduction of the probability of an error of the first kind by choosing a more stringent level of significance increases the probability of an error of the second kind.

$\sigma = 10$ the probability that the test will reject $m = 90$ when in fact $m = 80$ or $m = 100$ is approximately 0.98 when the test is conducted at the 5 per cent level of significance. It follows from this that with very small numbers of observations the chances of committing an error of the second kind are generally great. Otherwise stated, if 50 or more observations had yielded the above deviate, 1.25, we should have greater confidence that m was actually close to 90.

Example 2: Next, consider the question whether media A and B are equivalent. To do this, we test the hypothesis that the data for the two media represent independent random samples from normal populations with standard deviation $\sigma_A = \sigma_B = 10$, and equal means $m_A = m_B$, the admissible alternatives being that the data represent independent random samples from normal populations with $\sigma_A = \sigma_B = 10$ but unequal means $m_A \neq m_B$.

To test this hypothesis the normal deviate

$$[12] \quad x = (\bar{X}_1 - \bar{X}_2)/\sigma(1/N_1 + 1/N_2)^{\frac{1}{2}} = (95.6 - 83.6)/10(2/5)^{\frac{1}{2}} = 1.90$$

is calculated from the data. N_1 and N_2 are the number of observations from A and B, respectively, $\bar{X}_1$ and $\bar{X}_2$, the corresponding averages, and σ, the postulated common standard deviation. The deviate obtained in the present instance is close to, but does not exceed the 5 per cent level, 1.96, so that at this significance level the hypothesis tested should be tentatively accepted, and we might conclude that the two media are essentially equivalent. In practice, however, one would be reluctant to place great confidence in the conclusion that the data represent independent random samples from identical normal distributions. First, as already noted, the chances of an error of the second kind are great with so few observations. Second, the value obtained for the deviate is highly dependent on the value taken for $\sigma_A = \sigma_B$. With the new media, for example, it might be that $\sigma_A = \sigma_B = 8$, and this would lead to a deviate of 2.37. Since this value exceeds the 5 per cent significance level by a good margin, we should reject that portion of the hypothesis which states that $m_A = m_B$, had we postulated a standard deviation of 8.

Example 3: Although we might not place much confidence in the assumption that both σ_A and σ_B are equal to the σ observed with the standard medium, theoretical considerations might suggest the less restrictive assumption that σ_A and σ_B are equal. If so, "Student's" t test can be used. Suppose we ask the same question as in Example 2 but make no assumptions about the true value of $\sigma_A = \sigma_B$. Formally, we state this: Test the hypothesis that the data for media A and B constitute independent random samples from identical normal populations, the admissible alternatives being that the two parent populations are normal with identical standard deviations but different means.

We calculate the statistic,

$$[13] \quad t = (\bar{X}_1 - \bar{X}_2)/s(1/N_1 + 1/N_2)^{\frac{1}{2}}$$

where

$$[14] \quad s^2 = [\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2]/(N_1 + N_2 - 2).$$

X_1 and X_2 represent individual observations from A and B, respectively; the other symbols have the same meanings as in equation 12. This test, developed by Fisher (31), is an extension of a result obtained earlier by "Student" (90). The probability distribution of t depends on its degrees of freedom, which here are $N_1 + N_2 - 2$; tables of the significance levels of the distribution are available (30, 35, 79, 84).

For purposes of calculation it is often convenient to evaluate the summations from

$$\Sigma(X - \bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/N$$

[15] $$\Sigma(X_1 - \bar{X}_1)^2 = 100^2 + 110^2 + 85^2 + 90^2 + 93^2 - 478^2/5 = 377.2$$

$$\Sigma(X_2 - \bar{X}_2)^2 = 261.2$$

therefore,

$$s^2 = (377.2 + 261.2)/8 = 79.80, \quad \text{and}$$
$$t = (95.6 - 83.6)/\sqrt{79.8(2/5)} = 2.12$$

This value does not exceed the 5 per cent point, which for 8 degrees of freedom is 2.306, so that at this level of significance the hypothesis tested may be accepted tentatively, *i.e.*, the two media are equivalent. The risk of an error of the second kind discussed in connection with Example 2 has equal relevance here.

Example 4: If the method of obtaining the data is such that there exists a pairing of the values, the foregoing tests should not be employed. Instead, one should base the test on the successive differences between the pairs, testing whether the true mean difference is zero. For example, if in the yeast experiment the values 100 and 85 were determined in one run, 110 and 96 at another and so on, we test the hypothesis that the *differences* may be regarded as independent random observations from a normal population with *zero* mean, the admissible alternatives being that the *differences* are independent random observations from a normal population with non-zero mean.

The appropriate t to use to test this hypothesis is

[16] $$t = \bar{d}/(s/\sqrt{N})$$

where

[17] $$d = X_1 - X_2, \qquad \bar{d} = (\Sigma d)/N = \bar{X}_1 - \bar{X}_2$$

[18] $$s^2 = \Sigma(d - \bar{d})^2/(N - 1),$$

and the degrees of freedom are $N - 1$.

In the present instance $t = 12/(2.3452/\sqrt{5}) = 11.44$ which greatly exceeds the 1 per cent level (4.604) for 4 degrees of freedom. Accordingly, if in the long run, one is willing to risk committing an error of the first kind less than 1 time in 100, the discrepancy between observation and hypothesis in the present instance may be regarded as sufficient to warrant rejection of the hypothesis tested.

The hypothesis is rejected as a whole and further consideration is necessary

before one particular aspect is blamed. In the present instance, we have focussed our attention primarily on the equivalence of the media for yeast propagation and would, accordingly, be inclined to accept the alternative that the true mean difference was not zero, *i.e.*, medium A is superior to medium B. By construction, the t test is most powerful with respect to this class of alternatives, so that this explanation is generally the one to be adopted. But other alternatives should not be disregarded, such as: (a) non-randomness of sampling; (b) lack of independence of the successive differences; (c) non-normality of the common population. Fortunately, as a test of the hypothesis that the true mean difference is zero, the t test is relatively insensitive to moderate departures from normality. The issues of randomness and independence can be taken care of in the design and conduct of the research. It is for this reason that it is always desirable to introduce an element of randomization in experimental design. Similar remarks apply to the rejection of the hypotheses discussed in Examples 1 to 3.

When more than two samples of the data are at hand, pair-wise comparisons among all possible pairs with t tests do not provide a satisfactory method of testing the hypothesis that they are all from the same normal population, with the alternatives that their true means may differ. *Analysis of variance* provides the tests which are the extensions of t tests appropriate to such cases.

ANALYSIS OF VARIANCE AND DESIGN OF EXPERIMENTS

In applied phases of agricultural research the experimenter must frequently overcome handicaps which quantitatively at least appear more formidable than the corresponding ones of the laboratory scientist. These include:

1. Heterogeneity of the experimental material—this arises for example, from wide differences in soil fertility or pronounced variation in stock animals.
2. Restriction of replication—because of the expense and other considerations, the size and number of experimental plots or replicate animals is definitely limited.
3. Length of experiment—an experiment usually lasts through a growing season or reproductive period so that one or two experiments a year is the most one can hope to make. Moreover, because of unfavorable weather, depredation of animals, or of other accidents, the entire experiment may be lost.

It is understandable then that the complex type of experiment early found favor with the agronomist and animal husbandryman—if experimentation is limited, as many variables as practical should be included in order to gain maximum information when the experiment succeeds. The meaning of data from such experiments, however, was very obscure, and it was not until Fisher and his group at the Rothamsted Experimental Station developed the methods for analysis of variance that any real basis for comprehensive interpretation became available.

Although the particular handicaps referred to may not bother the laboratory scientist, their counterpart is constantly with him in various guises, and it would be unfortunate if he overlooked the powerful tools now at his disposal for analysis of data merely because the tools were originally developed for field trials. The principles used are perfectly general

and are equally applicable to data from investigations on pig feeding, fertilizer treatment, or bacterial nutrition.

Unfortunately, the exposition of the subject has been confined almost entirely to examples taken from field plots, animal production, or plant and animal breeding studies. Consequently, scientists unacquainted with the particular terminology or type of problem of such investigations may obtain only a hazy idea of the real nature of the analysis and may be inclined to regard the entire matter as a somewhat mysterious hocus-pocus useful only in applied agriculture where uncontrollable factors and lack of replication prevent use of the precise techniques of the laboratory. Actually, the development of the analysis of variance has allowed the introduction of the complex experiment, or "factorial design" as it is nowadays called, into other branches of science rather than restricting it to applied studies where it was formerly tolerated only because nothing better was at hand. The advantage of the complex experiment is not only its efficiency, *i.e.*, its economy of time and expense for a given amount of information, but also its insistence on close attention to experimental design and its ability to bring out information through the 'interaction' terms which either cannot be obtained or only with great effort in the single factor type of study. *More often than not, the interactions are precisely the information wanted if the results are to have significance for conditions other than the carefully standardized ones of the experiment.* (See Chapter VII in Fisher's *Design of Experiments* (33) for further discussion of the technical advantage of factorial design.)

Fisher suggests, "... perhaps it is worth while stating an impression I have formed—that the analysis of variance, which may perhaps be called a statistical method, because the term is a very ambiguous one—is not a mathematical theorem, but rather a convenient method of arranging the arithmetic." (Discussion of Wishart's (118) paper.) Although this is probably a modest understatement, it is a point of view that should be kept in mind for an understanding of the steps taken in an analysis. These may be enumerated:

1. A *total* sum of squares is calculated by squaring the deviations of the respective observations from their common mean without any attention being paid to treatments, strains, or other factors.
2. This sum of squares is divided into its components, thus assigning to each factor in the experiment its proper share of the observed variation among the observations.
3. A variance estimate from each factor is determined by dividing its share of the sum of squares by the appropriate degree of freedom.
4. The observed inequalities between certain of the variance estimates are tested for significance.

The general nature of the analysis of variance can be appreciated by noting certain properties of an arbitrary set of numbers arranged in k groups with n numbers in each group:

		Observation number							Total	Mean
		1	2	3		j		n		
group	1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1j}	$\cdots$	x_{1n}	T_1	$\bar{x}_1$
	2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2j}	$\cdots$	x_{2n}	T_2	$\bar{x}_2$
		.....								
		.....								
	i	x_{i1}	x_{i2}	x_{i3}	$\cdots$	x_{ij}	$\cdots$	x_{in}	T_i	$\bar{x}_i$
		.....								
	k	x_{k1}	x_{k2}	x_{k3}	$\cdots$	x_{kj}	$\cdots$	x_{kn}	T_k	$\bar{x}_k$
									T	$\bar{x}$

If the mean of the entire group is $\bar{x}$, and $\bar{x}_i$ is the mean for the ith group, it can be shown that the following equation holds:

$$[19] \qquad \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2.$$

That is, the sum of the squares of the deviations of the individual numbers (x_{ij}) from their general mean can be divided into two parts: (1) the sum of the squares of the deviations of the numbers in each row from the mean of that row summed for all the rows; (2) the sum of the squares of the deviations of each row mean from the general mean multiplied by the number of items in the row. In our example the rows have an equal number, but this is not a necessary restriction.

So far nothing has been said about the nature of the numbers since equation 19 holds independent of what they represent. Now let us consider them to be results of an experiment in which k treatments have been used and each treatment has been represented n times. The first term on the right hand side when divided by $k(n - 1)$ provides an estimate of σ^2 from variation *within the treatments*; the second term divided by $(k - 1)$ provides an estimate of σ^2 from variation *between treatment means*. If the k samples (k sets of n determinations each) are from normal populations with equal variances, *i.e.*, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2 = \sigma^2$, then a test of whether the means of these populations are equal can be devised. For if the variance estimate from *between treatment means* is denoted by A and the variance estimate from *within treatments* by B, it can be demonstrated that A and B are both unbiased estimates of σ^2 when the true treatment means are identical. It remains only to find a method for testing whether any two estimates of a variance, both subject to sampling errors, are estimates of the same true variance, σ^2. This is done by defining a new variable, $F = A/B$, and determining its distribution when A and B are both estimates of the same variance. As would be expected, this distribution depends on the values of $(k - 1)$ and $k(n - 1)$, the respective degrees of freedom of A and B. From the distributions of F, one can determine the probability that a given value of F will be exceeded through chance fluctuations when A and B are estimates of the same variance. Such information is condensed in tables, so that an observed value of F can be readily tested for significance by consulting the table under the proper degrees of freedom (84).

A few examples of varying degrees of complexity will illustrate how the method is used. The following data were obtained in a nitrogen fixation experiment:

CULTURE	N IN MILLIGRAMS			T_i
A	61	36	71	168
B	39	46	42	127
C	17	28	34	79
D	22	15	43	80

The triplicate determinations are not paired in any way through design of experiment. Is there reason for concluding that the mean nitrogen contents of these cultures differ? Our 'null' hypothesis (see Appendix) is that the samples are from the same population. The question asked may be rephrased in statistical terminology: Do the independent estimates of the variances from *between* and *within* treatments differ significantly? To answer this, the following estimates of variance are calculated: The *total sum of squares* is determined from equation 15 with the summation extending over all N observations and with $T = \Sigma X$

$$\Sigma(X - \bar{X})^2 = \Sigma X^2 - \frac{T^2}{N} = 61^2 + 36^2 + 71^2 \cdots 22^2 + 15^2 + 43^2 - 454^2/12$$

$$= 3069.7$$

Then, the *sum of squares between treatments* (different cultures)

$$[20] \quad \sum_i n_i(\bar{X}_i - \bar{X})^2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$$= (168^2 + 127^2 + 79^2 + 80^2)/3 - 454^2/12 = 1821.7$$

Having calculated these sums of squares in this simple case, we can set up the analysis immediately since the sum of squares *within treatment* is the difference between the *total* and that *between treatments*.

VARIATION	D.F.	SUM OF SQUARES	MEAN SQUARE	F	5% POINT
Between treatments	3	1,821.7	607.2	3.89	4.07
Within treatment (error)	8	1,248	156.0		
Total	11	3,069.7			

From the table we find that with 3 and 8 degrees of freedom, F may be expected to be at least 4.07, 5 times in 100 through chance alone, when the samples are from populations having a common mean and variance. Therefore, the observed differences in the estimates of variance are not large enough to be significant at the 5 per cent level, and we may conclude that these data do not conflict with our null hypothesis: that the samples are from the same population. It should be observed that in this simple case, the analysis of variance is merely an extension of "Student's" t test for the difference between two means. Actually, each of the means could be tested against the others by this method, but by combining the data in an analysis of variance we obtain greater precision since our estimate of error is now based on 8 degrees of freedom instead of 4 as would be the case if the t test were used. This is one of the main advantages of the design suggested by the analysis of variance, *viz.*, that in a given experiment, more treatments with fewer replications can be used, because the estimate of error will be based on all the samples, not merely on those of the two particular means which are to be compared.

Consider now a slightly more complex type of experiment in which two factors are varied. The following data from Thorne, Neal, and Walker (100) summarize the respiratory quotients of different species of the root nodule bacteria growing in a basic medium in which the source of nitrogen was varied:

SOURCE OF NITROGEN	R. MELILOTI	R. TRIFOLII	R. LEGUMINOSARUM	R. JAPONICUM	R. PHASEOLI	MEAN	T_i
Sodium nitrate	1.21	1.15	1.03	1.03	1.12	1.108	5.54
Ammonium chloride	1.18	1.15	1.07	1.08	1.13	1.12	5.61
Asparagin	1.05	1.08	1.06	0.92	1.19	1.06	5.30
Yeast extract	0.94	0.90	0.94	0.78	1.02	0.916	4.58
Mean	1.095	1.07	1.025	0.952	1.115		21.03
T_j	4.38	4.28	4.10	3.81	4.46		

Total sum of squares

$$1.21^2 + 1.18^2 + \cdots + 1.19^2 + 1.02^2 - 21.03^2/20 = 0.234$$

Sum of squares between means of species

$$\frac{4.38^2 + 4.28^2 + 4.10^2 + 4.46^2 + 3.81^2}{4} - 21.03^2/20 = 0.067$$

Sum of squares between means of nitrogen source

$$\frac{5.54^2 + 5.61^2 + 5.30^2 + 5.48^2}{5} - 21.03^2/20 = 0.133$$

The remainder, $0.234 - 0.067 - 0.133 = 0.34$, is due to residual variation and will be used for estimate of experimental error. This use assumes that any contribution to the residual variation due to interaction of species and source of nitrogen is negligible in comparison with the portion due to error of experiment, so that practically all of the observed variance in this term may be ascribed to random experimental error. This assumption may of course not be true and could easily be tested by replicating the estimations of the respiratory quotients.[26] The analysis of variance is now easily set up:

VARIATION	*D.F.*	SUM OF SQUARES	MEAN SQUARE	*F*	1% POINT
Species	4	0.067	0.0168	5.94	5.41
Nitrogen source	3	0.133	0.0443	15.65	5.95
Residual	12	0.034	0.00283		

It is evident that the value of F for both species of organism and source of nitrogen are much larger than would be expected to occur by random sampling

[26] Thorne, Neal, and Walker (private communication) did use replicates, but the complete data are not included in their paper. For purposes of illustration we have used the data actually given there.

even once in 100 times. We conclude then that our original null hypothesis, that the data may be regarded as independent samples from the same normal population (in particular, that the differences in species and nitrogen source had no effect), is incorrect. Accordingly, we accept the alternative with regard to which the test is most powerful, namely, that the differences in species and in nitrogen source did affect the outcome. We may want to know now which of the treatment means differ from the others. As shown in the table, the estimated variance of a single determination is 0.00283, so that a difference between an arbitrary pair of means of *sources of nitrogen* would be tested by

TABLE 8

Logarithms of milligrams nitrogen fixed by varieties of Medicago sativa inoculated with strains of Rhizobium meliloti

STRAIN OF RHIZOBIUM	VARIETIES OF MEDICAGO SATIVA					
	Hairy Peruvian		Ladak		Grimm	
	Expt. I	Expt. II	Expt. I	Expt. II	Expt. I	Expt. II
101	0.92	1.53	0.36	0.83	1.38	1.43
	0.97	1.55	0.32	0.86	1.28	1.53
105	1.09	1.40	1.01	1.37	1.40	1.58
	0.94	1.51	0.74	1.27	1.34	1.42
107	1.34	1.48	1.27	1.48	1.32	1.49
	1.44	1.52	1.10	1.22	1.45	1.46
111	0.94	1.42	1.06	1.60	1.37	1.34
	1.22	1.12	1.04	1.56	1.43	1.34
115	1.02	1.39	1.25	1.18	1.19	1.29
	1.06	1.12	1.18	1.21	1.35	1.29
129	1.35	1.43	1.30	1.39	0.83	1.44
	1.41	1.61	1.11	1.30	0.73	1.54

dividing by an estimated standard error of $\sqrt{0.00283(2/5)} = 0.0336$ to get a t with 12 degrees of freedom. The 5% point for t with 12 degrees of freedom is 2.18, hence a difference of $(2.18)(0.0336) = 0.073$ between an arbitrary pair of *source* means may be regarded as significant at the 5 per cent level. Similarly, $(2.18)(0.0377) = 0.082$ constitutes the 5 per cent level of significance for differences between an arbitrary pair of *species* means.

Certain reservations about the use of such comparisons should be noted. When the difference between a particular pair of means is compared with the corresponding 'minimal significant difference', as the foregoing calculated differences are often called, the interpretation of the outcome depends upon whether the decision to compare these particular means was reached before or after an examination of the data. If before, then, when the observed difference exceeds the minimal significant difference, it may be regarded as indicating a

real, underlying difference. If after, then a difference exceeding the minimal significant difference should be regarded as merely pointing to a comparison which may warrant special attention in further research. Also, observe that among m means only $(m - 1)$ independent comparisons can be made.

The final example to be considered deals with a more complex type of analysis and illustrates the importance of the interaction terms. Burton and Wilson (14) investigated in greenhouse trials the nitrogen-fixing ability of three varieties of *Medicago sativa* L. when inoculated with six strains of *R. meliloti*. The experiments were repeated during different seasons of the year to determine whether this factor affected the results. The logarithms of the quantity of nitrogen fixed per pot of 10 plants for two experiments are summarized in table 8. Only representative calculations which show how the variances are estimated will be given. From table 8 the total sum of squares, $\Sigma(X - \bar{X})^2 = 5.0429$, is determined as has been already illustrated. Next, a new table is made similar to table 8 but in which the duplicate samples have been combined. The items in this second table will be referred to as X_2 ; we calculate the total sum of squares for this table by the following:

[21] $$\frac{\Sigma X_2^2}{2} - \frac{T^2}{N}$$

where Σ denotes the summation over all values of X_2 , and $T = \Sigma X_2$ = sum of all original observations. Note that the correction term, the square of the total of a table divided by N, will be the same for all steps in a given analysis. The sum of squares obtained by formula 21 is 4.7039, and the difference between this and the original total sum of squares, $5.0429 - 4.7039 = 0.3390$ is that due to 'error', since by combining the duplicates we have eliminated the variation due to this source. This error sum of squares will have 36 degrees of freedom because each of the 36 pairs of duplicates will contribute 1 degree of freedom. We now make another table in which the effect of experiment is disregarded by adding together corresponding items from the two experiments:

STRAIN OF BACTERIA	VARIETY OF HOST PLANT			
	Hairy Peruvian	Ladak	Grimm	T_i
101	4.97	2.37	5.62	12.96
105	4.94	4.39	5.74	15.07
107	5.78	5.07	5.72	16.57
111	4.70	5.26	5.48	15.44
115	4.59	4.82	5.12	14.53
129	5.80	5.10	4.54	15.44
T_j	30.78	27.01	32.22	90.01

Total sum of squares in this table is

$$\frac{\Sigma X_4^2}{4} - \frac{T^2}{72} = \frac{4.97^2 + 4.94^2 \cdots 5.12^2 + 4.54^2}{4} - \frac{90.01^2}{72} = 2.7054$$

in which X_4 indicates that each term is composed of 4 corresponding items of the original data and Σ here denotes the summation over all values of X_4. The total sum of squares in this table is made of three factors: *Strain of organism* = $\sum_{i=1}^{6} T_i^2/12 - T^2/72 = 0.6033$ with 5 degrees of freedom; *Variety of plant* = $\sum_{j=1}^{3} T_j^2/24 - T^2/72 = 0.6032$ with 2 degrees of freedom; and *Interaction* of these two factors, $V \times S = 2.7054 - 0.6033 - 0.6032 = 1.4989$. The interaction term will have 10 degrees of freedom (5×2), as can be verified by the fact that this table has a total of 17 degrees of freedom and 7 of these are used by the simple factors. To obtain the effect of *Experiment* and its interaction with variety of plant, another table is constructed in which the items are classified according to these two categories as:

	HAIRY PERUVIAN	LADAK	GRIMM	TOTAL
Experiment I	13.70	11.74	15.07	40.51 (T_1)
Experiment II	17.08	15.27	17.15	49.50 (T_2)
Difference	3.38	3.53	2.08	8.99

The total sum of squares of this table is composed of: that due to *Variety* (which has been already determined), that due to *Experiment*, and that due to *Interaction* of variety and experiment. These could be calculated in the usual way, but for tables of this type ($2 \times n$), a more rapid method is:

Sum of squares due to experiment

$$[22] \qquad (T_2 - T_1)^2/2N' = 8.99^2/72 = 1.1225$$

with one degree of freedom, where T_1 and T_2 denote the totals for Experiments I and II respectively, and where N' is the number of items represented by each total, in this case $N' = N/2$, where N is the total number of observations in table 8.

Sum of squares due to interaction

$$[23] \qquad \frac{\Sigma d^2}{2k'} - \frac{(T_2 - T_1)^2}{2N'} = \frac{3.38^2 + 3.53^2 + 2.08^2}{24} - 1.1225 = 0.0530$$

where d is the difference between the total for a given variety in Experiment II and in Experiment I, each such total being the sum of k' original observations, and Σ denotes summation over varieties. The entire table has 5 degrees of freedom, of which 2 belong to *Variety*, 1 to *Experiment*, and 2 to their *Interaction*.

A similar table is constructed in which the categories are *Strain of organism* and *Experiment*. The interaction term, *Strain* $\times$ *Experiment* is determined in exactly the same manner, giving 0.2735 with 5 degrees of freedom. The

sums of squares for the several factors and their first order interactions are now added and this sum, 4.3053, taken from the total sum of squares for the X_2 table, 4.7039. The difference, 0.3986, is the sum of squares belonging to the triple (second order) interaction, *Variety* × *Strain* × *Experiment*. This has 10 degrees of freedom (5 × 2 × 1), since the total for all treatments is 35, and 25 of these have been used by the simple factors and first order interactions. The analysis of variance can now be set out as follows:

VARIATION	*D.F.*	SUM OF SQUARES	MEAN SQUARE	*F*	5% POINT	1% POINT
Variety	2	0.6032	0.3016	32.02	3.27	5.26
Strain	5	0.6033	0.1207	12.81	2.48	3.59
Experiment	1	1.1225	1.1225	119.16	4.12	7.41
V × E	2	0.0530	0.0265	2.81	3.27	5.26
S × E	5	0.2735	0.0547	5.81	2.48	3.59
V × S	10	1.4989	0.1499	15.91	2.13	2.90
V × S × E	10	0.5495	0.0550	5.84	2.13	2.90
Error	36	0.3390	0.00942			
Total	71	5.0429				

The interpretation should give no difficulty. The values of F corresponding to the variance due to the single factors, *Variety* and *Strain*, exceed the 1% point indicating that at least one variety of the host plant was superior to the others independent of the strain of bacteria used as the inoculum, and that at least one strain of the bacteria was better than the others independent of the host plant. Obviously the experiment factor was significant since the design was such that more nitrogen would be fixed in one experiment than the other. Considering the first order interaction terms, it appears that the varieties responded identically in the two experiments, but the strains did not. The important interaction term is the *Variety* × *Strain*; the high value of F corresponding to it shows that the efficiency of a strain in fixing nitrogen varies with the host plant used—thus establishing "host plant specificity" for this plant-bacterial group (14). Of interest is that this interaction of variety and strain varied with the experiment as shown by the significant F value for the second order interaction term, V × S × E. The possible significance of this point (verified in further experiments) for symbiotic nitrogen fixation is discussed in the original paper. Once significance for a factor is established, some indication of the particular bacterial strains and varieties of host plant responsible can be obtained by calculating the appropriate error from the error of the single sample, 0.00942, and testing *means* or *totals* by the usual t formula.

This example illustrates several important aspects of the analysis of variance which should be considered in somewhat more detail before finishing our discussion.

1. *The equality of variances within treatments.* The requirement that all the observations should be of equal precision, *i.e.*, have the same standard devia-

tion, needs emphasis as it is sometimes ignored. Cochran (21) stresses the importance of this requirement and illustrates by examples how a mechanical application of analysis of variance to data of varying precision can lead to absurd conclusions. In the present experiment, the treatments, including time of experiment, caused a wide variation in the nitrogen content of the plants. The difference in nitrogen content, of course, is merely another way of stating that the plants differed in size. *A priori*, one would not suppose that *variation* among a population of plants of one size would equal that of another population in which the plants were 2 to 3 times as large; experience with plants under the conditions used in this experiment has verified this conclusion. Since under these conditions the growth of the plants is approximately logarithmic, it seems reasonable to assume that if the original data, total nitrogen in 10 plants, are transformed by taking the logarithm, the variances of the transformed data would not greatly differ among the different treatments. The logical transformation is not always so evident as in this case (53). Thus, percentage should be transformed to an angle, $\theta = \arcsin \sqrt{p}$, where p denotes an observed *proportion*. Employed originally by Fisher (29), the application of this transformation to experimental data has been promoted by Zubin (125), Bliss (5, 6), and Clark and Leonard (16). Of special importance to bacteriologists is the square root transformation advanced by Bartlett (1, 2) for use in connection with variables having Poisson distributions. If suspicion exists that the variances are unequal, some function of the observations in which equal variances might be expected should be taken before proceeding with the analysis. Bartlett (2) has shown how to test a set of variance estimates for homogeneity.

2. *Design of experiment.* It is essential that the various treatments have an equal chance of being exposed to those factors in the experiment which are not under control. This is a matter of *design of experiment*. Little more can be noted here than some of the methods available for insuring proper design for statistical treatment of the data; details are given in the monograph of Fisher (33) as well as in any modern statistical text which deals with agricultural experimentation. In the nitrogen fixation studies, only 36 pots were used in each experiment, as this number could be readily placed in a rather small space on the greenhouse bench. The position of a given pot in the area used was determined by chance (drawing of a card or use of a table of random numbers). Had the experiment required more pots, *e.g.*, 4 replicates instead of 2, so that the space occupied in the greenhouse would be rather extensive, the introduction of blocks would be advisable. The greenhouse bench would be divided into 4 blocks in such a way that the conditions (light, temperature, drafts) in a given block would be reasonably constant, then one pot of each particular treatment placed within each block, its exact position being determined by chance as before. In this case the replicates of a particular treatment are not interchangeable, but are grouped with the corresponding replicate of the other treatments in the same block. Thus the block becomes one of the 'treatments'. The sum of squares due to this factor is removed from the total in the statistical analysis,

which prevents the error variance from increasing because of differences in the environment of the different blocks.

This technique is extremely important in field trials in which soil heterogeneity is one of the chief sources of variance. By proper experimental design, the variance due to differences in soil fertility in different parts of the experimental plot can be estimated and allowed for in the analysis. This increases the permissible number of replicates, but imposes a limit on the number of treatments, namely that which can be placed within the relatively homogeneous block. An interesting extension of thus correcting for soil heterogeneity is the Latin square which can be used if the experiment is designed so that the number of replicates equals the number of treatments. As can be seen in the following example of a four-fold Latin square in which A, B, C, and D represent treat-

B	A	D	C
A	D	C	B
C	B	A	D
D	C	B	A

ments, a particular treatment appears exactly once in each row and column. In the analysis, sums of squares due to differences between rows and to differences between columns are segregated thus making allowance for gradients of soil fertility, *etc.*, in two directions within the field. Although the Latin square has been used primarily in field plot trials, obvious applications suggest themselves in other research areas, *e.g.*, in bacteriology, arrangement of cultures in an incubator so as to be able to make allowances for temperature gradient.

Another point concerned with experimental design illustrated by the final example deals with the estimate of the error. If single samples had been taken as was done in the preceding example, instead of duplicates, the row labelled *Error* in the analysis of variance table would be absent, and the V × S × E interaction would have to serve as the best available estimate of the experimental error. If this had been done, the mean squares due to strain differences and to interaction of strain with experiment would have been judged 'non-significant'. Although perhaps of secondary interest in this particular experiment, the resulting difference in interpretation emphasizes the desirability of replication and the danger of presuming that a certain interaction is non-existent in order to use the corresponding mean square as an estimate of experimental error.

3. *The error variance.* An analysis of variance will not generally be needed to evaluate the relative merits of treatments differing widely in effectiveness, and, indeed, will be unavailable in such cases if the observations corresponding to the respective treatments differ in precision. The refinements of analysis of

variance are needed principally when the differences between the treatments are slight, and, fortunately, in such cases the observations are generally of approximately equal precision so that the analysis of variance technique *is* available. When treatment differences are not great, efficient analysis will aid mateirally in this detection. Furthermore, this analysis of variance provides additional advantages. First, we obtain an estimate of experimental error, for comparing any two treatment means, which is based on all the observations in the experiment rather than one based just on the observations corresponding to the treatments whose means are being compared. Hence, the test of significance has greater power, *i.e.*, has a larger chance of detecting any real difference between the treatments. Second, factors which are known to affect the result, and which in classical experimentation are kept constant in order to make possible a determination of experimental error, may be varied within the limits imposed by the design employed and allowed for in the analysis, thereby giving reality to an experiment which otherwise might suffer from idealization possible only in the laboratory.

To illustrate these points, let us consider an experiment in which four media are to be compared for their ability to bring about some desired growth response in the commercial production of yeast. Suppose that laboratory facilities would allow 16 determinations to be made with each medium. If all media were kept in some carefully controlled laboratory environment, then the analysis of variance would be:

Test of 4 media in one environment

Variation	*Degrees of Freedom*
Between media	3
Within media (error)	60

The difference between any two media will here be compared with an estimate of experimental error based on 60 degrees of freedom, whereas a pair-wise comparison using only the data for the media compared would employ an estimate of error based on 30 degrees of freedom. Since such a large number of replications are involved, the increase in precision by use of analysis of variance is slight, as can be seen from the fact that the 5 per cent level of t for 30 degrees of freedom is 2.04, and for 60 degrees of freedom is 2.00. Using the pair-wise comparison as a standard, it is seen that by planning to use analysis of variance, the number of determinations to be made on each medium could be reduced to 8 to get comparable accuracy, since the 5 per cent significance level of t for $4 \times 7 = 28$ degrees of freedom is 2.05.

It would be impractical, however, to make such an experiment in which the media were kept in a carefully controlled laboratory environment, since it is well established that the relative suitability of a medium for an organism varies with factors in the environment such as temperature, aeration, size of inoculum. Suppose each of these factors was introduced into the experiment at two levels, then duplicates of each treatment would be possible in the 64 cultures since $2(4 \times 2 \times 2 \times 2) = 64$. The analysis of variance would then be:

Test of 4 media in different environments

Variations	*Degrees of Freedom*	*Variations*	*Degrees of Freedom*
Main Effects:		*2nd Order Interactions:*	
Between media	3	M × T × A	3
Between temperatures	1	M × T × I	3
Between aerations	1	M × A × I	3
Between inocula	1	T × A × I	1
1st Order Interactions:		*3rd Order Interaction:*	
M × T	3	M × T × A × I	1
M × A	3		
M × I	3	*Between Duplicates (Error)*	32
T × A	1		
T × I	1		
A × I	1		

The *actual* experimental error will be no larger in the complex experiment, but its estimate, being based on 32 degrees of freedom instead of 60, will have lost some precision, but it is as precisely determined as would have been the case in a pair-wise analysis with 16 determinations for each medium in a single environment. This loss, however, is more than compensated for by the greatly increased information obtained concerning the influence of temperature, aeration, and size of inoculum on the response of yeast in different media together with the various interactions. The paper of Brandt (9) may be consulted for the working of an actual problem such as the one outlined.

An interesting example of the employment of analysis of variance in the statistical control of a laboratory technique is given by James and Sutherland (55, 56, 57) in their studies on the accuracy of the plate count in enumerating soil microörganisms. They investigated a number of factors which might be expected to affect the counts, such as aliquot of soil taken, method of dilution, of pouring plates, and of incubating. Their analysis indicated that both aliquot taken and dilution[27] were important in affecting the count, but the other details of technique investigated had less influence. Though they did not feel justified in making definite recommendations on the basis of their findings, they suggested that if one is limited to a certain number of plates, a more accurate estimate will be obtained if the number of aliquots and dilutions is increased at the expense of replication of a single dilution. The fact that different dilutions frequently gave rise to different estimates was considered, and a method for correcting the estimates so as to be interchangeable was suggested (56).

REGRESSION AND CORRELATION

A primary function of all research is to determine *relationships* between two or more quantities. To know *how* phenomena are related is essential to all scien-

[27] To avoid confusion in our discussion of the statistical control of plate counts, we have disregarded errors arising from technique as contrasted with the sampling error. In actual practice, however, it is recognized that errors in dilution, *etc.*, may become just as important in affecting the reliability of a count as the other. Jennison and Wadsworth (58) discuss this aspect of variation and have furnished a table for correction of dilution errors for various deviations in pipette and dilution blanks.

tists from the theorist who integrates the relationships into hypotheses regarding the *why* of nature to the technician who wants to know the value of one quantity from observations on another. Probably the most obvious way to judge the relationship between two variables[28] is by plotting the values on graph paper and noting the trend. Usually an effort is made to obtain a linear relationship as this simplifies interpretation and use. If the original data do not yield a straight line, its trend often suggests the proper function, and by suitable transformations a linear relationship can be secured. Having plotted the data so that they appear to be linear, the investigator may merely draw in the line which to his eye appears to fit them. Although this is probably not objectionable if the fit of the observations to the line is very close, it introduces an undesirable subjective element and an opportunity for bias. Sometimes the fit is fictitiously good, because of the scale used for plotting (37), and disagreement regarding interpretation ensues.

Statistical theory points out that the criterion of goodness of fit which should be adopted in a given instance depends upon the nature of the random variation affecting the variables, and provides objective methods of fitting which lead to the best fitting line as judged by the appropriate criterion. In biological research, methods for obtaining a relation between two variables which it is hoped will be sufficiently close to the true relationship for the purposes in mind involve the following steps[29]:

(a) It is assumed that the pairs of observational points, $(x_1y_1) \cdots (x_ny_n)$, differ from the 'true' points as a result of biological variation and errors of measurement in either x or y or both. 'True' is used in the sense that for a fixed value of x observed values of y will be randomly distributed about a central value, called the 'true' value, the exact nature of which depends upon the character of the distribution. When the distribution is normal, the *mean* is termed the 'true' value.

(b) From theoretical considerations or from the appearance of the graph some mathematical relationship between x and y is assumed. We shall restrict ourselves to consideration of the linear type:

$Y = \alpha + \beta x$ and $X = \gamma + \delta y$, in which the manner of conducting the experiment usually determines which is to be the independent variable (26).

(c) Estimates of the constants, *e.g.*, a and b for α and β, are chosen which will make the resulting line 'best' fit the observations. When the random variation is normal with the same standard deviation throughout the range of x and y considered, the fit is 'best' when the sum of the squares of the deviations of the observations from the line chosen is a minimum. We shall restrict the present discussion to the case where normal random variation is manifest in the observed

[28] In this paper attention will be confined to simple regression coefficients since the extension of the methods to problems involving 3 or more variables can be found in any of the standard references given in the bibliography.

[29] A fuller discussion of the concepts and principles involved has been given by Eisenhart (26).

values of y, but the values of x are exactly determined, so that an observed point can deviate from the 'true' point in the y direction only.

(d) Tests of significance are then carried out to determine how good the fit actually is; the outcome of these will be the basis for deciding whether the chosen function can adequately describe the actual relation between the x's and y's.

The methods for carrying out these steps can best be followed by working through a specific example. Table 9 provides data on the fixation of nitrogen by inoculated red clover plants kept in an atmosphere containing H_2. When

TABLE 9

*Fixation of nitrogen by inoculated clover plants in presence of H_2**

ARRAY	TIME IN DAYS (x)	LOG MG N FIXED (y)	SUM OF y ARRAYS (T_i)
1	0	0.520	
		0.546	1.678
		0.612	
2	17	0.843	
		0.844	2.522
		0.835	
3	31	1.090	
		1.189	3.478
		1.199	
4	50	1.484	
		1.559	4.539
		1.496	
T_x.....	294		
T_y.....			12.217

* $pN_2 = 0.15$; $pO_2 = 0.2$; $pH_2 = 0.15$; pHe = 0.5 atm.; $n_i = 3$ in all arrays.

the original data were plotted, a logarithmic function was suggested which was verified by plotting *log mg N* against *time in days*. It should be noted that in fitting these data it is clear that *time in days* constitutes the independent variable, since its values are determined by the will of the experimenter, and, provided the latter can count accurately, its values are not subject to experimental error. The postulated relationship is:

[24] $$Y = \alpha + \beta X$$

where Y = mg N fixed, X = time in days.
The constants in this equation are estimated by:

[25] $$a = \bar{y} - b\bar{x}, \text{ and}$$

$$[26] \qquad b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma x(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

where $\bar{x}$ and $\bar{y}$ denote the mean values of the 'observed' values of x and y, respectively, and Σ denotes summation of all pairs of observations. In actual calculation, use is made of the following identity where N = number of pairs of observations:

$$[27] \qquad \Sigma(x - \bar{x})(y - \bar{y}) = \Sigma xy - (\Sigma x)(\Sigma y)/N,$$

$= 0(0.520 + 0.546 + 0.612) + 17(0.843 + 0.844 + 0.835) \cdots$ etc. $- (294)(12.217)/12 = 377.642 - 299.316 = 78.326$

From equation [15],

$\Sigma(x - \bar{x})^2 = 3(0^2 + 17^2 + 31^2 + 50^2) - 294^2/12 = 11250 - 7203 = 4047$

hence, $b = 78.326/4047 = 0.01935$

$a = \bar{y} - b\bar{x} = 12.217/12 - 0.01935(294/12) = 0.544$, and the 'best' line is: $Y = 0.544 + 0.01935X$.

For testing the significance of these constants, the total sum of squares $\Sigma(y - \bar{y})^2$, must be divided into its several components appropriate to the test, and an analysis of variance made. By equation 15, $\Sigma(y - \bar{y})^2 = 0.520^2 + 0.546^2 + \cdots 1.559^2 + 1.496^2 - 12.217^2/12 = 1.53552$. For testing the significance of regression, which means testing whether b is significantly different from zero, $\Sigma(y - \bar{y})^2$ is divided into two parts: first, that due to regression, *i.e.* deviations of points on the regression line from the mean of the y's, which equals

$$[28] \qquad \Sigma(Y - \bar{y})^2 = b^2\Sigma(x - \bar{x})^2 = (0.01935)^2(4047) = 1.5153,$$

and the remainder which accounts for the deviations of the observed points from the regression function, $\Sigma(y - Y)^2 = 1.5355 - 1.5153 = 0.0202$.

The objection may well be raised that testing for regression in this example is unnecessary and somewhat artificial, since it is obvious from mere inspection of the raw data that the quantity of nitrogen fixed increases with time. Although it is true that in this particular instance it is superfluous to test whether b differs from zero, this is not always so, and indeed many times, it is the most significant test made. Hence, in order to illustrate the method and also, because some of the values will be needed in the test for linearity, we include this step. The analysis is:

VARIATION	SUM OF SQUARES	*D.F.*	MEAN SQUARE	*F*	5% POINT
Regression	1.5153	1	1.5153	750.1	4.96
Deviations from regression	0.0202	10	0.00202		

As would be expected F is well beyond the 5% point.

Although the foregoing analysis confirms the existence of regression as measured by a linear function, it does not indicate the adequacy of a straight line to represent the actual relationship. To test this, we divide the total sum of squares into two parts, one of which represents that portion due to the means

of the arrays differing from $\bar{y}$, (and therefore to differences between the arrays) and the other, that portion due to the scatter of values within the several arrays.

$$[29] \qquad \underset{\text{Total}}{\Sigma(y - \bar{y})^2} = \underset{\substack{\text{Between} \\ \text{arrays}}}{\Sigma n_i(\bar{y}_i - \bar{y})^2} + \underset{\substack{\text{Within} \\ \text{arrays}}}{\Sigma\Sigma(y - \bar{y}_i)^2}$$

The total degrees of freedom is $(n - 1)$ of which $(q - 1)$ belongs to the *Between Arrays* and $(n - q)$ to the *Within Arrays*. In the foregoing equation n_i denotes the number of samples in the ith array with mean $\bar{y}_i$, with $\Sigma n_i = n$, and with q number of arrays. The double summation signs indicate that the sum of squares for the deviations within arrays is to be summed for all q arrays. The sum of squares due to differences *Between Arrays* is further broken up into a part due to deviations of means of arrays from the regression line with $(q - 2)$ d.f. and a part due to the regression itself (1 d.f.):

$$[30] \qquad \Sigma n_i(\bar{y}_i - \bar{y})^2 = \Sigma n_i(\bar{y}_i - Y)^2 + b^2\Sigma(x - \bar{x})^2$$

The linearity of regression is tested by comparing the mean square corresponding to deviations of means of arrays from the regression line with the mean square corresponding to *Within Arrays*. Since

$$[31] \qquad \begin{aligned} \Sigma n_i(\bar{y}_i - \bar{y})^2 &= \Sigma T_i^2/n_i - T_y^2/N \\ &= \frac{1.678^2 + 2.522^2 + 3.478^2 + 4.539^2}{3} - \frac{12.217^2}{12} = 1.5205 \end{aligned}$$

then $\Sigma n_i(\bar{y}_i - Y)^2 = 1.5205 - 1.5153 = 0.0052$
and $\Sigma\Sigma(y - \bar{y}_i)^2 = 1.5355 - 1.5205 = 0.0150$.
The analysis of variance accordingly is:

VARIATION	SUM OF SQUARES	D.F.	MEAN SQUARE	F	5% POINT
Deviations of means of arrays from line.	0.0052	2	0.0026	1.37	4.46
Within arrays.....	0.0150	8	0.0019		

As the value of F is definitely less than that of the 5 per cent point, it is concluded there is no evidence of departure from linearity.

When a value of b has been determined in each of two independent experiments, an extension of the t test can be used to determine whether the two values of b differ significantly:

$$[32] \qquad t = \frac{(b_1 - b_2)}{\left[\dfrac{(N_1 - 2)s_1^2 + (N_2 - 2)s_2^2}{N_1 + N_2 - 4}\{1/\Sigma(x_1 - \bar{x}_1)^2 + 1/\Sigma(x_2 - \bar{x}_2)^2\}\right]^{\frac{1}{2}}}$$

in which

$$[33] \qquad (N - 2)s^2 = \Sigma(y - \bar{y})^2 - b^2\Sigma(x - \bar{x})^2$$

and t has $(N_1 + N_2 - 4)$ d.f. This formula should be used only when s_1^2 and s_2^2 do not differ significantly, otherwise the u or v tests of Welch (105) are available. The necessary data for testing estimates of a and b obtained in some experiments with clover plants grown in atmospheres with different partial pressures of H_2 follow:

pH_2	a	b	N	$(N-2)s^2$	$\Sigma(x-\bar{x})^2$	$\bar{x}$
0.15	0.544	0.01935	12	0.0202	4047	24.50
0.35	0.565	0.01606	9	0.0171	3231	21.77

The first step is to compare the variances by calculating $F = (0.0171/7)/(0.0202/10) = 1.21$, which is well below the 5 per cent point of 3.15 for 7 and 10 d.f. As the variances appear to be homogeneous, the t test may be applied:

$$t = \frac{0.01935 - 0.01606}{\left[\dfrac{0.0202 + 0.0171}{12 + 9 - 4}\left\{\dfrac{1}{4047} + \dfrac{1}{3231}\right\}\right]^{\frac{1}{2}}} = 3.21 \text{ with } 17 \text{ d.f.}$$

Since this value of t exceeds the 1 per cent point, it appears that the observed difference in the b's is significant.

The difference between the a's may be tested by means of the following formula when s_1^2 and s_2^2 do not differ significantly.

$$[34] \qquad t = \frac{a_1 - a_2}{\left[\dfrac{(N_1 - 2)s_1^2 + (N_2 - 2)s_2^2}{N_1 + N_2 - 4}\left\{\dfrac{1}{N_1 - 2} + \dfrac{\bar{x}_1^2}{\Sigma(x_1 - \bar{x}_1)^2} + \dfrac{1}{N_2 - 2} + \dfrac{\bar{x}_2^2}{\Sigma(x_2 - \bar{x}_2)^2}\right\}\right]^{\frac{1}{2}}}$$

$$= \frac{0.565 - 0.544}{\left[\dfrac{0.0171 + 0.0202}{9 + 12 - 4}\left\{\dfrac{1}{7} + \dfrac{21.77^2}{3231} + \dfrac{1}{10} + \dfrac{24.50^2}{4047}\right\}\right]^{\frac{1}{2}}} = 0.7$$

The difference between the a's is clearly not significant, which means, of course, that log mg. nitrogen at $t = 0$ was the same in each series, *i.e.*, the two tests started together.

Regression statistics have numerous applications in bacteriology and allied fields several of which will be considered briefly.

Calculation of k values.. In bacteriological research it is often advantageous to use the *rate* of growth (rate of respiration, rate of nitrogen fixation, *etc.*,) rather than *total* growth (total respiration, total nitrogen fixed, *etc.*). Since these functions in many instances increase logarithmically with time, the traditional measure of growth rate is the k value defined: $k = (1/t) \ln [(a + y)/a]$ in which a represents the growth at $t = 0$, and y the increase after time t. If log $(a + y)$ is plotted against t, the slope of the resulting line multiplied by 2.303 will estimate k. Thus, in the example just discussed, the specific rate constant of nitro-

gen fixation for a pH_2 of 0.15 atm. is: 0.01935 × 2.303 = 0.0445. Not only can the best estimate of k be made but also the significance of observed differences in k values can be determined by testing the b's from which they were derived. The papers of Wilson and his associates furnish several examples of this use of regression coefficients (60, 113, 116, 120, 121).

Estimation of k values by this method is greatly facilitated if experiments are planned so that the calculations are reduced. For example, nitrogen fixation by *Azotobacter* can be estimated indirectly in a Warburg microrespirometer by measuring the increase in rate of respiration with time. For routine determinations in our laboratory, a standard method has been adopted in which five readings are taken hourly. Under these conditions, both Σx and $\Sigma(x - \bar{x})^2$ equal 10; N is 5 and $\bar{x}$ is 2. Reference to formulae 25 and 26 shows that both a and b can be rapidly calculated since only Σy and Σxy must be evaluated, the other terms being determined mentally. Calculation of the error in b requires somewhat more effort, but if a calculating machine is available, all the statistics can be determined *without bias* as rapidly as b is estimated by the usual graphical procedure.

After a period of time, it may not be necessary to determine the error variance, s^2, except for a check. We have calculated variances from over 100 trials and shown them to be 'homogeneous', *i.e.*, all belong to the same population, so that their mean, $\bar{s}^2$, provides a reliable estimate of the expected variance under the prescribed conditions.[30] From $\bar{s}^2$ and $\Sigma(x - \bar{x})^2 = 10$, it was determined that two k's must differ by 13 to 17 per cent in order to be significant. Thus in routine work under the 'standard' conditions, only the b's are determined; from these the k's are calculated and compared. Occasionally, the variance, s^2, is estimated to make certain that the method is under statistical control. If significantly excessive (or deficient) variances are obtained, the technique is examined for possible interfering factors. Not only does this serve as a red light for trouble, but it provides an objective test for the learning process in new students. Before he starts his research, a beginner makes several practice runs; he continues these until the error in his b's is commensurate with that established by competent workers.

Test of a hypothesis. Frequently a hypothesis under test dictates the method of plotting the data and usually by mathematical manipulation the suggested function can be put in linear form. This is illustrated by data from Wilson, Burris and Lind (115) in figure 6. The hypothesis is that the initial steps in nitrogen fixation by *Azotobacter* can be represented by the enzyme mechanism first formulated by Michaelis and Menten (27). Lineweaver and Burk (61) showed that if the hypothesis is correct a straight line should result when the reciprocal of the rate of fixation ($1/k$) is plotted against the reciprocal of the partial pressure of nitrogen ($1/pN_2$). Likewise, if at a given pN_2 the velocity *relative* to the maximum velocity is k^*, then pN_2/k^* should be a linear function of pN_2. The figure shows that straight lines were obtained by both methods of plotting. The fit of the points to the line at the left does not appear to be so good, but this results primarily from the scale used. Statistical tests demonstrated that a straight line was satisfactory in both instances.

Another test of this hypothesis based on regression statistics was also described in the same paper (115). If the hypothesis holds, the rate of fixation should be less at a pN_2 of 0.2 atm. than at 0.8 atm. Calculation showed that the

[30] The test for homogeneity among several estimated variances constitutes another application of the X^2 distribution (see p. 122 ff.). The details of this test can be obtained in a textbook on statistics, *e.g.*, Rider's (79, p. 102), or from the original (2).

error in k would mask the expected difference in an *individual* experiment. But, by combining the results from a number of experiments the expected difference was demonstrated. Moreover, by increasing the number of observations in a single trial, the error was decreased so that significant differences were detected in the individual experiments. Wilson and his collaborators (116) have also used regression statistics for determining whether an inhibitor of an enzyme system acts competitively or non-competitively.

Estimation of μ values. In enzyme studies, definition of the physical-chemical characteristics of the system investigated is frequently useful—for example, to determine if the enzyme system in one organism is reasonably similar to a corre-

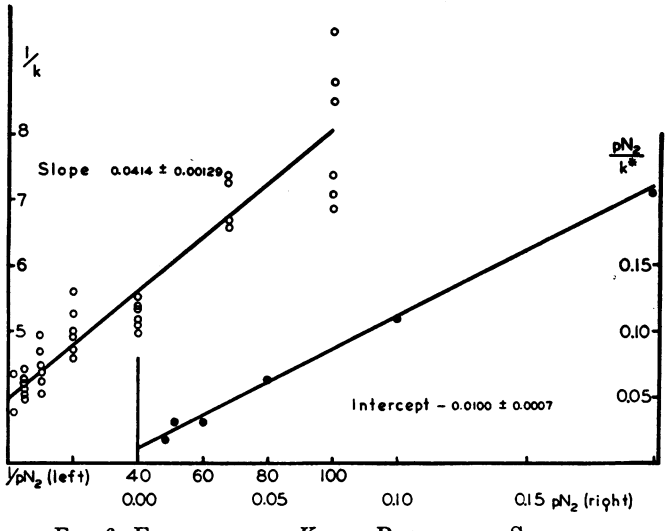

FIG. 6. ESTIMATION OF K_{N_2} BY REGRESSION STATISTICS

On the hypothesis that the enzyme mechanism is the Michaelis-Menten type (27), a straight line should result when the reciprocal of the rate of reaction, $1/k$, is plotted against the reciprocal of the substrate concentration, $1/pN_2$ (left). Likewise, if pN_2 is divided by the rate of reaction relative to the maximum rate, under the conditions of the experiment, and this value, pN_2/k^*, is plotted against pN_2, a linear relationship obtains (right). Data are from Wilson, Burris and Lind (115).

sponding system in another (27, chap. X). One important characteristic is the response of the system to changes in temperature. The theory of chemical kinetics suggests that if the log of the rate of reaction is plotted against the reciprocal of the absolute temperature, a straight line should result whose slope measures the 'energy of activation' of the compound undergoing reaction. In chemistry this is called E, but in biology it is usually denoted by μ to signify that it may not be necessarily identified with an energy of activation but may represent a complex of factors. With some systems, *e.g.*, dehydrogenases, this method of plotting gives quite satisfactory linear relationships, but with more complex reactions, such as respiration, the empirical function suggested by Beleradek (13) is often better. For many purposes it is immaterial whether a theoretical interpretation is placed on these slopes; the important fact is that they can be determined with known precision and can be compared by a method free

of subjective bias. Examples of this use are described by Tam and Wilson (97), and by Burris and Wilson (13).

Biological assay.[31] Figure 7 represents a most important type of application of statistics to biological phenomena: the assay of a compound based on the

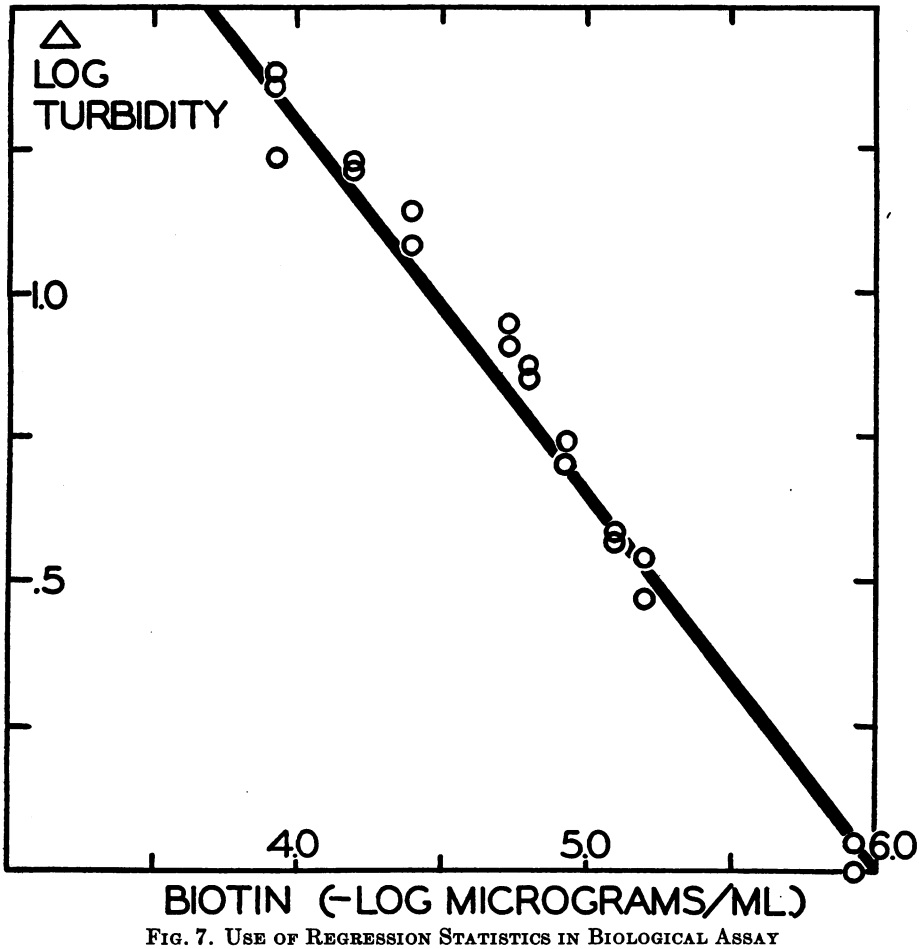

FIG. 7. USE OF REGRESSION STATISTICS IN BIOLOGICAL ASSAY

The graph is a standardization curve relating the growth of *Rhizobium trifolii* to the biotin concentration of the medium. The growth is estimated by the increase in the turbidity of a supension, measured in a photoelectric colorimeter. From such a graph, biotin in an unknown could be estimated by observing the effects of different quantities added to the basic medium on the growth of these bacteria.

[31] Space does not permit a thorough discussion of this important field for application of statistics to biological problems. The outstanding contributions of Bliss (3, 4, 7), Gaddum (38), Wilcoxon and McCallan (111), and others during the past decade has removed much of the sting of Burn's (11) 1930 observation that "Biological assay, as carried out by the majority of the workers in the world, still remains a subject for amusement or despair, rather than for satisfaction and self-respect." The cited references discuss the special methods developed for experimental design and analysis of the data together with examples of their use, including such fields of bacteriological interest as: toxicity tests on fungicides, response of animals to the administration of therapeutic drugs, bio-assay of antitoxin preparations, *etc.*

response of a biological agent. In this particular case *Rhizobium trifolii* was grown in colonies on a synthetic agar medium to which various levels of a sample of Kögl's biotin had been added; after a suitable period of incubation, the turbidity of six of these colonies suspended in 10 ml of water was determined in an Evelyn photoelectric colorimeter. It was found that under the conditions used in the assay, plotting Δlog turbidity against −(log biotin concentration in μg/ml) gave a straight line. Estimation of this line resulted in the relation:

$$Y = 3.917 - 0.652X$$

in which $X = -\log$ biotin concentration and $Y = \Delta\log$ turbidity. In actual experiment, however, the turbidity is determined after a certain quantity of material to be assayed is added to the medium, the biotin concentration being calculated from this reading. For convenience, the equation is solved for X:

$$X = 6.01 - 1.534Y$$

The point to be emphasized is that even though X will be the dependent variable in actual use, the fitting of the line must be done with it as the independent variable, since the experimental design requires this procedure (26).

CORRELATION COEFFICIENT

The correlation coefficient between two quantities measured simultaneously is:

$$[35] \qquad r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2\Sigma(y - \bar{y})^2}} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{N\sigma_x\sigma_y}$$

This coefficient is particularly useful for measuring the association between two variables when these can be classed according to quantitative standards. (The treatment is readily extended to more than two variables in multiple correlation.) The statistic, r, is obviously related to the regression coefficient but differs from the latter in that its value is absolute in the sense that it is independent of the units used in the measurements. In cases where the data do not indicate which variable is independent, two regression lines are calculated, one which measures the regression of y on x with slope b_{yx}, and another which measures the regression of x on y with slope, b_{xy}. It can be shown that

$$[36] \qquad r^2 = (b_{yx})(b_{xy}),$$

i.e., r is the geometric mean of the regression coefficients. The value of r ranges from +1 (complete direct dependence) through 0 (complete independence,[32] to −1 (complete inverse dependence).

The distribution of r itself is very skew in small samples and changes its form

[32] When r is calculated from a set of values depicting a complete independence of x and y, r will be zero, but when $r = 0$, it does not follow that x and y are completely independent only that they are not linearly dependent. Thus $r = 0$ for a set of points lying on, and equally spaced around, the circle $x^2 + y^2 = 4$. For an elaboration see Rider (79).

rapidly as ρ, the true correlation coefficient of which r is the estimate, changes. For this reason it is difficult to correct for the skewness in estimating the probable range of an observed value of r. However, r can be transformed to a statistic,

$$[37] \qquad z = 1/2[\ln(1 + r) - \ln(1 - r)] = (1.15) \log \frac{(1 + r)}{(1 - r)},$$

whose distribution is nearly normal with a standard deviation of $1/\sqrt{(N - 3)}$. It follows then that the significance of an observed r will depend on the number, N, of paired observations. Fisher (30) furnishes a table which gives the values which r must reach for different values of N and for different levels of significance. For significance at the 5 per cent level, r must be at least 0.63 if based on 10 pairs of samples, but only 0.28 if based on 50 pairs. To test for significance between two observed values of r, these are transformed into the corresponding z's, and since the standard deviation of the latter is known, the difference can be tested in the usual way using a table of the normal probability curve.

Although r is a useful statistic for a quantitative measure of association, its use, or rather abuse, by some investigators has led to much nonsense. The source of most of this is the false reasoning that a significant correlation between two quantities reveals a causal relationship. Although the natural scientist has not erred with this piece of faulty logic so frequently as the investigator in the social sciences, the former is by no means guiltless. A measure of the improvement that can be made in the error of estimation through knowledge of the value of a correlated variable is given by $1 - r^2$. This quantity measures the percentage of variance retained; obviously unless r is rather high (order of 0.8) there is little improvement. Since the arithmetic of correlation is very similar to that already illustrated with the regression coefficient, it is unnecessary to discuss this in detail; any standard text on statistics will provide examples of the methods for calculation. The use of the correlation coefficient is so wide-spread that illustrations are probably familiar to all; three typical applications from bacteriological literature will be cited.

Edwards and Rettger (25) found that the maximum growth temperature of 104 strains of bacteria representing 18 species was closely correlated with the minimum temperature of destruction of indophenol (cytochrome) oxidase (0.843), catalase (0.845), and succinodehydrogenase (0.774). Martin (65) surveyed different types of Arizona soils for presence of *Azotobacter* and for their nitrogen-fixing ability, *p*H, and content of certain salts. He found significant negative correlation between the mg nitrogen fixed per gram of soil and the water-soluble sodium and calcium. Rather unexpectedly, no correlation was noted between nitrogen fixed and phosphate content or pH; sulfate and chloride content were correlated with nitrogen fixation only through their association with calcium and sodium. Recently, Vaughn and Levine (102) determined the correlation between significant characteristics of "Intermediate" cultures of the coliform bacteria and, on the basis of the findings, recognized two species which were allocated to the genus *Escherichia*.

TESTING FOR AGREEMENT BETWEEN OBSERVED AND EXPECTED FREQUENCIES

Often in experimental work decision must be made as to whether a given series of observed frequencies corresponds to that implied by some hypothesis, *e.g.*, in animal or plant breeding, does the F_2 generation follow the 3:1 Mendelian ratio? Early investigators were inclined to lean heavily on 'experience' to decide whether an observed discrepancy could reasonably be regarded as fortuitous. Such a procedure usually has a high subjective bias. In 1900, Karl Pearson (75) provided an objective procedure when he published his *Chi Square* criterion for testing goodness of fit. Through further work by Pearson and

TABLE 10

Comparison of distribution of nitrogen fixed by clover plants in agar with theoretical values from normal curve

RANGE, MG NITROGEN	OBSERVED FREQUENCY f_o	THEORETICAL FREQUENCY f_t	$f_o - f_t$	$\frac{(f_o - f_t)^2}{f_t}$
<0.30	3	5.8	−2.8	1.35
0.30–0.45	5	6.2	−1.2	0.23
0.45–0.60	9	10.4	−1.4	0.19
0.60–0.75	22	15.6	+6.4	2.62
0.75–0.90	29	21.2	+7.8	2.87
0.90–1.05	25	24.9	+0.1	0.00
1.05–1.20	23	26.4	−3.4	0.44
1.20–1.35	20	25.1	−5.1	1.04
1.35–1.50	17	21.2	−4.2	0.83
1.50–1.65	16	16.0	0.0	0.00
1.65–1.80	11	10.7	+0.3	0.01
1.80–1.95	7	6.4	+0.6	0.06
1.95–2.10	4 } 9	3.4 } 6.1	+2.9	1.38
>2.10	5 }	2.7 }		
Total.....	196	196		11.02

$n = 10$ $P = 0.35$

others, notably R. A. Fisher, the field of application of the X^2 criterion has been greatly broadened and its interpretation clarified.

The Chi Square Test for Goodness of Fit. In table 10, an observed distribution of nitrogen fixed by clover plants in agar is compared with the distribution corresponding to the normal curve with mean and variance, estimated from the data, of $\bar{x} = 1.129$ and $\Sigma (x - \bar{x})^2/(N - 1) = 0.1888$ respectively. Do the observed frequencies in the respective classes, f_o, agree within the limits of sampling fluctuations with the theoretical frequencies, f_t, corresponding to this normal curve? To answer this question one calculates

$$[38] \qquad X^2 = \Sigma \frac{(f_o - f_t)^2}{f_t} .$$

when Σ denotes summing over the cells of the table. The accuracy of the test is improved by combining classes in the 'tails' to bring the theoretical frequency

up to 5 or more, as indicated in table 10. The distribution of X^2 depends on n, the number of degrees of freedom involved; tables of the significance levels of X^2 are available (30, 35, 79, 84, 98).

In this example there are, after combining, 13 classes of frequencies, so that X^2 possesses $13 - 3 = 10$ degrees of freedom, since the normal curve chosen was selected to have the same total frequency, the same mean, and the same variance thereby introducing 3 constraints which absorb 3 degrees of freedom. The value of X^2 observed, 11.02, corresponds to a probability, P, of 0.35; that is, if the distribution of nitrogen fixed *is* normal, comparison with a normal curve fitted as above would be expected to yield a X^2 as large or larger 35 times in 100. Therefore, since X^2 does not exceed the 5 per cent significance level, the test does not indicate that the normality hypothesis should be discarded. However, it should be noted that 2 of the 13 components (underlined values in table 10) contribute approximately half of the total X^2. That marked skewness of the observed distribution is responsible for this feature is apparent when the data and fitted curve are plotted (117). In such a case one is reluctant to accept the normality hypothesis even in the light of a 'favorable' value of X^2, and should reserve judgment until further data are at hand.

K. Pearson (78) emphasizes that, although the X^2 test will enable the experimenter to determine whether a given curve (or type of distribution) will reasonably describe the observations and may even allow some choice between alternative graduation curves, the 'better' curve as judged by the higher value of P does not necessarily represent the distribution from which the material was drawn. It requires a large-sized sample to discriminate between alternative curves.

A second example of the X^2 test is provided by the data in table 2 in which the distribution of bacteria on the squares of a Petroff-Hausser counting chamber is compared with that based on the proper Poisson distribution. The number of frequency classes is 8, therefore X^2 has 6 d.f. as one degree each is lost through keeping equal the population totals and the means of the two distributions. The probability corresponding to X^2 is 0.43, and we may conclude that the data do not conflict with the hypothesis that they came from a Poisson series.

A useful application of the X^2 criterion depends on the fact that a series of independent X^2's may be summed to form a total X^2 which possesses degrees of freedom equal to the sum of the degrees of freedom of its components. Thus, routine testing in a laboratory may be checked occasionally by summing the X^2 made over a period and testing the sum so obtained. Wilson and Kullman (114) used this method of statistical control to check the accuracy of counts of the root nodule bacteria made with the Petroff-Hausser chamber method under routine laboratory procedures. In each trial 144 squares were counted, and during the period the species of organism, density of suspension, and type of medium were varied. For each of the 50 trials, a X^2 was calculated by the same procedure as illustrated in table 2 and the sum tested. Since the number of degrees of freedom involved was outside the range of published significance levels, use was made of the fact that for large values of n, the distribution of $\sqrt{2X^2}$ is approximately normal with unit standard deviation about a mean of $\sqrt{2n-1}$, so that the difference, $\sqrt{2X^2} - \sqrt{2n-1}$ can be regarded as a normal deviate. In this particular case a normal deviate of +0.57 was obtained, which is well below that for the 5 per cent level of significance.

A further use of X^2 is for testing independence in a contingency table. In such a table an individual is classified in two (or more) different ways and the question is: Are the two methods of classification independent? Contingency tables provide a rapid and simple method for detecting associations in enumeration data and are of especial value when either or all of the classifications is qualitative. A 2×2 contingency table is shown in table 11. The data are from the study of McCarter, Getz and Stiehm (62) on the comparative response of different classes of students to intracutaneous injections of purified protein derivatives (P.P.D.) from the avian and human types of the tubercle bacillus. The Short Course students are boys, all from farms; the Freshmen are first-year male students at the University, predominantly from urban homes.

The 'expected' values for each cell when calculated on the hypothesis of independence require the four frequencies to be proportional; hence they can be determined from the marginal values. For example, in the first cell the expected value is $(497 \times 739)/1026 = 357.98$; the other values are automatically determined since the marginal totals of expected and observed must be equal. This means that the X^2 found will have only one degree of freedom. Since X^2 is 5.58 which corresponds to a P value of 0.02, the rejection of the hypothesis of independence is indicated at the 5 per cent level of significance.[33] The cause of the high value of X^2 apparently is the excess of Short Course students who reacted positively to the avian P.P.D. Further information was obtained from the responses of the students when tested with both human and avian P.P.D. The results (table 12) gave a X^2 value of 8.10; as calculation of three cells in a single

[33] The value of X^2 for a 2×2 contingency table,

a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	$a + b + c + d = N$

may be evaluated, without calculating the expected values, from the formula

$$X^2 = \frac{N\,(a\,d - b\,c)^2}{(a + b)\,(c + d)\,(b + d)\,(a + c)}$$

When expected frequencies less than 500 occur in a 2×2 table, the 'Yates correction for continuity' should always be applied. This consists of decreasing by 1/2 the frequencies which exceed their expected values, and increasing by 1/2 the frequencies which fall short of their expected values. These steps may be shortcut by noting that

$$\text{corrected } X^2 = \frac{N\,(\,|\,ad - bc\,|\,N/2)^2}{(a + b)\,(c + d)\,(b + d)\,(a + c)}$$

where $|\,ab - bc\,|$ denotes the value of $(ad - bc)$ *taken positive*, so that the correction, $-N/2$, always reduces the magnitude of the quantity to be squared.

In the present case the corrected X^2 is 5.26 which is still significant at the .05 level.

No satisfactory correction of this sort has been found for the general $r \times c$ table. For further discussion of this correction and the additional refinements in the case of a 2×2 table see Fisher and Yates (35).

row will fix all the other values if the marginal totals remain unchanged, there are 3 d.f. In general, a table of r rows and c columns has $(r - 1)(c - 1)$ d.f. As the probability is again less than 0.05, the hypothesis of independence is doubtful. As can be seen in the table, the largest contribution to X^2 arises from the excess of Short Course students who are human P.P.D.-negative and avian-positive.

TABLE 11

*Contingency table showing response of students to avian P.P.D.**

	AVIAN +	AVIAN −	TOTALS
Freshmen	(357.98) 341 0.805	(381.02) 398 0.757	739
Short Course	(139.02) 156 2.074	(147.98) 131 1.948	287
Totals	497	529	1,026

$X^2 = 5.58$ $n = 1$ $P = 0.02$

* In both contingency tables the expected values for each cell is given in parentheses; the contribution of X^2 from each cell is underlined.

TABLE 12

Contingency table showing response of students to human (H) and avian (A) P.P.D.

	H+ A+	H+ A−	H− A+	H− A−	TOTALS
Freshmen	(173.63) 176 0.032	(3.43) 3 0.054	(177.05) 165 0.820	(384.89) 395 0.265	739
Short Course	(29.37) 27 0.191	(0.57) 1 0.324	(29.95) 42 4.848	(65.11) 55 1.570	125
Totals	203	4	207	450	864

$X^2 = 8.10$ $n = 3$ $P = 0.045$

D^2 Test for Binomial Distributions.[34] A rapid method of testing for agreement with binomial sampling—and which is the only method practicable when

[34] The index of dispersion appropriate to the binomial distribution which we have denoted by D^2 was introduced by an author using the pseudonym "Mathetes" (66) following an analogy with previous work of Fisher, Thornton, and Mackenzie (34). Its use as a test of sampling technique has been illustrated by Fisher (30, sec. 19); Snedecor (84, sec. 9.6); Snedecor and Irwin (85); and by others. Cochran (19), Haldane (46) and Welch (104) have studied its sampling distribution in very small samples from binomial populations.

the samples are of different sizes—is based on the *dispersion index* appropriate to the binomial distribution, namely:

$$[39] \qquad D^2 = \Sigma \frac{(x_i - N_i p')^2}{N_i p'(1 - p')}$$

where N_i denotes the number of individuals in the i^{th} sample, $(i = 1, 2, \cdots, k)$, x_i is the number of the individuals in the i^{th} sample which possess the characteristic under investigation, and $p' = \Sigma x_i / \Sigma n_i$ is the observed *proportion* of individuals with this characteristic in all the data at hand, *i.e.*, in all k samples lumped together. If all k samples are of the same size, so that $N_i = N$ for all i, then [39] simplifies to

$$[40] \qquad D^2 = \frac{N\Sigma(x_i - \bar{x})^2}{\bar{x}(N - \bar{x})} = N \frac{\Sigma x_i^2 - (\Sigma x_i)^2/k}{\bar{x}(N - \bar{x})}$$

where $\bar{x} = \Sigma x_i / k$ is the average number of individuals with the characteristic in question per sample.

This index of dispersion, D^2, is essentially a criterion for judging whether the variance of the observed frequency distribution is in agreement with the variance of the binomial distribution *having the same mean.* When sampling is in accordance with the binomial distribution, the expected value of D^2 is $(k - 1)$, and its sampling distribution about this mean is well represented by the tabular X^2-distribution for $(k - 1)$ degrees of freedom provided k is small compared with ΣN_i, *i.e.*, provided that either the individual samples are large, or, if they are small, that there are many of them. In consequence, the tabulated significance levels of X^2 can be used for testing the statistical significance of an observed value of D^2.

For data arranged in a frequency table, such as table 1, a form of equation 39 more convenient for calculation is

$$[41] \qquad D^2 = \frac{\Sigma f_x x^2 - \dfrac{(\Sigma f_x x)^2}{\Sigma f_x}}{\bar{x}(N - \bar{x})}$$

where $\bar{x} = (\Sigma f_x x)/(\Sigma f_x)$. For the data in table 1: $N = 100$, $k = \Sigma f_x = 113$, $\Sigma f_x x = 673$, and $\Sigma f_x x^2 = 4779$, yielding $D^2 = 137.612$ as compared with an expected value of 112. Since tables of the significance levels of X^2 do not go as high as 112 degrees of freedom, whether the observed D^2 is significantly greater than its expected value has to be tested here by employing $\sqrt{2X^2} - \sqrt{2n - 1}$ as a normal deviate with unit standard deviation and taking $X^2 = D^2$ and $n = k - 1$. In the present instance, $\sqrt{275.224} - \sqrt{223} = +1.66$ which is

These writers, with the exception of Welch, denote this index of dispersion by X^2; Welch uses D. We have used D^2 to avoid confusion with the X^2 goodness-of-fit criterion, adopting D^2 instead of D because the former carries the implication that its value is always positive or zero.

just significant at the .05 level.[35] In brief, the variance of the observed frequency distribution exceeds significantly the variance of the binomial distribution with the same mean.

In the preceding example, the observed D^2 was significantly larger than the value expected on the hypothesis of binomial sampling. Although values of D^2 may frequently be obtained which are less than the values expected in binomial sampling, generally these will not differ widely from the expected value. If, however, values considerably less than the expected values occur, the significance of the discrepancies may be judged by noting whether D^2 is *less* than the 0.95 level of X^2. Should such 'significantly small' values of D^2 be a common occurrence, it is well to seek an explanation of the unusual uniformity of the samples. For example, had a significantly small value of D^2 been obtained in the study on monocytes (table 1), one might have inquired whether some physiological factor controlled their distribution in the blood so that this distribution was more uniform than would occur in random mixing.

SUMMARY

On the basis of the discussion given in the main body of the text the following points should be emphasized:

1. Statistical analyses provide no substitute for proper and precise experimental technique. The quantitative relationships to be derived are not altered through any statistical magic; their accuracy is that of the observations. As Fisher has said, "The statistician must be treated less like a conjurer whose business is to exceed expectation, than as a chemist who undertakes to assay how much of value the material submitted to him contains" (Rothamsted report for 1933). As was illustrated in several instances, however, statistical analysis can be of great assistance in providing a measure of the adequacy of a particular method used in practice.

2. Although statistical analysis will divulge only those facts present in the observations, such analysis furnishes a tool for extracting information inherent in the data but not readily evident by mere inspection. As a corollary to this, statistical methods allow quite complicated experiments to be designed in which the influence of each of several variables on some particular phenomenon as well as their interactions can be simultaneously determined. Such complex experiments are extremely valuable for saving of time, labor and money, and

[35] We are testing here whether D^2 significantly *exceeds* its expected value so the .05 significance of D^2 is judged by Fisher's .10 significance level (1.64485) for a normal deviate with unit variance. The reader with some statistical experience may be curious to know why, that we have gone to the trouble of finding the expected frequencies, we do not test the agreement of the observed and expected frequencies with the X^2-test of goodness-of-fit. To meet the conditions of applicability of the X^2-test we should have to group together (as indicated in table 1) several of the frequencies at the tails on account of the small numbers involved, thereby reducing the sensitivity of the X^2-test. Fisher (30, sec. 19) discusses the relative usefulness of the two tests.

provide a truer representation of the actual situation than several experiments in which the several factors are varied one at a time.

3. Statistical considerations will suggest the proper design of an experiment—proper not only in the sense that the results are most readily amenable to statistical treatment, but also from the viewpoint of economy and efficiency. A corollary to this advantage is that the statistician should be consulted before the experiment is performed rather than asked to do the impossible: make valid conclusions from the data of a poorly designed experiment.

4. Statistical theory emphasizes the necessity of repeating an experiment. Statistical analysis is of little use on the individual experiment unless something is known about the properties of the population; usually this is most accurately obtained through repeated trial under similar circumstances. Although most experimenters appreciate the fact that results "have weight" only when they have been obtained in several experiments, without recourse to statistical considerations it is often difficult to determine how many repetitions are needed. Often, by statistical procedures, the experimenter may decide beforehand by using his previous experience whether the results of a contemplated investigation will be worth the effort, time and expense required to obtain an unequivocal answer.

5. Statistical measures in conjunction with statistical theory provide a means for condensing information derived from large-scale experimentation.[36] The essential information in a large mass of data, whose tabulation would require much space and whose very size may intimidate a reader wishing to make his own interpretation, can often be summarized in the form of a few statistical measures (*e.g.*, number of observations taken, mean with its standard deviation, a regression coefficient with its standard error, an analysis of variance) with little loss of relevant knowledge.

APPENDIX AND EXPLANATORY COMMENTS

As with other branches of science, statistics possesses a technical vocabulary in which common words are used in a sense which often proves confusing to those unfamiliar with the special meaning. Many of these are concisely defined by mathematical formulas which are supplied in the text; others need descriptive explanation which is given in the following glossary of terms:

Argument: one of the independent variables upon which a tabled function depends, the values of which are given at the margin of the table and make it possible to locate the corresponding tabled values, *e.g.*, in a table of logarithms, the function tabled is $\log N$ and N is the argument.

Confidence Intervals and Fiducial Limits: it has long been realized that a single value computed from a sample as an estimate of a parameter θ has very little chance of actually equalling θ, and that some sort of range to indicate the probable accuracy of the estimate is needed. Thus, with an observed proportion, p_o, the estimate of the true proportion, p, was given as $p_o \pm \lambda\sigma_o$ where $\sigma_o = \sqrt{(p_o q_o)/N}$ and λ was chosen to correspond to some probability P, and a statement often made was: "The probability the true proportion, p, lies outside the limits $p_o - \lambda\sigma_o$ and $p_o + \lambda\sigma_o$ is less than or equal to P." So far as we have been

[36] Further discussion and illustration of this point are given in the *A.S.T.M. Manual on Presentation of Data*, American Society for Testing Materials (260 S. Broad St., Philadelphia, Pa.), 3rd printing, August, 1940.

able to determine, E. B. Wilson (112) first pointed out that such a statement is erroneous. He stressed that it is p_o which is a chance variable, not p, and that a correct statement must take some such form as: If it is not true that $p_l < p < p_u$, where p_l and p_u depend on p_o and P, then the probability of our observing the value p_o itself, or any more improbable value, is less than P.

Wilson showed in the above case how to calculate the limits p_l and p_u, but did not follow up his note with a general discussion of the procedure to be followed in other cases. The first general treatment was that of R. A. Fisher (32), who appears to have been unaware of Wilson's note. Fisher terms such limits *fiducial limits*, and his development is in terms of what he calls *fiducial probability*. About 1930, J. Neyman independently initiated his theory of *confidence intervals* (see ref. 70, where a list of the principal papers on fiducial limits and confidence intervals is given) and for some time it seemed to many statisticians that the two approaches were equivalent, since in the cases considered they had led to identical ranges for the parameter being estimated. A further paper by Fisher (33b) threw doubt on the equivalence of the two approaches, which doubt increased as more papers appeared. The two approaches have been contrasted by Neyman (70), and, while they lead to identical results in many instances, there are cases of disagreement, so they should no longer be regarded as equivalent. Neyman's approach seems to be more general, and for this reason, as well as from a greater familiarity with it, we have adopted the method of confidence intervals in the present paper.

Degrees of Freedom: the number of variates upon which a quantity depends minus the number of constraints upon these variates, *e.g.*, if N independent x values are concerned, then $\Sigma(x - c)^2$ has N degrees of freedom provided c is in no way determined by the values of the x's themselves; on the other hand, $\Sigma(x - \bar{x})^2$, where $\bar{x} = \frac{1}{N}\Sigma x$ = average of x values, has only $(N - 1)$ degrees of freedom since the quantities $(x - \bar{x})$ are constrained by the equation $\Sigma(x - \bar{x}) = 0$.

e: base of natural logarithms; is a constant like π. Common logarithms use 10 as a base. In this paper ln signifies that the base is e, log, that the base is 10.

Entry: a quantity appearing in the body of a table, *e.g.*, the tabled values of log N are entries.

Likelihood (as introduced by R. A. Fisher)*:* the likelihood of a particular value θ' of a parameter θ in the light of an observed sample, is proportional to the probability of this sample when $\theta = \theta'$; the largest likelihood can be assigned the value unity by convention, if desired; in comparing two particular values, θ' and θ'', of a parameter θ, importance attaches only to the ratio of their likelihoods, not to the values of the respective likelihoods. Thus, while we may know nothing about the relative frequency with which $\theta = \theta'$ and $\theta = \theta''$ in a particular type of research, it may be an inescapable fact that the probability of the observed sample is three times as great when $\theta = \theta'$ as when $\theta = \theta''$. "If we need a word to characterize this relative property of different values of $[\theta]$, I suggest that we may speak without confusion of the *likelihood* of one value of $[\theta]$ being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity $[\theta]$ would in fact yield the observed sample." (28, p. 326). Likelihood also differs from probability in that it is not capable of being summed or integrated; whereas the sum of the probabilities of all possible values of a random variable is unity, the sum of the likelihoods, computed from a particular sample, of the admissible values of a parameter will in general be infinite. An exact knowledge of the likelihood of different values of θ tells us nothing whatever about the 'probability' that θ will fall in any given range. Indeed, in well conducted research, θ will not have a probability 'distribution', but will be confined to a single value, the identity of which is unknown, and the object of the research will be to estimate this value as closely as possible.

Method of Maximum Likelihood: estimating a parameter from a sample by choosing the value of the parameter which has the largest likelihood, as defined above, when calculated

from the sample at hand; the estimate so obtained is termed the *optimum estimate* of the parameter, or its *maximum likelihood estimate.*

When the parameter θ can itself be regarded as a random variable drawn from a superpopulation in which all values of θ are equally probable *a priori*, then the value of θ which has the highest 'probability' of 'producing' the observed sample—this *a posteriori* 'probability' being calculated with the aid of Bayes' Theorem—is identical with the maximum likelihood estimate of θ from this sample. This property was employed by Gauss in 1809 to justify his development of the method of least squares by a formulation identical with that now used in the method of maximum likelihood. Later (1829), however, he gave less stress to this argument through the conviction that maximizing the probability was less important than minimizing the injurious effects of the actual errors of estimation (see Fisher, (33b) p. 249).

The justification of the method of maximum likelihood by Fisher and his followers comes from this later viewpoint. Fisher (28, 30a) has shown that, of the estimates calculated from large samples, the one obtained by maximizing the likelihood is in general the one for which the intrinsic accuracy is greatest, and that, when there exist estimates, called *sufficient statistics*, which exhaust the information in the sample about the parameter, then these are the maximum likelihood estimates. Thus, use of the method of maximum likelihood is justified by the advantageous properties of the estimates to which it leads. Further elaboration can be found in two expository papers by Fisher (33a, b), the former being supplemented by the critical comments of many noted statisticians, of which J. Neyman's are especially interesting.

Null Hypothesis: a hypothesis, relating to the parent population(s) of the research data under consideration, and whose acceptance or rejection is to be based on the agreement or non-agreement between some of its logical consequences and corresponding aspects of the research data. Obviously a null hypothesis must be sufficiently specific to permit the deduction of some criterion of agreement between observations and this hypothesis. For example, the assertion that a drug "reduced the mortality" from a certain disease "*10 per cent*" could constitute a null hypothesis. Mere assertion that it "reduced the mortality" could *not* constitute a null hypothesis, however, since this hypothesis is not sufficiently specific to permit an evaluation of the agreement of observations with the hypothesis. In this case, the null hypothesis is usually taken to be that the drug produced *no reduction* in mortality from the disease, the contradiction of which by experimental results would lead to the inference that the treatment reduced the mortality.

Parameter: in the strictest sense, a constant in the mathematical formula specifying a hypothetical population the value of which serves to distinguish a specific population from others having the same functional form, and by varying the value of which different populations of the same general family can be specified, *e.g.*, m, is the parameter of the Poisson distribution, the general term of which is

$$\frac{e^{-m}m^x}{x!}$$

More generally, any descriptive quantity such as an average, correlation coefficient, *etc.*, relating to a population and which serves to distinguish this population from other conceivable populations.

Parent Population: used in connection with a particular sample to designate the population from which it was drawn.

Population or Universe: the aggregate of all individuals or objects which, by reason of some characteristics in common, may logically be regarded as comprising the set of objects under consideration. When the objects are susceptible of complete enumeration, the population is *finite*; otherwise, *infinite*. Thus, a set of laboratory animals constitutes a finite population, whereas the temperatures at which a species of bacteria will grow forms an infinite population. A finite population may consist of so large a number of items as

to permit being treated as infinite without sensible error, *e.g.*, yeast cells growing in a vat. The foregoing refers to *real* populations. Often *hypothetical* populations are employed. An example is the abstract population of offspring which a geneticist regards as producible by repeated matings of specified character, although some individuals of such a population can never be produced; furthermore, the population exists in the abstract even if no matings are carried out at all.

In bacteriology it is often desirable to regard a quantity of liquid as a population, *e.g.*, a sample of water or milk. In such cases it appears somewhat artificial to regard the molecules as the individual objects comprising the population, and it seems more natural to regard it as an aggregate of arbitrary volumetric units of the liquid, such as milliliters. Although the latter approach is often convenient, two aspects of it should be noted: the arbitrary size of the constituent 'objects' and their momentary identity under any sort of mixing process.

Random: an operation, such as drawing a sample from a population or arranging pots in a greenhouse, is performed in a *random* manner when its execution is such that *a priori* each and every possible outcome has an equal chance of occurrence. Thus, a sample drawn from a population in such a manner that *a priori* each and every individual of the population has an equal chance of being included is termed a *random sample*. Since any sample whatever can be obtained either by a random or by a non-random operation, it is important to note that *it is the operation of drawing the sample which is random and not the sample itself*.

The presence of randomness, in sampling and in laying out experimental arrangements, is required for the validity of statistical tests of significance. Accordingly, operations which are to be performed at random should be faithfully randomized with the aid of a table of random numbers, dice, drawing cards out of a hat, *etc.*—mere absence of conscious system is not enough to insure freedom from bias. For illustrations of sampling bias arising in samples selected without attention to randomization, see Cochran and Watson (22) and Yates (122).

Replicates: the subdivisions of an experiment which are similar with respect to some factor under investigation although they may be dissimilar with respect to other pertinent characteristics. For example, the animals inoculated with a particular one of several organisms being tested constitute the replicates of that inoculum although they may come from different litters or correspond to different age groups, *etc.*

Σ: the sum of, *e.g.*, if there are N values of x, then $\frac{1}{N}\Sigma x$ denotes the average (arithmetic mean).

Sample: a finite portion of a population.

Statistic: any quantity calculated from an observed sample with a view to characterizing the parent population; a value calculated from a sample as an estimate of a parameter of the parent population, *e.g.*, the mean of a sample is a statistic in that it provides an estimate of the mean of the parent population. A statistic used to estimate a parameter need not be the quantity in the sample which directly corresponds to the parameter in the population, *e.g.*, the range of a sample (*i.e.*, largest observation minus the smallest) provides an estimate of, and hence is a statistic for, the standard deviation of the population, as is also the standard deviation of the sample. Infinitely many statistics could be devised for estimating a particular parameter from an observed sample, but comparatively few of these would be of practical utility, and of these there are often strong theoretical reasons for a particular choice.

Statistically Significant: a discrepancy between some property of a sample and that which is expected on the basis of a particular null hypothesis is *statistically significant* if the probability of a discrepancy as bad or worse arising solely from sampling fluctuations (*i.e.*, from errors of measurement, biological variation, *etc.*) admissible under the null hypothesis is less than some preassigned quantity, α, known as the *level of significance*. The value of α usually adopted is 0.05, and discrepancies statistically significant on this basis are said to be *significant at the .05 (or 5 percent) level of significance*. Other values of α

are also used; the desiderata to be taken into consideration in choosing a level of significance for a particular line of research are discussed in the text. Early writers used P to denote the level of significance, and this notation is still employed by many biologists, but the use of α for this purpose has definite advantages when tests of significance are considered from the viewpoint of Neyman and Pearson, so that today α is employed by the majority of American writers on mathematical statistics.

It should be noted that the statistical significance of a result depends solely on the probability of occurrence of results equally or more discrepant as a consequence of sampling fluctuations admissible under the null hypothesis, and is not in itself an indication of the practical significance of the result. For example, an observed correlation coefficient, r, of .10 calculated from a sample of 500 is significantly different from zero at the .05 level and suggests the existence of a real correlation of the order of .10 between the variables under consideration. If the *complete independence* of these variables is important, this contradictory evidence is of practical significance. On the other hand, if it is desired to predict the values of one variable from that of the other, this correlation is practically worthless, for it means that $100(1 - r^2) = 99\%$ of the variation in each variable is independent of variation in the other.

Variable: a quantity which in a given context can assume different values in different individual cases; the antithesis of a constant, the latter being a quantity which in a given context can have only a single value (which may or may not be known.)

Variate: a variable; when used in the plural the word often refers to the particular values of a single variable corresponding to individual cases, especially when these values are unknown, *e.g.*, the heights of ten individuals, denoted by $x_1, x_2, \cdots, x_{10}$, are ten variates, but height symbolized by x is the only variable.

REFERENCES

(1) BARTLETT, M. S. 1937 The square root transformation in analysis of variance. J. Roy. Stat. Soc., Suppl., **3**, 68–78.

(2) BARTLETT, M. S. 1937 Some examples of statistical methods of research in agriculture and applied biology. J. Roy. Stat. Soc., Suppl., **4**, 137–170.

(3) BLISS, C. I. 1935 The calculation of the dosage-mortality curve. Ann. Applied Biol., **22**, 134–167.

(4) BLISS, C. I. 1935 A comparison of dosage-mortality data. Ann. Applied Biol., **22**, 307–333.

(5) BLISS, C. I. 1937 The analysis of field experimental data expressed in percentages. (In Russian, with summary and legends accompanying tables in English). Plant Protection, fasc., **12**, 67–77.

(6) BLISS, C. I. 1938 The transformation of percentages for use in the analysis of variance. Ohio J. Sci., **38**, 9–12.

(7) BLISS, C. I. 1941 Biometry in the service of biological assay. Ind. Eng. Chem., Anal. Ed., **13**, 84–88.

(8) BORTKIEWICZ, L. v. 1898 Das Gesetz der kleinen Zahlen. B. G. Teubner, Leipzig.

(9) BRANDT, A. E. 1937 Factorial design. J. Am. Soc. Agron., **29**, 658–667.

(10) BUCHBINDER, L., SOLOWEY, M., AND SOLOTOROVSKY, M. 1941 Studies on microorganisms in simulated room environments. IV. The effect of survival on the pathogenic properties of streptococci: mouse virulence. With a statistical appendix by E. B. Phelps. **J. Bact., 42**, 615–630.

(11) BURN, J. H. 1930 The errors of biological assay. Physiol. Rev., **10**, 146–169.

(12) BURN, J. H. 1937 Biological Standardization. Oxford University Press. Humphrey Milford, London.

(13) BURRIS, R. H., AND WILSON, P. W. 1939 Respiratory enzyme systems in symbiotic nitrogen fixation. Cold Spring Harbor Symposia on Quantitative Biology, **7**, 349–361.

(14) BURTON, J. C., AND WILSON, P. W. 1939 Host plant specificity among the medicago in association with root-nodule bacteria. Soil Sci., **47**, 293–302.

(15) CAIRNS, W. D. 1918 Note on the geometrical mean as a *B. coli* index. Science, n.s., **47**, 239–240.
(16) CLARK, A., AND LEONARD, W. H. 1939 The analysis of variance with special reference to data expressed as percentages. J. Am. Soc. Agron., **31**, 55–66.
(17) CLARK, H. 1927 On the titration of bacteriophage and the particulate hypothesis. J. Gen. Physiol., **11**, 71–81.
(18) CLOPPER, C. J., AND PEARSON, E. S. 1934 The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, **26**, 404–413.
(19) COCHRAN, W. G. 1936 The X^2 distribution for the binomial and Poisson series with small expectation. Ann. Eugenics, **7**, 207–217.
(20) COCHRAN, W. G. 1938 An extension of Gold's method of examining the apparent persistence of one type of weather. Quart. J. Roy. Meteor. Soc., **64**, 631–634.
(21) COCHRAN, W. G. 1938 Some difficulties in the statistical analysis of replicated experiments. Empire J. Exptl. Agric., **6**, 157–175.
(22) COCHRAN, W. G., AND WATSON, D. J. 1936 An observer's bias in the selection of shoot-heights. J. Exptl. Agr., **4**, 69–76.
(23) DE MOIVRE, ABRAHAM 1733 *Approximatio and Summam Terminorum Binomii* $(a + b)$ *in Seriem expansi*, pp. 1–7, a second supplement (dated November 12, 1733) to his *Miscellans Analytica* first published several years before.
(24) DURHAM, F. M., GADDUM, J. H., AND MARCHAL, J. E. 1929 Reports on biological standards. II. Toxicity tests for novarsenobenzene (neosalvarsan). Privy Council. Med. Research Council (Brit.), Spec. Rept. Series, No. 128. *Note.* The tables of the binomial probability distribution included in this report have been reprinted as Appendix III in Burn (12).
(25) EDWARDS, O. F., AND RETTGER, L. F. 1937 The relation of certain respiratory enzymes to the maximum growth temperatures of bacteria. J. Bact., **34**, 489–515.
(26) EISENHART, C. 1939 The interpretation of certain regression methods and their use in biological and industrial research. Ann. Math. Stat., **10**, 162–186.
(27) ELVEHJEM, C. A., AND WILSON, P. W., *et al.* 1939 Respiratory Enzymes. Burgess Publishing Company, Minneapolis.
(28) FISHER, R. A. 1922 On the mathematical foundations of theoretical statistics. Trans. Roy. Soc. (London), **222A**, 309–368.
(29) FISHER, R. A. 1922 On the dominance ratio. Proc. Roy. Soc., Edinburgh, **42**, 321–341.
(30) FISHER, R. A. 1925 Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh. And subsequent editions.
(30a) FISHER, R. A. 1925 Theory of statistical estimation. Proc. Cambridge Phil. Soc., **22**, 700–725.
(31) FISHER, R. A. 1926 Applications of "Student's" distribution. Metron, **5**, 90–104.
(32) FISHER, R. A. 1930 Inverse probability. Proc. Cambridge Phil. Soc., **26**, 528–535.
(33) FISHER, R. A. 1935 The Design of Experiments. Oliver and Boyd, Edinburgh.
(33a) FISHER, R. A. 1935 The logic of inductive inference. J. Roy. Stat. Soc., **98**, 39–82.
(33b) FISHER, R. A. 1936 Uncertain inference. Proc. Am. Acad. Arts Sci., **71**, 245–258.
(34) FISHER, R. A., THORNTON, H. G., AND MACKENZIE, W. A. 1922 The accuracy of the plating method of estimating the density of bacterial populations. With particular reference to the use of Thornton's agar medium with soil samples. Ann. Applied Biol., **9**, 325–359.
(35) FISHER, R. A., AND YATES, F. 1938 Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, Edinburgh.
(36) FRY, T. C. 1928 Probability and its Engineering Uses. D. Van Nostrand, New York.
(37) FULMER, E. I., AND BUCHANAN, R. E. 1929 The thermal increments and critical temperatures of biological reactions. Proc. Soc. Exptl. Biol. Med., **26**, 446–449.
(38) GADDUM, J H. 1933 Reports on biological standards. III. Methods of biological

assay depending on a quantal response. Privy Council. Med. Research Council, (Brit.) Special Rept. Series, 183.

(39) GEE, L. L. AND SARLES, W. B. 1942 The disinfection of trout eggs contaminated with *Bacterium salmonicida*. J. Bact., **44**, 111–126.

(40) GORDON, R. D. 1938 Note on estimating bacterial populations by the dilution method. Proc. Natl. Acad. Sci., U. S., **24**, 212–215.

(41) GORDON, R. D. 1939 Estimating bacterial populations by the dilution method. Biometrika, **31**, 167–180.

(42) GORDON, R. D., AND ZOBELL, C. E. 1938 Note on the successive dilution method for estimating bacterial populations. Zentr. Bakt. Parasitenk., II Abt., **99**, 318–320.

(43) GREATHOUSE, G. A., KLEMME, D. E., AND BARKER, H. D. 1942 Determining the deterioration of cellulose caused by fungi. Ind. Eng. Chem., Anal. Ed., **14**, 614–620.

(44) GREENWOOD, M., AND WHITE, J. D. C. 1908 A biometric study of phagocytosis with special reference to the "opsonic index." Biometrika, **6**, 376–401.

(45) GREENWOOD, M., AND YULE, G. U. 1917 On the statistical interpretation of some bacteriological methods employed in water analysis. J. Hyg., **16**, 36–54.

(46) HALDANE, J. B. S. 1939 The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small. Biometrika, **31**, 346–355.

(47) HALVORSON, H. O., AND MOEGLEIN, A. 1940 Application of statistics to problems in bacteriology. V. The probability of occurrence of various experimental results. Growth, **4**, 157–168.

(48) HALVORSON, H. O., AND ZIEGLER, N. R. 1933 Quantitative Bacteriology. Burgess Publishing Company, Minneapolis.

(49) HALVORSON, H. O., AND ZIEGLER, N. R. 1933 Application of statistics to problems in bacteriology. I. A means of determining bacterial population by the dilution method. J. Bact., **25**, 101–121.

(50) HALVORSON, H. O., AND ZIEGLER, N. R. 1933 II. A consideration of the accuracy of dilution data obtained by using a single dilution. J. Bact., **26**, 331–339.

(51) HALVORSON, H. O., AND ZIEGLER, N. R. 1933 III. A consideration of the accuracy of dilution data obtained by using several dilutions. J. Bact., **26**, 559–567.

(52) HARMSEN, G. W., AND VERWEEL, H. J. 1936 The influence of growth-promoting substances upon the determination of bacterial density by the plating-method. Zentr. Bakt. Parasitenk., II Abt., **95**, 134–150.

(53) HOTELLING, H., AND FRANKEL, L. R. 1938 The transformation of statistics to simplify their distribution. Ann. Math. Stat., **9**, 87–96.

(54) JAMES, N., AND SUTHERLAND, M. L. 1939 The accuracy of the plating method for estimating the numbers of soil bacteria, actinomyces, and fungi in the dilution plated. Can. J. Research, **170**, 72–86.

(55) JAMES, N., AND SUTHERLAND, M. L. 1939 The accuracy of the plating method for estimating the numbers of bacteria and fungi from one dilution and from one aliquot of a laboratory sample of soil. Can. J. Research, **170**, 97–108.

(56) JAMES, N., AND SUTHERLAND, M. L. 1940 Effect of numbers of colonies per plate on the estimate of the bacterial population in soil. Can. J. Research, **180**, 347–356.

(57) JAMES, N., AND SUTHERLAND, M. L. 1940 Fluctuations in numbers of bacteria in soil. Can. J. Research, **180**, 435–443.

(58) JENNISON, M. W., AND WADSWORTH, G. P. 1940 Evaluation of the errors involved in estimating bacterial numbers by the plating method. J. Bact., **39**, 389–397.

(59) LAPLACE, P. S. 1812 Theorie Analytique des Probabilités. Paris. And later editions.

(60) LIND, C. J., AND WILSON, P. W. 1941 Mechanism of biological nitrogen fixation. VIII. Carbon monoxide as an inhibitor for nitrogen fixation by red clover. J. Am. Chem. Soc., **63**, 3511–3514.

(61) LINEWEAVER, H., AND BURK, D. 1934 The determination of enzyme dissociation constants. J. Am. Chem. Soc., **56**, 658–666.
(62) MCCARTER, J., GETZ, H. R., AND STIEHM, R. H. 1938 A comparison of intracutaneous reactions in man to the purified protein derivatives of several species of acid-fast bacteria. Am. J. Med. Sci., **195**, 479–493.
(63) MCCRADY, M. H. 1915 The numerical interpretation of fermentation-tube results. J. Infectious Diseases, **17**, 183–212.
(64) MCCRADY, M. H. 1918 Table for rapid interpretation of fermentation-tube results. Public Health J., **9**, 201–220.
(65) MARTIN, W. P. 1940 Distribution and activity of *Azotobacter* in the range and cultivated soils of Arizona. Agr. Expt. Sta. Tech. Bull. No. 83, Arizona, 335–369.
(66) "MATHETES" 1924 Statistical study on the effect of manuring on infestation of barley by gout fly. Ann. Appl. Biol., **11**, 220–235.
(67) MATUSZEWSKI, T., NEYMAN, J., AND SUPIŃSKA, J. 1935 Statistical studies in questions of bacteriology, Part 1. The accuracy of the "dilution method". J. Roy. Stat. Soc., Suppl., **2**, 63–82.
(68) MATUSZEWSKI, T., SUPIŃSKA, J., UND NEYMAN, J. 1936 Über die Wahrscheinlichkeit der Reinkulturisolierung aus einer Petrischale. Zentr. Bakt. Parasitenk., Abt. II, **95**, 45–53.
(69) NEYMAN, J. 1939 On a new class of "contagious" distributions, applicable in entomology and bacteriology. Ann. Math. Stat., **10**, 35–57.
(70) NEYMAN, J. 1941 Fiducial argument and the theory of confidence intervals. Biometrika, **32**, 128–150.
(71) NEYMAN, J., AND PEARSON, E. S. 1928 On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika, **20A**, 175–240, 263–294.
(72) NEYMAN, J., AND PEARSON, E. S. 1933 On the problem of the most efficient tests of statistical hypotheses. Trans. Roy. Soc. (London), **231A**, 289–337.
(73) NEYMAN, J., AND PEARSON, E. S. 1936 Contributions to the theory of testing statistical hypotheses. I. Unbiased critical regions of type A and type A_1. Stat. Research Mem., **1**, 1–37.
(74) PEARSON, E. S. 1939 Note on the inverse and direct methods of estimation in R. D. Gordon's problem. Biometrika, **31**, 181–186.
(75) PEARSON, K. 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag., 5th ser., **50**, 157–175.
(76) PEARSON, K. 1924 Historical note on the origin of the normal curve of errors. Biometrika, **16**, 402–404.
(77) PEARSON, K. 1930 Tables for Statisticians and Biometricians. Edited by Karl Pearson. Vol. I and II. 3 ed. Cambridge University Press.
(78) PEARSON, K. 1935 Statistical tests. Nature, **136**, 296–297, 550.
(78a) REED, J. L. 1925 Report of Advisory Committee on Official Water Standards. *B. coli* densities as determined from various types of samples. Public Health Repts. **40**, I, 704–716.
(79) RIDER, P. R. 1939 An Introduction to Modern Statistical Methods. John Wiley and Sons, New York.
(80) SAVAGE, G. M., AND HALVORSON, H. O. 1941 The effect of culture environment on results obtained with the dilution method of determining bacterial population. J. Bact., **41**, 355–362.
(81) SHEWHART, W. A. 1931 Economic Control of Quality Control of Manufactured Products. D. Van Nostrand Co., New York.
(82) SHEWHART, W. A. 1939 Statistical Method from the Viewpoint of Quality Control. Graduate School, United States Department of Agriculture, Washington.
(83) SIMON, L. E. 1941 An Engineer's Manual of Statistical Methods. John Wiley and Sons, New York.

(84) SNEDECOR, G. W. 1940 Statistical Methods applied to Experiments in Agriculture and in Biology. Iowa State College Press, Ames.
(85) SNEDECOR, G. W., AND IRWIN, M. R. 1933 On the Chi-square test for homogeneity. Iowa State College J. Sci., **8**, 75–81.
(86) SOPER, H. E. 1914 Tables of Poisson's exponential binomial limit. Biometrika, **10**, 25–35.
(87) STEIN, M. F. 1919 The interpretation of *B. coli* test results on a numerical and comparative basis. J. Bact. **4**, 243–265.
(88) STEIN, M. F. 1922 The bacteria coli test. Relation between positive results in samples of one and ten cubic centimeters. Eng. Contr., **57**, 445–446.
(89) "STUDENT" (W. S. GOSSET) 1907 On the error of counting with a haemacytometer. Biometrika, **5**, 351–360.
(90) "STUDENT" 1908 The probable error of a mean. Biometrika, **6**, 1–25.
(91) SUPIŃSKA, J. 1934 Comparison of the precision with which one can determine the number of microorganisms in a unit of volume by means of certain modifications of the dilution method. (Polish, with French summary). Med. Doświadczalna i Spoleczna 18.
(92) SUTHERLAND, M., AND JAMES, N. 1938 The accuracy of the plate count of suspensions of pure cultures of bacteria in sterile soil. Can. J. Research, **160**, 305–312.
(93) SWAROOP, S. 1938 Numerical estimation of *B. coli* by dilution method. Indian J. Med. Research, **26**, 353–378.
(94) SWAROOP, S. 1940 Error in the estimation of the most probable number of organisms by the dilution method. Indian J. Med. Research, **27**, 1129–1147.
(95) SWAROOP, S. 1941 A modification of the routine dilution tests and tables showing the most probable number of organisms and the standard error of this number. Indian J. Med. Research, **29**, 499–510.
(96) SWAROOP, S. 1941 A consideration of the accuracy of estimation of the most probable number of organisms by dilution test. Indian J. Med. Research, **29**, 511–521.
(97) TAM, R. K., AND WILSON, P. W. 1941 Respiratory enzyme systems in symbiotic nitrogen fixation. III. The dehydrogenase systems of *Rhizobium trifolii* and *Rhizobium leguminosarum*. J. Bact., **41**, 529–546.
(98) THOMPSON, C. M. 1941 Table of percentage points of the X^2 distribution. Biometrika, **32**, 187–191.
(99) THORNDIKE, F. 1926 Applications of Poisson's probability summation. Bell System Tech. J., **5**, 604–624.
(100) THORNE, D. W., NEAL, O. R., AND WALKER, R. H. 1936. Physiological studies on *Rhizobium*. VIII. The respiratory quotient. Arch. Mikrobiol., **7**, 477–487.
(101) TIPPETT, L. H. C. 1932 A modified method of counting particles. Proc. Roy. Soc. (London), **137A**, 434–446.
(102) VAUGHN, R. H., AND LEVINE, M. 1942 Differentiation of the "intermediate" colilike bacteria. J. Bact., **44**, 487–505.
(103) WANG, SHU-HSIEN 1941 A direct smear method for counting microscopic particles in fluid suspension. J. Bact., **42**, 297–319.
(104) WELCH, B. L. 1938 On tests for homogeneity. Biometrika, **30**, 149–158.
(105) WELCH, B. L. 1937 The significance of the difference between two means when the population variances are unequal. Biometrika, **29**, 350–362.
(106) WELLS, W. F. 1918 The geometrical mean as a *B. coli* index. Science, n.s., **47**, 46–48.
(107) WELLS, W. F. 1919 Basis of the geometrical mean as a *B. coli* index. Science, n.s., **49**, 400–402.
(108) WELLS, W. F. 1919 The bacteriological dilution scale and the dilution as a bacteriological unit. Am. J. Pub. Health, **9**, 664–667.
(109) WELLS, W. F. 1919 On a standard system of bacteriological dilutions. Am. J. Pub. Health, **9**, 956–959.

(110) WELLS, P. V., AND WELLS, W. F. 1922 A standard bacterial index. J. Am. Water Works Assoc., **9**, 502–527.

(111) WILCOXON, F., AND MCCALLAN, S. E. A. 1939 Theoretical principles underlying laboratory toxicity tests of fungicides. Contrib. Boyce Thompson Inst., **10**, 329–338.

(112) WILSON, E. B. 1927 Probable inference, the law of succession, and statistical inference. J. Am. Stat. Assoc., **22**, 209–212.

(113) WILSON, P. W. 1940 The Biochemistry of Symbiotic Nitrogen Fixation. The University of Wisconsin Press, Madison.

(114) WILSON, P. W., AND KULLMANN, E. D. 1931 A statistical inquiry into methods for estimating numbers of rhizobia. J. Bact., **22**, 71–90.

(115) WILSON, P. W., BURRIS, R. H., AND LIND, C. J. 1942 The dissociation constant in nitrogen fixation by *Azotobacter*. Proc. Natl. Acad. Sci., U. S., **28**, 243–250.

(116) WILSON, P. W., LEE, S. B., AND WYSS, O. 1941 Mechanism of symbiotic nitrogen fixation. V. Nature of inhibition by hydrogen. J. Biol. Chem., **139**, 91–101.

(117) WILSON, P. W., WENCK, P. AND PETERSON, W. H. 1933 A statistical study of nitrogen fixation by clover plants. Soil Sci., **25**, 123–143.

(118) WISHART, J. 1934 Statistics in agricultural research. J. Roy. Stat. Soc., Suppl. **1**, 26–61.

(119) WOLMAN, A., AND WEAVER, H. L. 1917 A modification of the McCrady method of the numerical interpretation of fermentation-tube results. J. Infectious Diseases, **21**, 287–291.

(120) WYSS, O., AND WILSON, P. W. 1941 Mechanism of biological nitrogen fixation. VI. Inhibition of *Azotobacter* by hydrogen. Proc. Natl. Acad. Sci., U. S., **27**, 162–168.

(121) WYSS, O., LIND, C. J., WILSON, J. B., AND WILSON, P. W. 1941 Mechanism of biological nitrogen fixation. VII. Molecular H_2 and the pN_2 function of *Azotobacter*. Biochem. J., **35**, 845–854.

(122) YATES, F. 1935 Some examples of biased sampling. Ann. Eugenics, **6**, 202–213.

(123) YULE, G. U. 1910 An Introduction to the Theory of Statistics. Chas. Griffin and Co., London. And subsequent editions—11th with M. G. Kendall in 1937.

(124) ZIEGLER, N. R., AND HALVORSON, H. O. 1935 Application of statistics to problems in bacteriology. IV. Experimental comparison of the dilution method, the plate count, and the direct count for the determination of bacterial populations. J. Bact., **29**, 609–634.

(125) ZUBIN, J. 1935 Note on a transformation function for proportions and percentages. J. Applied Psychol., **19**, 213–220.