ON TESTS OF THE SIGNIFICANCE OF DIFFERENCES IN DEGREE OF POLLUTION BY COLIFORM BACTERIA AND ON THE ESTIMATION OF SUCH DIFFERENCES

By H. J. BUCHANAN-WOLLASTON

Principal Naturalist on the Staff of the Ministry of Agriculture and Fisheries

(*Working at the Laboratory of the Freshwater Biological Association, Wray Castle, Ambleside*)

(With 3 Figures in the Text)

CONTENTS

Preface

It is not possible, within the bounds of a short paper intended for practical bacteriologists, to explain fully the difficult fundamental theory of modern statistical tests of significance.

The sections of this paper are necessarily unequal as to difficulty. It is suggested that bacteriologists who wish to apply statistical tests to their data but who have not had sufficient statistical training to understand their theory completely should omit some of the sections on first reading, concentrating on sections which deal with the problems most likely to arise in practice.

The *Introduction* is important since it comprises a short discussion on the concept of the 'null hypothesis' on which are based all statistical tests of significance.[1] Without clear conception of the particular null hypothesis

[1] See Fisher (1935).

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

involved in a statistical test and of the meaning of 'deviation from expectation based on the truth of that null hypothesis an investigator cannot possibly know what he is testing for significance.

The last two paragraphs of § 1 should be read in conjunction with Table 4 and the use of Table 3 should be practised, so that any observed difference in the total number of fertile tubes between any pair of samples may be given the corresponding value of P from Table 3. P should be considered as having a meaning corresponding to that which it has in the case of the χ^2-test, t-test, z-test or normal test of the significance of the difference between two means. § 5 should also be read.

§ 2 is easy to understand, and in it is explained a method which is very useful in testing the significance of a series of differences in degree of pollution by coliform bacteria. § 4 should certainly be mastered. Its methods are sure to be needed. § 7 is important and should be read in conjunction with Chapter 5 of Fisher (1934). The methods described are easy to apply. Table 2 of Pearson (1930) is necessary. § 9 may be considered as an appendix to § 7 and should be read in conjunction with it. The methods of §§ 7 and 9 are applicable to a very large number of practical problems. § 8 contains a suggestion as to sampling.

The method described in § 3 is of interest when a long series of differences is being examined. The theory is rather more complicated than that of the other methods described.

Introduction

It will be assumed that the degree of pollution of a sample of water by coliform bacteria is estimated in the usual way from the number of positive tubes of MacConkey lactose broth after inoculation and incubation at a suitable temperature.

It is often more important to be able to decide whether two samples of water are significantly different in degree of pollution than it is to make an estimate of the degree of pollution pertaining to each of them. Sometimes also it is of practical importance to be able to make a fairly accurate estimate of a difference in degree of pollution. This is particularly the case when it is wished to compare the results of two methods of treatment of the inoculated tubes. Here, probably, we have to deal with a long series of paired counts, and it may be useful to be able to calculate a factor from these to be applied to future estimates of pollution deduced from the results of one method so as to obtain approximately the results which the other would have given.

The problems of estimating a difference in pollution and of estimating its significance are different problems. It is not even necessary that the difference should be measured in the same way in the two cases. Indeed, the test of significance of a series of differences may depend only on a test of consistency in sign, this involving no measure of the difference.

The test of the significance of a difference between two samples is a test of the acceptability of the 'null hypothesis' that the two are identical as to pollution. This hypothesis is acceptable only if the two samples may be regarded

reasonably as having been taken from the same body of water, namely, a mixture of equal parts of the two kinds of water compared, the observed difference being such as might often occur by chance in that case. The criterion of acceptability may be defined as the probability that, *if the null hypothesis were true*, a difference at least as great as that observed would occur. No particular method of measuring the difference is implicit in this concept, but the possibility of determining the value of the criterion in the case of a single pair depends on whether the random sampling distribution of the chosen measure can be determined. If this is possible the case is simple when only two samples are to be compared, for here only one kind of difference is involved, namely, that in degree of pollution, however this be measured. When, however, we make a series of comparisons, expectation implied by the assumption that the null hypothesis is true in the case of every pair takes a more complicated form. We have, then, an expected *distribution* of differences, and marked deviation of the observed distribution of differences from that expected, of whatever kind that deviation might be, would indicate that acceptance of the truth of the null hypothesis that the members of each pair were taken from the same body of water was unreasonable. The null hypothesis may take other forms. For example, let us assume that we have a long series of paired sets of inoculated tubes, the members A and B of each pair having actually been taken from the same bottle of water, all the members A, however, having been treated by a method of incubation different from that used for the members B. Here the null hypothesis takes the form, 'The difference in treatment has had no effect, the observed distribution of differences agrees well with the expected distribution'. Deviation from expectation may take many forms. It may be that the members A show consistently greater pollution than the members B. There may be too many large differences or too many small differences without regard to sign. Perhaps there is an undue degree of skewness. It is generally possible to determine the type of deviation met with and to test separately the significance of such components of deviation. Often these are of more interest than general unspecified deviation from the 'expected' distribution.

The necessity, in testing the acceptability of the null hypothesis with single pairs, for exact determination of the expected distribution has led me to reject the measure of difference in degree of pollution generally used, namely, the difference in the so-called most probable number of bacteria per 100 ml., or M.P.N., and to employ as a measure of deviation from identity in degree of pollution the total difference in the number of fertile tubes, which I shall designate N.F.T. The random sampling distribution of the difference in N.F.T. is readily determinable for the numbers of tubes generally applied, the laws of chance involved being particularly simple.[1] The use of this measure has other advantages also when compared to any measure which depends on estimates of M.P.N. Thus, no estimate even approximately correct of a difference

[1] For further discussion on theory involved see § 8.

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

in M.P.N. can be made if, in one of the compared sets, all the tubes are fertile or all sterile. This is a common occurrence. Further, sets of tubes in which the dilution technique varies cannot be combined in a composite test if M.P.N. be the measure in use. Omission of cases of these two kinds need not be made if difference in N.F.T. be our measure of deviation for the purpose of estimating the value of our criterion. The determination of the expected distribution is dealt with in our next section.

1. Calculation of the probability of any total difference in N.F.T.

We shall consider first the case of two sets of five tubes at a single dilution.

Though the distribution of numbers of bacteria in a tube should follow Poisson's law, yet, when an observation can take only two forms, sterile and fertile, and the total number of the two together is limited to 10, it is clear that the random sampling distribution of the proportion of sterile tubes, obtained by inoculating the ten tubes from the common hypothetical sample, is given by the terms in the expansion of the binomial $(q+p)^{10}$, where p is the chance of a sterile tube. The relation between p and m, where m stands for M.P.N., is given by the equation

$$p = e^{-am},$$

in which a is the fraction of 100 ml. of water with which a tube is inoculated and $q = 1-p$. The binomial specifies the type of distribution from which our two counts have originated. We do not know the value of p for certain, but an estimate of p, the only estimate available, is given by the observed proportion of sterile tubes in the two samples combined. We now have to distribute the ten tubes into two sets of five and find out in how many different ways this may be done so that each particular difference in number of fertile tubes occurs. The proportional number of ways in which this distribution may be performed so as to give a particular difference is the probability of that difference. In this enquiry the value of p must be assumed to be invariable, in accordance with the generally accepted statistical rule as to 'degrees of freedom'. Only those pairs of samples in which the total number of sterile tubes is $10p$ must be included in the distribution. Though the action of chance is limited to mere redistribution of the tubes we have into two sets, the easiest procedure for calculating the probability of obtaining any given difference is to assume that the two sets are independent except that the total number of sterile tubes in the two is $10p$. Thus the proportional number of times any particular number of sterile tubes in one set will occur with any particular number of sterile tubes in the second set will be given by the appropriate product term in the expansion of the product

$$(q+p)^5 (q+p)^5,$$

only those terms being included which correspond to the aforesaid total number of sterile tubes.

Let us consider the case when the total number of sterile tubes in the two sets is 4. We write down the terms under each other with the possible differences over the top. Each product is written under the difference to which it applies, and no term is written down for which the indices of p and q are other than 4 and 6 respectively. The plan is shown in Table 1.

The relative probabilities of differences of 0, ± 2 and ± 4 steriles are therefore equal, respectively, to

$$100q^6p^4, \quad 100q^6p^4 \quad \text{and} \quad 10q^6p^4.$$

To obtain the absolute probabilities these results must be divided by the probability of obtaining 4 steriles in ten tubes, namely,

$$\frac{10!}{4!\,6!}q^6p^4 \quad \text{or} \quad 210q^6p^4.$$

Table 1. *Differences*

0	±1	±2	±3	±4	±5
$1q^5$	$5q^4p$	$10q^3p^2$	$10q^2p^3$	$5qp^4$	$1p^5$
$1q^5$	$5q^4p$	$10q^3p^2$	$10q^2p^3$	$5qp^4$	$1p^5$
$100q^6p^4$		$50q^6p^4$		$5q^6p^4$	
		$50q^6p^4$		$5q^6p^4$	
$100q^6p^4$		$100q^6p^4$		$10q^6p^4$	

Table 2

	d ... 0	±1	±2	±3	±4	±5
2	$0\dot{5}$	—	$0\dot{4}$	—	—	—
3	—	$0\dot{8}\dot{3}$	—	$0\dot{1}\dot{6}$	—	—
4	0·47619	—	0·47619	—	0·04762	—
5	—	0·79365	—	0·19841	—	0·00794
6	0·47619	—	0·47619	—	0·04762	—
7	—	$0\dot{8}\dot{3}$	—	$0\dot{1}\dot{6}$	—	—
8	$0\dot{5}$	—	$0\dot{4}$	—	—	—

The probabilities of differences, in number of sterile tubes, of 0, 2 and 4 are thus respectively equal to

$$\frac{100}{210}, \quad \frac{100}{210} \quad \text{and} \quad \frac{10}{210} \quad \text{or} \quad 0{\cdot}4762, \quad 0{\cdot}4762 \quad \text{and} \quad 0{\cdot}04762.$$

It will be seen that the value of p does not enter into the calculation, its only effect being that of limiting the totals. It is a very simple matter to calculate the probabilities of all possible differences when there are 2, 3, 4, 5, 6, 7 or 8 steriles in the ten tubes. The probability of a given difference is exactly the same with a total of s steriles in n as it is with a total of $n-s$ steriles in n, and therefore it does not matter whether a difference is considered as applying to fertile tubes or to sterile tubes. We shall henceforth consider that the difference applies to fertile tubes. The probabilities of the various differences are given in Table 2 for two parallel sets of five tubes each. In Table 2, s stands for the sum of the numbers of fertile tubes in the two sets of five, d is the difference, in number of fertile tubes, between the two sets,

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

and the entries give p, the probability of $\breve{d}$. Any difference with no corresponding entry for p is impossible under the restrictions imposed by theory.

Table 3

Class	Sums S_1	S_2	S_3	Probability of total difference in N.F.T. equal to or greater than: 0 or 1	2 or 3	4 or 5	6 or 7	8 or 9	10 or 11	12 or 13	14 or 15
a	0	0	1	1							
b	0	0	2	1	0·$\dot{1}$						
c	0	0	3	1	0·1$\dot{6}$						
d	0	0	4	1	0·52381	0·04762					
e	0	0	5	1	0·20635	0·00794					
f	0	1	1	1	0·5						
g	0	1	2	1	0·$\dot{2}$						
h	0	1	3	1	0·58$\dot{3}$	0·08$\dot{3}$					
i	0	1	4	1	0·28571	0·02381					
j	0	1	5	1	0·60317	0·10714	0·00397				
k	0	2	2	1	0·59260	0·09876					
l	0	2	3	1	0·31481	0·03704					
m	0	2	4	1	0·62963	0·14285	0·01058				
n	0	2	5	1	0·33862	0·05026	0·00176				
o	0	3	3	1	0·63889	0·15278	0·01389				
p	0	3	4	1	0·36111	0·06349	0·00397				
q	0	3	5	1	0·65277	0·17261	0·02049	0·00066			
r	0	4	4	1	0·65874	0·18254	0·02380	0·00112			
s	0	4	5	1	0·38096	0·07671	0·00680	0·00019			
t	0	5	5	1	0·66534	0·19135	0·02758	0·00160	0·00003		
u	1	1	1	1	0·25						
v	1	1	2	1	0·6$\dot{1}$	0·$\dot{1}$					
w	1	1	3	1	0·$\dot{3}$	0·041$\dot{6}$					
x	1	1	4	1	0·64286	0·15476	0·01190				
y	1	1	5	1	0·35514	0·05551	0·00194				
z	1	2	2	1	0·34568	0·04938					
A	1	2	3	1	0·65741	0·17593	0·01852				
B	1	2	4	1	0·38624	0·07672	0·00529				
C	1	2	5	1	0·66931	0·19444	0·02601	0·00088			
D	1	3	3	1	0·39584	0·08334	0·00695				
E	1	3	4	1	0·68055	0·21229	0·03372	0·00198			
F	1	3	5	1	0·41269	0·09655	0·01057	0·00033			
G	1	4	4	1	0·42064	0·10317	0·01246	0·00056			
H	1	4	5	1	0·69048	0·22883	0·04175	0·00349	0·00009		
I	1	5	5	1	0·42835	0·10947	0·01459	0·00082	0·00002		
J	2	2	2	1	0·66392	0·18656	0·02195				
K	2	2	3	1	0·40535	0·09054	0·00823				
L	2	2	4	1	0·68606	0·22162	0·03762	0·00235			
M	2	2	5	1	0·42153	0·10357	0·01215	0·00039			
N	2	3	3	1	0·69136	0·22994	0·04167	0·00309			
O	2	3	4	1	0·43695	0·11640	0·01632	0·00088			
P	2	3	5	1	0·70039	0·24551	0·04989	0·00492	0·00015		
Q	2	4	4	1	0·70459	0·25309	0·05404	0·00591	0·00025		
R	2	4	5	1	0·45090	0·12878	0·02086	0·00162	0·00004		
S	2	5	5	1	0·70875	0·26029	0·05820	0·00703	0·00037	<0·00001	
T	3	3	3	1	0·44445	0·12270	0·01853	0·00116			
U	3	3	4	1	0·70899	0·26057	0·05819	0·00694	0·00033		
V	3	3	5	1	0·45789	0·13492	0·02321	0·00199	0·00006		
W	3	4	4	1	0·46429	0·14096	0·02560	0·00245	0·00009		
X	3	4	5	1	0·71671	0·27459	0·06656	0·00931	0·00065	0·00002	
Y	3	5	5	1	0·47048	0·14680	0·02810	0·00297	0·00014	<0·00001	
Z	4	4	4	1	0·72034	0·28139	0·07077	0·01056	0·00084	0·00003	
α	4	4	5	1	0·47647	0·15266	0·03061	0·00354	0·00021	<0·00001	
β	4	5	5	1	0·72389	0·28791	0·07491	0·01188	0·00104	0·00004	<0·00001
γ	5	5	5	1	0·48228	0·15833	0·03319	0·00413	0·00027	<0·00001	<0·00001

Note. The odd differences apply to letter classes in which $S_1+S_2+S_3$=an odd number.

When three sets of five paired tubes at three different dilutions are dealt with the calculation of the probabilities of all possible total differences, though

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

carried out by a simple extension of the process shown for the case of a single set of five pairs, is rather a tedious business. I have, however, done the necessary calculations and the results are incorporated in Table 3. An entry in Table 3 does not, however, give p, the probability of the corresponding difference itself, but P, the probability that the difference is *at least as great* as that shown in the top row vertically above the entry. Each entry thus includes the probability of the corresponding difference and the probabilities of all greater differences possible under the restriction that the sum of the numbers of fertile tubes in the two members of each pair at each different dilution is as given under S_1, S_2 and S_3 in the columns on the left. The chance of a given difference when any sum is equal to 1 is equal to that when the sum is equal to 9, 2 corresponds to 8, 3 to 7 and 4 to 6. Therefore no sum greater than 5 occurs in the table of sums. The use of the table may be best explained by way of examples and I shall use those given in Table 4.

Table 4

	10 c.c.	1 c.c.	0·1 c.c.	P
		Example 1		
Sample A	0	0	1	0·346
Sample B	1	2	1	
Sum	1	2	2	
		Example 2		
Sample A	0	1	1	0·04989
Sample B	3	4	1	
Sum	3	5	2	
		Example 3		
Sample A	5	2	2	0·36111
Sample B	5	5	2	
Sum	10	7	4	

In Example 1 the sums are 1, 2 and 2 and d, the total difference, is 3. Entering Table 3 at $S_1=1$, $S_2=2$, $S_3=2$, we find that $P=0{\cdot}346$ approximately. Thus at least as great a difference will occur by chance rather more than once in three trials when there is no real difference. In Example 2 the sums are 3, 5 and 2. In Table 3 the sums are placed in order of magnitude, and we therefore enter the table at 2, 3 and 5 and find for the difference of 6 that $P=0{\cdot}04989$. The difference is therefore significant if judged by the usual standard.[1] For Example 3 we have sums of 10, 7 and 4 or 0, 3, 4 and $P(\pm 3)=0{\cdot}36111$.

2. On tests of the significance of a series of differences between paired samples. Binomial test of significance of consistency in sign

Let us designate the subsamples on one side of each pair by A, those on the other side by B. If we have a long series of paired samples which do not show many significant differences when tested separately by way of Table 3 but in which the difference, $B-A$, is positive in the great majority of cases

[1] See § 5.

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

we may test whether the two sets, A and B, may reasonably be considered to be equally polluted or, in other cases, whether two methods of treatment may be considered as having given the same results, by finding out the probability of at least as great an excess in positive or negative signs as that observed. If n stands for the number of pairs, we have to sum the terms of the binomial $(\frac{1}{2}+\frac{1}{2})^n$ beyond and including the terms corresponding to the observed excess. Both tails of the binomial must be summed. The difference considered may be that between values of M.P.N. or that between values of N.F.T. I think that in carrying out the binomial test of signs by itself it is preferable to consider the difference in M.P.N.[1]

Dr L. F. L. Clegg, a member of the staff of the Ministry of Agriculture and Fisheries at the research station at Conway, has kindly supplied me with data suitable for illustrating the application of various statistical methods including the binomial test. Dr Clegg's data refer to the results of two different methods of treatment of water samples of which the degree of pollution by faecal coliform bacteria is being investigated. Each pair of samples, A and B, was taken from the same body of water, the tubes for A being incubated immediately at 44° C., while the tubes B were incubated first at 37° C., further tubes being inoculated from those proving fertile at 37° C. and incubated at 44° C. We have the results of the comparison of 353 pairs, if pairs with the same M.P.N. on each side be omitted. We have therefore to sum the appropriate terms of the binomial $(\frac{1}{2}+\frac{1}{2})^{353}$. When n is fairly large the normal distribution is a sufficiently close approximation to the binomial, particularly when the binomial is symmetrical as in the present case. Table 13 gives the probability of at least as great an excess in positive or negative signs as that shown in the uppermost horizontal row, with values of n from 3 to 20. For values of n greater than 20 the normal distribution may be used except in cases near the border line of significance. In border-line cases with n between 20 and 30 a correction for continuity should be applied to allow for the fact that the normal distribution is continuous while the binomial is discontinuous, there being a finite probability for each occurrence in the binomial while, in a continuous distribution, the probability of any exact occurrence x is infinitesimal. Thus x should be considered as being at the inner limit of the finite probability beyond and including which the total is required. For a deviation of 2 from the mean, x is taken as $1\frac{1}{2}$, 3 is replaced by $2\frac{1}{2}$, 71·5 by 71, and so on. The standard deviation of the binomial is equal to $\sqrt{(npq)}$, which in the present case is equal to $\sqrt{\frac{353}{4}}$ or 9·4 approximately. The mean of the distribution is 176·5. It was found that the method used for samples B gave the greater estimate of M.P.N. in 248 cases and the lesser estimate in 105 cases. The term 248, 105 is distant 71·5 units from the mean. Thus $\frac{x}{\sigma}=\frac{71{\cdot}5}{9{\cdot}4}$, which is greater than 7. From a table of the normal error function[2] we find that the probability

[1] See § 7. [2] Pearson (1930).

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

of as great a difference as that observed is less than 10^{-11}, the two 'tails' of the normal integral being added together to obtain this probability. We may say that, practically speaking, it is impossible that the two methods of treatment have given the same results and that the observed difference is a chance effect.

An objectionable feature in the binomial test is that clearly some relevant information contained in the data is not used in the test. In the case of Dr Clegg's data, since the result is so definite, objection to the use of the binomial or normal distribution could be academic only. Sometimes, however, it might happen that, though samples *B* showed slightly the greater M.P.N. in a majority of cases, yet, in the others, the M.P.N. in samples *A* was considerably the greater. Here the result of the sign test if used alone might be misleading. The test of signs has, however, in common with the test of significance of a difference in N.F.T., one great advantage when compared with any test in which the magnitude of each difference in M.P.N. is taken into account, namely, that no pair in which the M.P.N. differs in *A* and *B* need be omitted from the test. In a later section I discuss the distribution of the difference between logarithms of M.P.N. which, in the case of Dr Clegg's data, proved to be approximately normal in form and therefore the normal distribution may be used in a test of the significance of the mean difference in the logarithms. The use of the logarithms in this way, however, necessitates the omission of all pairs in which either member had no positive tubes or 15 positive tubes. These clearly give information relevant to the question whether one set shows significantly greater pollution than the other, though the difference cannot be expressed logarithmically or accurately measured. Dr Clegg's data include the results from 129 such pairs.

It is obvious that an excess in positive signs is correlated with the difference giving rise to this excess, and it would doubtless be possible to work out a scale by which the value of the difference could be calculated from the excess. In the present case, however, it is preferable to estimate the mean difference by other methods, as applied in a later section.

If we are only interested in the significance of a mean difference, and if significance of this can be shown by way of the binomial test of signs, it would be waste of time to carry out further tests of significance. Often, however, in the case of a series of differences, it is of interest to find out whether there is any undue degree of discrepancy between an observed distribution of differences and that expected on the assumption of the truth of the null hypothesis. Such discrepancy does not always give rise to a significant mean difference. Again, in comparing the results of one method with that of another, it may be suspected that the effect of the difference in method is very uneven in its action. A general test of discrepancy is described in the next section.

3. On a general test of discrepancy between an observed distribution of differences and the expected distribution

Table 5 has been prepared to facilitate the application of a test of the significance of general discrepancy between observation and expectation in the case of a series of differences in N.F.T. Each entry in the table gives the probability of the difference, with particular sign, which is entered in the top row vertically above the entry, except in the case of a difference of zero, to which, of course, no sign is applicable. Table 5 gives p, the probability of the corresponding difference itself, and not P, the probability of at least as great a difference. The letters in the left-hand column are to facilitate reference to the conditions of possible variation applicable to cases in which the sums of fertile tubes in the two sets at three dilutions are those given in the next three columns under S_1, S_2, S_3. When a series of paired samples is being dealt with, against the data for each pair is written the appropriate letter and also the difference in N.F.T. Table 6 shows the procedure. The entries in Table 5 corresponding to each letter are then multiplied by the number of times that letter occurs in the data and the results entered under the corresponding differences, each vertical column of figures being then summed to give the expectation of that difference, positive, negative or zero, to which the column applies. I have used this method in examining Dr Clegg's data relating to water samples from wells and piped supplies, including chlorinated water. Samples A and B were treated respectively according to the two methods already described. There were 266 pairs of observations, and twenty-nine different conditions of variation occurred, that represented by the letter a being far the commonest. The quantities of the original water sample used in the case of chlorinated water were one tube of 50 c.c. and five tubes inoculated with 10 c.c. and five tubes inoculated with 1 c.c. For the other kinds of water there were five tubes at each dilution for both A and B, the inoculating quantities being respectively 10, 1 and 0·1 c.c. The data have been combined, justifiably, since the method used for limiting expected variation allows for variation in numbers of tubes and in degree of dilution.[1] The letters a, b and c occurred 69, 27 and 23 times respectively. The first three entries in the table of expectation were therefore as shown in Table 7, $B-A$ being considered a positive difference.

The final result was as shown in Table 8, differences greater than 4 being combined at each end of the distribution.

The result is shown graphically in Fig. 1, in which continuous lines join points corresponding to expected frequencies, broken lines pertain to observed frequencies. It will be seen that there is a very strong bias towards positive differences. The χ^2-test (Fisher, 1934) may be applied to see whether the deviation from expectation is significant; it should not, however, be applied

[1] When only one tube is used at a particular dilution in each set, $S=1$, 0 or 2. If $S=2$, Tables 3 and 5 are entered at $S=0$ for the dilution in question.

Table 5. *Probability of each total difference in* N.F.T.

Sums			Total difference							
S_1	S_2	S_3	0 or 1	2 or 3	4 or 5	6 or 7	8 or 9	10 or 11	12 or 13	14 or 15
0	0	1	0·5							
0	0	2	0·5	0·2						
0	0	3	0·416	0·083						
0	0	4	0·47619	0·23810	0·02380					
0	0	5	0·39683	0·09920	0·00397					
0	1	1	0·5	0·25						
0	1	2	0·38	0·1						
0	1	3	0·416	0·25	0·0416					
0	1	4	0·35715	0·13095	0·01190					
0	1	5	0·39683	0·24802	0·05158	0·00198				
0	2	2	0·40740	0·24692	0·04938					
0	2	3	0·34259	0·13889	0·01852					
0	2	4	0·37037	0·24339	0·06614	0·00529				
0	2	5	0·33069	0·14418	0·02425	0·00088				
0	3	3	0·36111	0·24306	0·06945	0·00694				
0	3	4	0·31945	0·14881	0·02976	0·00198				
0	3	5	0·34723	0·24008	0·07606	0·00991	0·00033			
0	4	4	0·34126	0·23810	0·07937	0·01134	0·00056			
0	4	5	0·30953	0·15213	0·03495	0·00330	0·00009			
0	5	5	0·33466	0·23699	0·08189	0·01299	0·00079	0·00001		
1	1	1	0·375	0·125						
1	1	2	0·38	0·25	0·05					
1	1	3	0·3	0·14583	0·02083					
1	1	4	0·35715	0·24405	0·07143	0·00595				
1	1	5	0·32243	0·14982	0·02678	0·00097				
1	2	2	0·32716	0·14815	0·02469					
1	2	3	0·34259	0·24074	0·07870	0·00926				
1	2	4	0·30688	0·15476	0·03571	0·00265				
1	2	5	0·33069	0·23744	0·08421	0·01257	0·00044			
1	3	3	0·30208	0·15625	0·03820	0·00347				
1	3	4	0·31945	0·23413	0·08929	0·01587	0·00099			
1	3	5	0·29366	0·15807	0·04299	0·00512	0·00016			
1	4	4	0·28968	0·15874	0·04535	0·00595	0·00028			
1	4	5	0·30952	0·23083	0·09354	0·01913	0·00170	0·00004		
1	5	5	0·28583	0·15944	0·04744	0·00688	0·00040	0·00001		
2	2	2	0·33608	0·23868	0·08231	0·01097				
2	2	3	0·29732	0·15741	0·04115	0·00412				
2	2	4	0·31394	0·23222	0·09200	0·01763	0·00118			
2	2	5	0·28924	0·15898	0·04571	0·00588	0·00019			
2	3	3	0·30864	0·23071	0·09414	0·01929	0·00154			
2	3	4	0·28153	0·16027	0·05004	0·00772	0·00044			
2	3	5	0·29961	0·22744	0·09781	0·02249	0·00239	0·00007		
2	4	4	0·29541	0·22575	0·09952	0·02407	0·00283	0·00013		
2	4	5	0·27455	0·16106	0·05396	0·00962	0·00079	0·00002		
2	5	5	0·29125	0·22423	0·10105	0·02558	0·00333	0·00018	<0·00001	
3	3	3	0·27	0·16088	0·05209	0·00868	0·00058			
3	3	4	0·29101	0·22421	0·10119	0·02563	0·00330	0·00016		
3	3	5	0·27105	0·16149	0·05586	0·01061	0·00096	0·00003		
3	4	4	0·26786	0·16166	0·05768	0·01157	0·00118	0·00005		
3	4	5	0·28329	0·22106	0·10402	0·02862	0·00433	0·00031	0·00001	
3	5	5	0·26476	0·16184	0·05935	0·01256	0·00142	0·00007	<0·00001	
4	4	4	0·27966	0·21948	0·10531	0·03010	0·00486	0·00041	0·00001	
4	4	5	0·26177	0·16191	0·06102	0·01353	0·00167	0·00010	<0·00001	
4	5	5	0·27610	0·21799	0·10650	0·03152	0·00542	0·00050	0·00002	<0·00001
5	5	5	0·25886	0·16198	0·06257	0·01453	0·00193	0·00013	<0·00001	<0·00001

Note. The odd differences apply to letter classes in which $S_1+S_2+S_3=$an odd number.

Table 6

A. Fertile			B. Fertile				
10 c.c.	1 c.c.	0·1 c.c.	10 c.c.	1 c.c.	0·1 c.c.	d	
3	1	0	4	0	0	0	h
5	1	1	5	1	0	−1	g
0	0	0	1	0	0	1	a
5	2	0	5	5	1	4	h
1	0	0	0	0	0	−1	a
5	2	0	3	0	0	−4	k

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

to the whole distribution for the following reason. In calculating the conditions of variation represented by the letters in Table 5 a degree of freedom is absorbed from each pair of variates. No other restriction is imposed on the possible deviation from expectation in each letter class other than that due to the total expected and the total observed being the same. The variation

Table 7

Difference ...	−(>4)	−4	−3	−2	−1	0	+1	+2	+3	+4	>+4
Letter class:											
$a \times 69$	—	—	—	—	34·5	—	34·5	—	—	—	—
$b \times 27$	—	—	—	6	—	15	—	6	—	—	—
$c \times 23$	—	—	1·916	—	9·583	—	9·583	—	1·916	—	—
Sum	—	—	1·916	6	44·083	15	44·083	6	1·916	—	—

Table 8

Difference ...	−(>4)	−4	−3	−2	−1	0	+1	+2	+3	+4	>+4
Expected (f)	0·9	2·9	10·2	25·6	68·6	49·7	68·6	25·6	10·2	2·9	0·9
Observed (f)	1	4	4	14	45	41	85	34	23	11	4

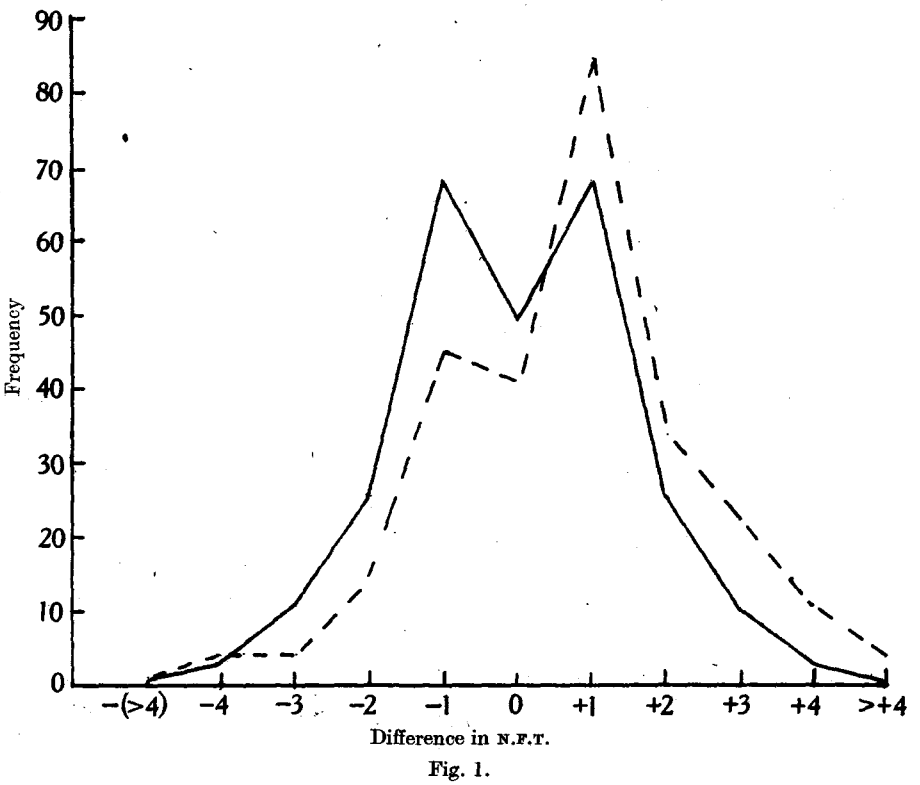

Fig. 1.

in each letter class is thus multinomial, and this implies that any frequency class may contain all the frequencies. Therefore only those letter classes may be combined which contain the same frequency classes. If, for example, we were to combine the letter class a with class b, the differences in class a, which are limited to $+1$ and -1, cannot fall in class b in which the only differences

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

are 0, +2 and −2. This combined distribution would not therefore be multinomial and the χ^2-test, with 4 degrees of freedom, could not be applied. The correct procedure is shown in Table 9 in which letter classes having the same frequency classes are combined. Table 9 includes expected and observed frequencies and the corresponding values of χ^2, with the number of degrees

Table 9

Letter class *a*

Difference ...	−1	+1
Expected (*f*)	34·5	34·5
Observed (*f*)	20	49
χ^2	6·0942	6·0942

Total $\chi^2 = 12{\cdot}1884$. D.F. = 1.

Letter classes *b*, *f*

Difference ...	−2	0	+2
Expected (*f*)	9·25	21·5	9·25
Observed (*f*)	4	20	16
χ^2	2·9797	0·1046	4·9257

Total $\chi^2 = 8{\cdot}0100$. D.F. = 2.

Letter classes *c*, *g*, *u*

Difference ...	−3	−1	+1	+3
Expected (*f*)	4·5972	18·9027	18·9027	4·5972
Observed (*f*)	1	16	19	11
χ^2	2·8147	0·4457	0·0005	8·9173

Total $\chi^2 = 12{\cdot}1782$. D.F. = 3.

Letter classes *d*, *h*, *k*

Difference ...	−4	−2	0	+2	+4
Expected (*f*)	1·6975	11·9877	21·6296	11·9877	1·6975
Observed (*f*)	3	8	15	14	9
χ^2	3·1231	1·3265	2·0320	0·3378	31·4150

Total $\chi^2 = 38{\cdot}2344$. D.F. = 4.

Other odd classes

Difference ...	−(>3)	−3	−1	+1	+3	>+3
Expected (*f*)	0·7266	5·5993	15·1740	15·1740	5·5993	0·7266
Observed (*f*)	1	3	9	17	11	2
χ^2	0·8552		2·5179	0·19611	7·0577	

Total $\chi^2 = 10{\cdot}6269$. D.F. = 3.

Other even classes

Difference ...	−ve	0	+ve
Expected (*f*)	5·7157	6·5692	5·7157
Observed (*f*)	4	6	8
χ^2	0·7359		

Total $\chi^2 = 1{\cdot}6981$. D.F. = 2.

Grand total of $\chi^2 = 82{\cdot}9360$. D.F. = 15.

Note. So as to exhibit particularly large deviations from expectation classes have been included in which the expectation is small. This may have exaggerated the significance of the value of χ^2 slightly.

of freedom applicable to each case. These values of χ^2 and numbers of degrees of freedom may respectively be summed for a composite test. A total value of χ^2 equal to 82·9 with 16 degrees of freedom is outside the range of the published χ^2 tables, but, in any case, P is less than 10^{-6}. A highly significant discrepancy from expectation is therefore shown.

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

It should be repeated that the comprehensive test detailed is a test of the significance of all kinds of deviation from expectation. That the main element in the discrepancy is due to a difference between means is shown graphically in Fig. 2, in which the observed frequency distribution is made symmetrical by summing the observed frequencies of +3 and −3, +2 and −2, and so on and distributing half the sum in each case equally on each side of zero difference which remains the same. This reduces the discrepancy very greatly. It should be understood, however, that though a calculated value of χ^2 may be used here as a measure of deviation from expectation the χ^2-*test* is no longer

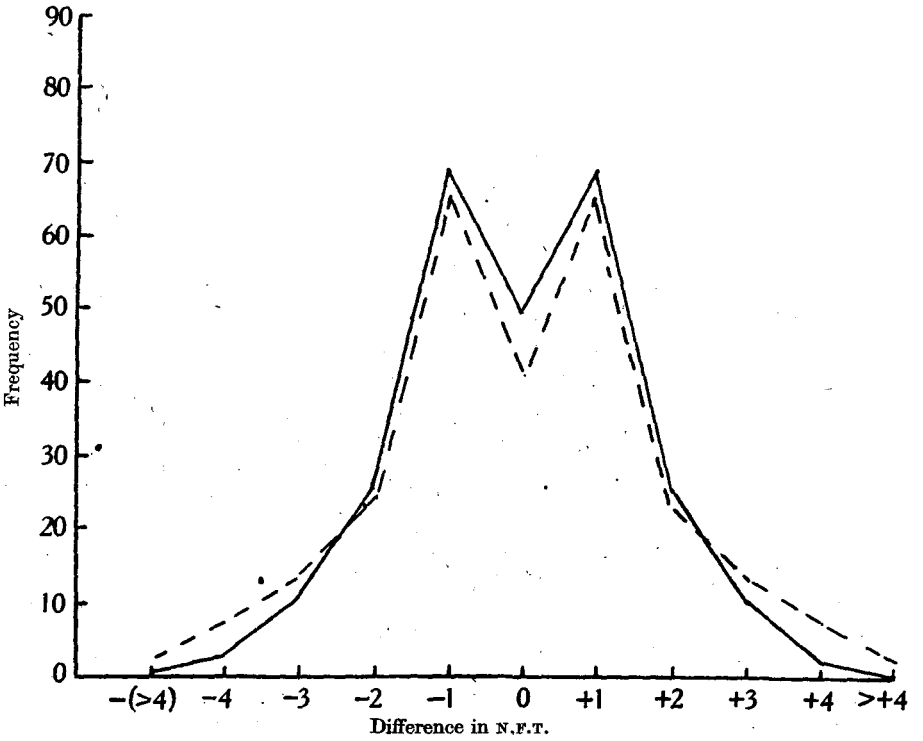

Fig. 2.

applicable. The χ^2-*distribution* would be realized only if the null hypothesis were true. We have no knowledge of the distribution of χ^2 when used as a measure of discrepancy from the null hypothesis when it is not true. By way of our original test we have been led to the conclusion that this hypothesis is very unlikely to be true, that it is entirely reasonable to assume that it is not true. If we wish to test the significance of any particular kind of discrepancy we must first formulate a null hypothesis which might be true even when the original hypothesis is not true. General discrepancy in the distribution does not necessarily imply difference between the means or difference between the standard deviations of the two series compared, therefore the

significance of these may be tested separately. We have already applied the binomial distribution in a test of the significance of the difference between means. Further tests for this are described in the next section, in which, however, it is differences in the logarithms of M.P.N. which are considered, and, as already explained, the use of these as variates entails the rejection of much relevant information. There appears to be only one kind of test which is applicable to the problem of the mean difference in N.F.T. and which makes use of all the relevant information on that question contained in the data. This is described in § 7, in which also examples are given of its application, and its philosophical basis is compared with that of another statistical test.

The question of the homogeneity of the components of the excess in positive signs to which we have already applied the binomial test may sometimes be of interest. We may ask, Is the excess in positive signs on the side of samples *B regular*, the observed irregularity being such as might be expected to arise by chance? Here the test of homogeneity described by Fisher (1934)

Table 10

B	Raw piped	Filtered	Raw wells	Chlorinated	Raw sea, lakes and reservoirs	Raw streams, springs and surface	Total
+	52	58	25	34	15	64	248
−	27	18	11	11	8	30	105
Sum	79	76	36	45	23	94	353

$\frac{a}{a+a}$, 0·3418, 0·2369, 0·3056, 0·2444, 0·3478, 0·3192 $= p$.

$\bar{p} = 0{\cdot}2975$, $\bar{q} = 0{\cdot}7025$.

ap ... 9·230, 4·265, 3·362, 2·688, 2·783, 9·576.

$S(ap) = 31{\cdot}24$.

$\chi^2 = 4{\cdot}785\ (0{\cdot}665) = 3{\cdot}182$. $P > 0{\cdot}5 < 0{\cdot}7$.

is applicable. Samples from the following kinds of water were used in the combined test of signs: (1) raw piped water, (2) filtered water, (3) raw well water, (4) chlorinated water, (5) untreated water from the sea and from lakes and reservoirs, (6) untreated water from surface sources, springs and streams. The numbers of times the M.P.N. of samples *B* was greater than that of samples *A* and the number of times that excess had the opposite sign for each kind of water are shown in Table 10. For the test all the subtotals must be regarded as fixed, therefore the number of degrees of freedom for the χ^2-test of homogeneity is 4. The values of the various items necessary in calculating χ^2 are also given in Table 10. The value of χ^2 is 3·182 and P, for 4 degrees of freedom, lies between 0·5 and 0·7. There is therefore no significant heterogeneity in the excess in sign for different kinds of water. Samples *B* may be expected to give a greater value of M.P.N. in 248 cases out of 353 on the average, deviations from this result being due to chance, this result being a guide, however, only when the conditions of an experiment approximate to those under which Dr Clegg's data were obtained.

Homogeneity in the excess of positive signs in the case of large differences

only may also be examined by way of the χ^2-test. If a large difference be defined by B being *at least* 5 times as great as A, cases in which the calculated M.P.N. for A is zero and that for B is 5 or more may be included.

Before leaving the subject of general tests of significance, it is necessary to refer once more to the comprehensive test of deviation from the expected distribution of differences in N.F.T. There is a hiatus between the test of the difference in N.F.T. between a single pair of samples, for which Table 3 is used, and the test for deviation from expectation in the case of a long series of differences, to which the χ^2-distribution is applicable. The χ^2-test is not, however, reliable when applied to cases in which the expectation in any frequency class is small, the lower limit being at least five. The hiatus may be filled in by making use of the exact multinomial distribution involved, but unfortunately description of the application of the multinomial is out of the question here as it would have to include a series of rather complicated diagrams. I have completed a paper on this subject and I hope to be able to give the reference to it, when published, in a future note to the Editor of this *Journal*. Many bacteriologists will be satisfied, however, by a test of the significance of a mean difference in N.F.T. such as that described in § 7. This test is very sensitive, and in its application, use is made of all relevant information contained in the data.

4. On testing the significance of a mean difference in the most probable number of bacteria per 100 ml., and on the choice of the most useful estimate of the mean difference

An estimate of a difference in M.P.N. is not of much service unless a fairly accurate estimate of the random sampling distribution of the difference is obtainable. Further, unless this distribution is approximately normal in form the practical usefulness of the standard error of the estimate as a measure of its variation is small. The M.P.N. itself has such a wide range that equal differences between values of M.P.N. do not mean at all the same thing when they occur at different parts of the range. The standard error of a difference is more or less nearly proportional to the mean of the values of M.P.N. between which the difference is taken. If only because of this it would seem more reasonable to investigate the difference between logarithms of M.P.N. rather than that between actual values. For practical purposes also an estimate of the difference between logarithms is the more generally useful. A difference between logarithms can be immediately translated into a ratio, and this ratio is equally applicable throughout the range of M.P.N. This is of particular advantage when the results of two different methods of treatment of water samples are being compared. It is more useful to be able to say that method B shows 5 times as many bacteria per 100 ml. than it is to give an actual difference in numbers which is only applicable at one point in the range. A similar advantage pertains to the standard error of the logarithmic difference.

Before testing the significance of a mean difference in logarithms of M.P.N. it is necessary to find out whether the distribution of differences is approximately normal in form.[1] I have carried out this investigation in connexion with Dr Clegg's data, to which reference has been made, with the following results.

Omitting unsuitable pairs and also results from chlorinated water, in which only one tube was used at the highest concentration, there were 251 pairs for comparison. Log differences were grouped in intervals of 0·2, the outer class intervals being +1·2 to +1·4 and −1·2 to −1·4 respectively. Replacing the lowest class character by 1, the next by 2 and so on, to simplify calculations, the frequency of differences is as shown at f in Table 11. The mean, measured from $\pm 1{\cdot}5$, is at 8·255 and is thus, when translated into logarithmic units, at +0·151, the estimate of the variance is 4·512, that of σ, 2·1241 in the units of Table 11. The expected frequencies, given by the best-fitting normal distribution, are shown at E in Table 11. The contribution to χ^2 for each class is also shown. The total value of χ^2 is 3·286, and P, for 7 degrees of freedom—

Table 11

x	1	2	3	4	5	6	7
E	0·15	0·66	2·31	6·48	14·7	26·7	39·2
f	1	1	3	4	10	28	45
χ^2		0·041			1·49	0·07	0·86

x	8	9	10	11	12	13	14
E	47·3	43·8	33·4	20·5	10·1	4·02	1·28
f	47	45	30	20	12	4	1
χ^2		0·03	0·34	0·01	0·36	0·089	

Total = 3·286. $P > 0{\cdot}8 < 0{\cdot}9$.

Note. In calculating the χ^2 total figures to 3 places of decimals were used.

three are absorbed in fitting the normal curve—lies between 0·8 and 0·9. The fit is therefore very satisfactory, and there is no valid reason why 'normal theory' should not be applied to test the significance of the mean log difference of +0·151. The estimate of the standard deviation of the mean difference is $\sqrt{\frac{4{\cdot}512}{251}}$ or 0·1341 in the units of Table 11. This is equal to 0·02682 in logarithmic units. The observed difference is therefore nearly 6 times its standard deviation, and so great a difference has a probability of less than 2×10^{-9}. We are justified in saying therefore that it is, practically speaking, impossible that the observed mean difference in the logarithms is only apparent and due to random sampling error. In Fig. 3 the observed distribution of log differences is compared graphically with the best-fitting normal distribution. It will be seen that the misfit, though only slight, is of a systematic character, being apparently due to positive skewness in the observed distribution. The cause of this is probably restriction of variation near the top and the bottom of the range owing to omission of pairs in which one member had no fertile tubes or

[1] See § 7.

no sterile tubes. If a dot diagram be made, showing each logarithm of M.P.N. for A on the scale of abscissae and the corresponding logarithm for B on the ordinate scale, the skewing effect of the omission of end pairs is clearly shown. The only way to eliminate this effect would be to include, for our estimates of the mean log difference and its standard error, only those pairs for which the 'dot' did not fall in a row or column which was cut off at zero or 1800 on either scale. This would involve the omission of a great number of pairs beyond those already omitted. I do not think that the observed skewness, though probably real, is of sufficient importance to necessitate these additional omissions.

Our estimates may be at once translated into ratios which should be of considerable service to bacteriologists who may wish to make use of method A

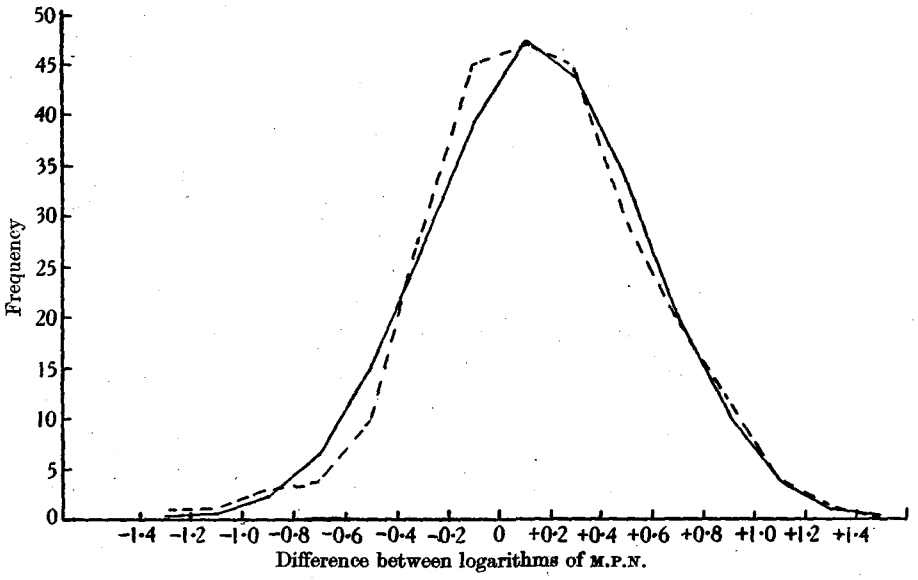

Fig. 3.

sometimes since it is simpler than method B and takes less time though giving less satisfactory results.

We have, for the mean ratio, '$B:A$', antilog 0·151 or 1·416. We may employ the usual level of significance of 0·025 in each direction from the mean to give convenient limits to the expected range of ratios. The standard deviation of the log differences is 0·42482. The value of x/σ corresponding to the 0·025 point is approximately 1·96. Multiplying this value by 0·42482 we have 0·8327. For positive differences we have 0·8327 + 0·1510 or 0·9837, for negative differences 0·8327 − 0·1510 or 0·6817 as the end-points of the 19:1 log range. We may expect then, on the average, that in nineteen cases out of twenty if the number given by method B be greater than that given by method A, it will not be more than about 9·6 times as great, whereas, in the reverse case,

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

the number given by method *A* will not be more than about 4·8 times the greater. This result cannot, of course, be used to 'raise' a zero count obtained by method *A* to the expected count by method *B*, but it may be useful in dealing with the mean of a series of counts obtained by method *A* or in cases in which it is of interest to compare a series of counts made by one worker using method *A* with those of another worker who used the other method.

It should be emphasized that the numerical results of our investigation into log differences in M.P.N. are applicable only to the methods *A* and *B as used by Dr Clegg*, with three dilutions and five tubes at each dilution. In other cases the distribution of log differences may not be approximately normal in form and, even if it is of this form, its standard deviation may vary widely with varying conditions. If, on further investigation, it is found that the distribution is generally of normal form, the well-known *t*-test may be applied to test the significance of the mean of quite a small series of log differences. The use of this test is clearly set forth in Fisher (1934), § 24, and therefore it is unnecessary to explain the test here. In our case Fisher's 'variate x' is a single log difference. It is advisable to use the *t*-test instead of the general normal test when the number of differences is less than 30.

5. ON THE CHOICE OF A LEVEL OF SIGNIFICANCE

The level of significance generally employed in both experimental and observational work is that of 0·05 or 0·025 in either direction. If an observation lies outside this point and the null hypothesis be rejected because of this it will be wrongly rejected if true in one case out of twenty. The choice of a level is, however, arbitrary, and the value should, I think, be varied according to circumstances, provided that in each particular set of circumstances the same level be adhered to. Standards of potability for water are not usually flexible, and in judging whether water is potable *any* apparent excess in pollution above the standard would therefore have to be considered significant. In carrying out investigations into change in degree of pollution from place to place or from time to time, it should be realized that comparison of a single pair of samples can yield little information on the question of the reality of a *small* difference, so that it would seem preferable to adopt a fairly high level of significance for single pairs. We should prefer to be fairly confident that we are distinctly more likely to be right than wrong in rejecting the null hypothesis, but I do not think it necessary that the degree of confidence should be as great as that given by the use of the 0·05 level. I would suggest that the level of 0·2 would be suitable for general bacteriological work. It is true that a higher level, for instance, that of 0·1, would be preferable for many reasons, but the range of differences in estimates of degree of pollution is so small that, if the higher level were used there would only be three or four 'stages' in the whole range of pollution which could be estimated from a given set of dilutions. It is therefore preferable to use a low level and against every result to give the sign of the difference in N.F.T. and the probability of the difference as given in

Table 3. If the results should be afterwards combined it is then a simple matter to look for consistency in sign which, as we have seen, may be very significant. If this procedure be adopted it is not generally necessary to concern ourselves with a hard and fast level of significance applicable to single pairs. This is only necessary in isolated cases.

6. On the results of a control experiment

Dr C. B. Taylor, who is working at Wray Castle on the bacteriology of water, has kindly handed me some data from a control experiment which he is carrying out on the method which I have called method *B* in connexion with Dr Clegg's data. In Dr Taylor's experiment ten tubes were inoculated at each dilution and incubated at 37° C., further tubes being inoculated from those proving fertile at 37° C. and incubated at 44° C. to confirm presence of faecal coliform bacteria. The original ten tubes inoculated at each dilution

Table 12

Classes uncombined

Difference ...	−5−	−4−	−3	−2	−1	0	+1	+2	+3	+4+	+5+
Expected (*f*)	0·16	0·96	1·48	4·91	4·86	8·25	4·86	4·91	1·48	0·96	0·16
Observed (*f*)	1	0	2	7	5	10	5	1	1	1	0

Even classes combined

Difference ...	−2−	0	+2+
Expected (*f*)	5·56	7·88	5·56
Observed (*f*)	7	10	2
χ^2	0·3730	0·5703	2·279

Odd classes combined

Difference ...	−1−	+1+
Expected (*f*)	7	7
Observed (*f*)	8	6
χ^2	0·143	0·143

Total of $\chi^2 = 3{\cdot}5083$. D.F. $= 3$. $P > 0{\cdot}3 < 0{\cdot}5$.

Note. '+5+' means '5 and greater than 5' and so on.

were divided haphazard into two sets of five, *A* and *B*. We have therefore a chance distribution of differences which should conform to the null hypothesis that there is no difference between the degree of pollution in tubes *A* and that for tubes *B*. There were thirty-three pairs and the letter classes of Table 5 represented were as follows, the number of times each letter occurred being given in brackets: *a* (1), *c* (3), *d* (3), *f* (1), *h* (6), *i* (6), *j* (2), *k* (2), *m* (4), *n* (1), *o* (1), *s* (1), *y* (1), *B* (1). Owing to the restriction, in the χ^2-test, that the expectation in any class must not be a very small number, considerable combination of the probabilities given under the letter classes in Table 5 is necessary. I have therefore combined the even letter classes into three frequency classes, namely −2 and negative differences greater than 2, zero differences, and differences of +2 and positive differences greater than 2, these being called respectively −2−, 0, and +2+. The odd letter classes were combined into two classes, namely −1− and +1+. The expected and observed frequencies of the various differences, the contribution to χ^2 from each class and the total value of χ^2 are shown in Table 12.

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

The misfit is only such as would be expected to occur about once in three trials. There is certainly no significant experimental error. It would seem that experimental error cannot very well influence a result from a series of differences in N.F.T. if the inoculation of the two series of tubes to be compared be carried out in a truly random fashion. It would not be wise, for instance, to inoculate all the tubes for A before those for B or to take them alternately. It would be preferable to randomize the order of inoculation and the position of the tubes in the incubator by tossing a coin in the case of each tube. In taking samples from different places for comparison over a period the order in which the places are visited should be randomized when possible, otherwise a biased error may occur, due perhaps to change of temperature or to some other change with time. Unbiased error cannot, as a rule, be eliminated owing to the fact that complete synchronization in sampling is usually impossible, but such unbiased error, if significant, must be considered as showing real differences at the times of sampling. The 'time effect' could, of course, be studied by way of a series of samples taken at fixed times in different places.

Table 13

Probability of an excess in positive or negative signs of at least

n	0 or 1	2 or 3	4 or 5	6 or 7	8 or 9	10 or 11	12 or 13	14 or 15	16 or 17
3	1	0·25							
4	1	0·625	0·125						
5	1	0·375	0·0625						
6	1	0·6875	0·21875	0·03125					
7	1	0·45313	0·125	0·01563					
8	1	0·72656	0·28906	0·07031	0·00781				
9	1	0·50781	0·17971	0·03906	0·00391				
10	1	0·75391	0·34375	0·10937	0·02148	0·00195			
11	1	0·54883	0·22657	0·06543	0·01173	0·00098			
12	1	0·77424	0·38770	0·14235	0·03848	0·00635	0·00049		
13	1	0·58105	0·26685	0·09018	0·02246	0·00342	0·00024		
14	1	0·79053	0·42395	0·17957	0·05737	0·01297	0·00183	0·00012	
15	1	0·60724	0·30176	0·11847	0·03516	0·00739	0·00100	0·00006	
16	1	0·80362	0·45450	0·21011	0·07681	0·02127	0·00418	0·00051	0·00003
17	1	0·62906	0·33154	0·14346	0·04904	0·01273	0·00235	0·00027	0·00002
18	1	0·81453	0·48068	0·23788	0·09625	0·03088	0·00754	0·00131	0·00015
19	1	0·64761	0·35928	0·16707	0·06357	0·01921	0·00443	0·00073	0·00008
20	1	0·82380	0·50344	0·26318	0·11532	0·04139	0·01179	0·00258	0·00040

Note 1. For $n=18$, $P(18)=0{\cdot}00001$; for $n=19$, $P(17)=0{\cdot}00008$; for $n=19$, $P(19)<0{\cdot}00001$; for $n=20$, $P(18)=0{\cdot}00004$, $P(20)<0{\cdot}00001$.

Note 2. When n is odd only odd excesses are possible.

7. On tests of the significance of a mean difference in N.F.T. and of a mean error from zero

The theoretical or philosophical basis of methods of testing the significance of a mean difference when we have no knowledge of the random sampling distribution of *each component difference* is not the same as when that distribution is known. In the former case the distribution must be assumed to be that of the observed differences or to be represented by the best-fitting normal distribution. If the observed distribution of differences is approxi-

mately normal in form the *t*-test may be applied. If not, the test described by Fisher (1935) under the title 'Test of a Wider Hypothesis' is applicable. It is of great interest to compare the results of a test in which knowledge of the form of the distribution of each component difference is left out of account with that arising from a test in which use is made of that knowledge. Some of Dr Taylor's data are suitable for this comparison. Dr Taylor has, over a long period, taken water samples each week from certain fixed stations and the

Table 14. *Application of t-test*

x		x^2
+1	(*a*)	1
−3	(*b*)	9
+1	(*c*)	1
+6	(*d*)	36
0		
+3	(*e*)	9
+2	(*f*)	4
+4	(*g*)	16
+9	(*h*)	81
+2	(*i*)	4

$+25 = S(x)$. $\bar{x} = 2{\cdot}5$.

$S(x^2) = 161$. $S(x-\bar{x})^2 = S(x^2) - \bar{x}S(x) = 98{\cdot}5$.

$$s_{\bar{x}}^2 = \frac{98{\cdot}5}{9 \times 10} = 1{\cdot}094. \quad s_{\bar{x}} = 1{\cdot}046.$$

$$t = \frac{2{\cdot}5}{1{\cdot}046} = 2{\cdot}390. \quad n = 9. \quad P < 0{\cdot}05 > 0{\cdot}02.$$

Fisher's test of a wider hypothesis

'A total difference equal to or greater than that observed will occur

	Frequency
With 0 negative signs	1
1 negative sign (*a, b, c, e, f, i*)	6
2 negative signs (*ac, af, ai, cf, ci*)	5
	12

If the observed total difference be negative an equal or greater negative difference will also occur 12 times. The total number of possible combinations is 512. Thus the probability of at least as great a total difference as that observed is equal to

$$\frac{24}{512} \text{ or } 0{\cdot}04687.$$

degree of pollution of the water at each station by coliform bacteria has been estimated. In some cases there has been an apparent general increase or decrease in pollution between one week and the next and it is interesting to find out which of these general apparent changes are significant.

In Table 14 the first column gives the differences between the N.F.T. at the 10 fixed stations on 7 October 1940 and the corresponding values of N.F.T. for the week before. Firstly let us assume that the only information we have about the distribution of differences is that given by the observations themselves and that it is reasonable to assume that the differences are normally distributed. We may then apply the *t*-test and shall find that P is slightly less than 0·05. We should judge therefore that the general increase in pollution is just significant. It will be seen that Fisher's exact test gives a very similar

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

result. Now let us consider the implications of the assumption that our only information about the distribution of differences is that given by the observations in the light of our accurate knowledge of the form of this distribution which has been given tabular shape in Tables 3 and 5. Consider the difference of +9. This large difference appears once in ten trials, therefore its probability is assumed to be 0·1, half as great as the difference of +1. The effect of the great width of the range of differences observed—from −3 to +9—is to give a large standard error and therefore to minimize the significance of t. The case is completely altered if our knowledge of the probability of so great a difference as +9 is taken into account. This difference falls in letter class 0 and, on reference to Table 3, we find that $P = 0{\cdot}00088$, very different from 0·1.

When our only knowledge of the form of the distribution of differences is that given by the observations we have to fit a theoretical or an empirical curve of distribution to these and investigate the question whether the range of variation of the mean of that curve is such that the mean is likely to be zero. If this is likely the observed mean is not significant. In the other case we have a *known* distribution, we have to fit our set of observations to it and investigate the question whether its mean is reasonably well estimated by the mean of our observations; if it is, the observed mean is not significant.

We must therefore choose some known symmetrical distribution with a mean of zero and fit our observed differences to it, their correct positions being assigned by the corresponding values of P which are given in Table 3. The distribution must be of a type for which the distribution of the mean is known and to which all the values of P given in Table 3 may be fitted. It is clear that the normal distribution is the most suitable for our purpose. Adjustment must be made, however, to allow for the fact that the normal distribution is continuous while the probabilities of our differences are finite. This adjustment may be based on the assumption that a difference in N.F.T. is really a continuous variable quantity which can only manifest itself in a short range of whole numbers. This is not an absurd assumption. A difference of two in ten tubes may be considered as a difference of 20 in 100 or of 200 in 1000. By increasing the number of tubes the difference may be made to approximate as nearly as we like to an infinitesimally variable quantity. We shall consider that the class '2' really comprises all differences between 1 and 3. Now any univariate distribution may be expressed on the normal scale if the normal abscissa x/σ for each class be placed at what would be the mean of that class on the normal scale.[1] The mean of the class '2' is given by

$$\frac{z_2 - z_4}{p(2)},$$

where z_2 is the normal ordinate corresponding to the value of $P(2)$ given in Table 3, z_4 is the corresponding ordinate for $P(4)$ and $p(2)$ is the probability

[1] I hope shortly to submit for publication a table of the normal abscissae equivalent to the entries in Table 3.

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

of a difference of 2 as given in Table 5 for the letter class in which the difference falls. To find z_2 in class h, for instance, we write

$$1-\tfrac{1}{2}P\,(2)=0{\cdot}70833=\tfrac{1}{2}(1+\alpha),$$

and against 0·70833 in Table 2 of Pearson (1930) we find the corresponding value of z to be 0·3429 approximately. For z_4 we have

$$1-\tfrac{1}{2}P\,(4)=0{\cdot}95833=\tfrac{1}{2}(1+\alpha)$$

and $z=0{\cdot}0893$ approximately. From Table 5 we find that $p\,(2)=0{\cdot}25$. For the required value of x/σ we have therefore

$$\frac{x}{\sigma}=\frac{0{\cdot}3429-0{\cdot}0893}{0{\cdot}25}=1{\cdot}014.$$

Table 15. *Test of significance of mean difference and mean error from zero*

d	Class	$P(d)$	$P(d+2)$	$p(d)$	$d(z)$	x/σ
+1	c	0·5	0·08333	0·4167	0·2450	+0·5880
−3	p	0·18052	0·03171	0·1488	0·1930	−1·2970
+1	e	0·5	0·10317	0·3968	0·2186	+0·5509
+6	o	0·00695	0	0·0070	0·0194	+2·7710
0	b	0·5	0·22222	0·2778	0·1000	±0·3559
+3	c	0·08333	0	0·0833	0·1540	+1·8490
+2	h	0·29167	0·04167	0·2500	0·2536	+1·0140
+4	m	0·07143	0·00529	0·0661	0·1204	+1·8220
+9	O	0·00044	0	0·00044	0·00156	+3·5450
+2	b	0·22222	0	0·2222	0·2989	+1·3450
						+12·1879

$\bar{x}/\sigma=+1{\cdot}21879.$ $\quad \sigma_{\bar{x}/\sigma}=\dfrac{1}{\sqrt{10}}=0{\cdot}3162.$ $\quad \dfrac{\bar{x}/\sigma}{\sigma_{\bar{x}/\sigma}}=3{\cdot}864.$ $\quad P<0{\cdot}000118.$

Mean error $=1{\cdot}51418.$ $\quad$ Standard error $=\sqrt{\dfrac{S(x/\sigma)^2}{10}}=1{\cdot}7967=\sigma.$

$\sigma_\sigma=\dfrac{1}{\sqrt{20}}.$ $\quad \dfrac{\text{Deviation}}{\sigma_\sigma}=\dfrac{0{\cdot}7967}{(\sqrt{20})^{-1}}=3{\cdot}563.$ $\quad P=0{\cdot}0001833$ approx.

It is generally sufficient to take the nearest tabulated value of $\frac{1}{2}(1+\alpha)$ for the calculation of x/σ, and to include only four significant figures for the final calculation. I have found the book of Tables (Bottomley, 1919) very convenient for these calculations as it includes logarithms, squares, square roots, reciprocals and many other functions.

The mean of the calculated values of the normal abscissa will be approximately normally distributed with standard deviation equal to $\sqrt{n}$, n being the number of differences.[1]

Table 15 shows the necessary procedure in calculating P, the probability of at least as great a mean difference, either positive or negative, as that observed. The data used in this example are the same as were used in our discussion on the t-test. It will be seen that the mean difference in N.F.T. is very highly significant, so great a difference would only occur by chance about once in 10,000 trials. Our knowledge of the exact distribution of a difference in N.F.T. is proved to be of great practical importance.

[1] The deviation from normal form is such that the significance of the mean is usually underestimated. If significance is shown therefore it is safe to accept the indication.

It may be thought that the adjustment for discontinuity described previously may possibly exaggerate the significance of a mean difference. I do not consider that this is a serious possibility. It might equally well be claimed that the smoothing effect of the adjustment helps to mitigate the limitations in variation imposed by the use of very small numbers of tubes. However, to take one step in investigating the possibility of exaggeration of significance, I applied the method displayed in Table 15 to Dr Taylor's control data which are discussed in § 6. P was found to be equal to 0·2113 approximately and the mean difference is therefore nowhere near significance when judged by the usual standard. The result is in close agreement with that of the binomial sign test in which also P was found to be equal to 0·2113 approximately. This appears to show that the mean of all differences without regard to sign was very near that expected. This question may be made the subject of a separate enquiry.

In calculating the mean error the mean x/σ for each zero difference must be included in the sum, though, as these values are signless, they have no effect on the mean difference. For Dr Taylor's control data the mean error was found to be 0·7582. The true mean error is given by

$$\frac{z_0}{\frac{1}{2}(1+\alpha)_0} = \frac{0{\cdot}39894}{0{\cdot}5} = 0{\cdot}7979.$$

The standard deviation of the observed differences is equal to $0{\cdot}7582 \times 1{\cdot}253$ or 0·9499, the true standard deviation being equal to $0{\cdot}7979 \times 1{\cdot}253$ or unity. The standard deviation of σ or σ_σ is equal to $\frac{1}{\sqrt{(2n)}}$ which is 0·125. The deviation, 0·0501, divided by σ_σ, is equal to 0·4007 and $P = 0{\cdot}689$. The observed mean error agrees very well therefore with expectation.

The enquiry into the significance of the mean error was also carried out in the case of Dr Taylor's data from fixed stations and the result is included in Table 15. The mean error is very highly significant. In this case the standard error was calculated direct from the sum of squares of x/σ.[1]

It should be understood that what has been tested is the mean error from zero, not the mean spread about the very significant mean difference. These two are not at all the same thing. A highly significant mean error or standard error from zero, if coupled with a highly significant *mean difference*, indicates homogeneity in the component differences and shows that there is a highly significant general effect extending over the component estimates of pollution. In the present case there is undoubtedly a marked general rise in the degree of pollution at the fixed stations. A significant mean difference, if coupled with a much less significant mean error from zero, would, as a rule, indicate that the effect was not general but due only to isolated large differences. If, however, there were significant consistency in sign of the differences the mean difference might be significant, the mean error from zero not significant.

[1] This method is preferable as tending towards underestimation of significance.

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

This would indicate a real but very slight general effect. In investigating the significance of a general effect extending over a series of observations the implications of the significance of these various kinds of deviation from expectation should be borne in mind. Where, as in the data displayed in Table 15, all tests agree in showing high significance, while the test of the mean difference shows very high significance, there is practical certainty that not only is there a general increase in degree of pollution but that this increase is large.

8. Theoretical bases of tests of the significance of a difference in N.F.T. and M.P.N. compared

I have chosen the difference in N.F.T. in preference to the difference in M.P.N. as the variate in tests of the significance of various kinds of deviation from expectation chiefly for the following reasons. (1) The theoretical basis is not only the simpler to understand but it is based more nearly on the laws of pure chance. (2) The exact distribution is *readily* determinable and the exact probability of every possible difference may be tabulated. (3) Every pair of samples in which the members differ from each other may be included.

If the degree of pollution in two original water samples is really exactly the same we are entitled to assume that the hypothetical common sample *at each dilution* consists of a mixture of equal parts of the two original samples diluted to the appropriate extent. No estimate of the degree of pollution of each of the original samples is involved. When a pipetteful is taken from any of the common samples for inoculation of a tube in set A or set B it is a question of pure chance which set the inoculated tube belongs to. We are dealing solely with laws of chance based on the binomial

$$(\tfrac{1}{2}+\tfrac{1}{2})^n.$$

This is the simplest implication of the null hypothesis that there is no difference in the degree of pollution of the two samples compared. It seems unnecessary to use a complicated function such as M.P.N. when the simplest function of the number of fertile tubes will meet the case equally well. One must, however, be able to assume that *at each dilution* the two subsamples are such as might have been taken from a mixture of the original samples. Therefore when two samples, A and B, are to be compared, it is necessary that, *for each dilution* of A, a new aliquot part be taken from the original A sample. Similarly for B. From the sampling point of view also this method is preferable to that of making up greater dilutions by steps from more concentrated suspensions.[1]

As for point (2), it is, of course quite possible to calculate the probability of any difference in M.P.N., since the standard error of a difference is calculable.

[1] When, for practical reasons, aliquot parts of the original samples have to be diluted before drawing pipettefuls for test, allowance should be made for additional chance variation by using higher levels of significance, e.g. 0·1 for single pairs, ·025 for series.

This is not, however, constant, and also there are so many possible differences that tabulation of the distribution seems out of the question.

My point (3) has already been dealt with.

It may be objected that, in applying the normal test of the significance of a difference in the logarithms of M.P.N. in a previous part of this paper, use was not made of knowledge of the probability of each difference. Calculation of all the probabilities involved would, however, have taken a prodigiously long time and it was not attempted for the reason that the chief interest in the logarithmic differences lies in their value in obtaining an estimate of the ratio of one mean to another and not in their application to the problem of estimating the significance of a mean difference.

9. ON TESTING THE SIGNIFICANCE OF TIME OR LOCALITY EFFECTS AND ON THE USE OF REPLICATED OBSERVATIONS

In § 7 and Table 15 a simple example is given of the application of the exact distribution of differences in N.F.T. in the examination of a general change in degree of pollution in time. Many other effects may be studied by methods similar to that described in § 7. Three of Dr Taylor's fixed stations which were referred to in that section and which we shall call stations 5, 6 and 7 respectively, are in Lake Windermere, station 5 being the nearest to a source of pollution, station 6 farther from it, station 7 farther off still. The decrease in degree of pollution as we pass from station 5 to station 7 varies from time to time and it is interesting to compare the significance of the decreases observed.

The strength of any effect, if the corresponding value of P be determinable, may be measured on the normal scale and the effects may thus be arranged in order of strength. Effects of every kind may be compared on this scale.

For each of the stations 5, 6 and 7 duplicate estimates of the degree of pollution by coliform bacteria were made. These form the control data which were discussed in § 6 and it was found that there was no significant variation between the members of each pair of observations. Whether this variation is significant or not, however, it is preferable to make use of all relevant information given by data and, in the present case, to use both members of each pair of observations separately rather than to assume that they give identical results.

We are more concerned with finding out if there is a significant falling off in degree of pollution as we move away from the source rather than in investigating the exact way in which the decrease varies from station to station. A very suitable selection of differences for testing the significance of the effect in question seems therefore to be that shown in Table 16. In this table the first sample taken at each station is designated A, the second B. If the degree of pollution at station 5 is greater than that at station 6, or that at 6 greater than that at 7 the difference is considered positive. The observations included in Table 16 were those for the month of April 1941. In Table 17 full details

Table 16

Date: 7 April 1941

Stations	d	Class	x/σ	
5 A and 6 A	+2	m	+0·8962	$\bar{x}/\sigma = 0{\cdot}8117$
5 B and 6 B	−1	a	−0·7979	$\frac{\bar{x}/\sigma}{\sigma_{\bar{x}/\sigma}} = \frac{0{\cdot}8117}{0{\cdot}4083} = +1{\cdot}988$
5 A and 7 A	+5	i	+2·6080	
5 B and 7 B	0	k	0	$P = 0{\cdot}04659$
6 A and 7 A	+3	g	+1·7060	
6 B and 7 B	+1	l	+0·4582	
			+4·8705	

Date: 15 April 1941

Stations	d	Class	x/σ	
5 A and 6 A	0	j	±0·2542	$\bar{x}/\sigma = 0{\cdot}8023$
5 B and 6 B	0	m	±0·2344	$\frac{\bar{x}/\sigma}{\sigma_{\bar{x}/\sigma}} = \frac{0{\cdot}8023}{0{\cdot}4083} = +1{\cdot}965$
5 A and 7 A	+2	h	+1·0140	
5 B and 7 B	+3	B	+1·2280	$P = 0{\cdot}0488$
6 A and 7 A	+2	b	+1·3450	
6 B and 7 B	+3	B	+1·2280	
			+4·8150	

Date: 21 April 1941

Stations	d	Class	x/σ	
5 A and 6 A	+5	y	+2·2370	$\bar{x}/\sigma = 1{\cdot}457$
5 B and 6 B	+5	p	+2·1650	$\frac{\bar{x}/\sigma}{\sigma_{\bar{x}/\sigma}} = 3{\cdot}659$
5 A and 7 A	+8	H	+3·2950	
5 B and 7 B	+2	d	+1·1290	$P = 0{\cdot}000357$
6 A and 7 A	+3	n	+1·3410	
6 B and 7 B	−3	l	−1·4190	
			+8·7480	

Date: 25 April 1941

Stations	d	Class	x/σ	
5 A and 6 A	+6	x	+2·8720	$\bar{x}/\sigma = 1{\cdot}537$
5 B and 6 B	+2	E	+0·7863	$\frac{\bar{x}/\sigma}{\sigma_{\bar{x}/\sigma}} = 3{\cdot}764$
5 A and 7 A	+6	x	+2·8720	
5 B and 7 B	+4	m	+1·8200	$P < 0{\cdot}00017$
6 A and 7 A	0	b	0	
6 B and 7 B	+2	x	+0·8738	
			+9·2241	

Table 17. *Details of calculations for* 25 *April* 1941

Stations 5 A and 6 A. $+6x$.

$1 - \frac{1}{2}P(d) = 0{\cdot}99405$, $1 - \frac{1}{2}P(d+2) = 1$. $\quad \frac{z - z'}{p(d)} = \frac{0{\cdot}01709}{0{\cdot}00595} = +2{\cdot}8720 = x/\sigma$.

Stations 5 B and 6 B. $+2E$.

$1 - \frac{1}{2}P(d) = 0{\cdot}65972$, $1 - \frac{1}{2}P(d+2) = 0{\cdot}89385$. $\quad \frac{z - z'}{p(d)} = \frac{0{\cdot}18413}{0{\cdot}23413} = +0{\cdot}7863 = x/\sigma$.

Stations 5 A and 7 A. $+6x$.

As for stations 5 A and 6 A. $= +2{\cdot}8720$.

Stations 5 B and 7 B. $+4m$.

$1 - \frac{1}{2}P(d) = 0{\cdot}92857$, $1 - \frac{1}{2}P(d+2) = 0{\cdot}99471$. $\quad \frac{z - z'}{p(d)} = \frac{0{\cdot}12036}{0{\cdot}06614} = +1{\cdot}8200$.

https://doi.org/10.1017/S0022172400012377 Published online by Cambridge University Press

of the calculation of some of the values required for 25 April are shown. As the standard deviation from zero was not investigated the value of the normal equivalent for the difference in stations 6 A and 7 A, being signless, was equated to zero in Table 16. There is a very interesting change in the strength of the effect studied between the first and the second halves of the month.

The values of the mean normal abscissa, each divided by its standard error, may be combined for comparison with other mean effects. Thus the monthly means may be combined to give a measure of the yearly effect on the normal scale. Such measures may be examined by any statistical methods appropriate to the normal distribution, for instance by the methods of correlation, regression, or analysis of variance.

SUMMARY

A detailed study is made of methods available for estimating the significance of difference in degree of pollution of water by coliform bacteria and of obtaining practically useful estimates of such differences.

For estimating the significance of a difference in a single pair of samples it was considered preferable to employ as the variate the difference in the total number of fertile tubes, or N.F.T., rather than the difference in the so-called most probable number of bacteria per 100 ml. A table (Table 3) is included by which the significance of any difference in N.F.T. may be determined at a glance. The variate is also useful in determining whether a series of differences may be regarded as homogeneous and such as might be expected to arise by chance in a high proportion of trials if corresponding members in a series of pairs of samples had been taken from sources identical as to their degree of pollution. Table 5 is included to facilitate the application of a comprehensive test of identity, in degree of pollution, of the members of each pair of samples.

The application of the binomial distribution in testing the significance of consistency in sign of a series of differences is explained, and Table 13 is included to facilitate the application of the binomial for a series of pairs twenty or less in number. The use of the normal distribution as an approximation to the binomial for pairs over twenty in number is explained.

The distribution of differences in the logarithms of the 'most probable number of bacteria per 100 ml.', or M.P.N., between samples in a certain series, *A*, and those of another series, *B*, in an experiment carried out by Dr L. F. L. Clegg, was found to be approximately normal in form. Thus the normal test of the mean logarithmic difference in M.P.N. is applicable and the *t*-test is applicable to a small series of differences. The advantages of employing the logarithmic difference as the variate are pointed out.

All tests applied to Dr Clegg's data agree in showing that the degree of pollution estimated from the results of a certain method, *B*, was higher than that indicated by the results of another method, *A*. According to method *A* the tubes were inoculated and incubated at once at 44° C., while according

to method B the inoculated tubes were incubated at 37° C., those proving fertile at that temperature being used for inoculating fresh tubes which were then incubated at 44° C. The mean ratio, $B : A$, was found to be 1·416 and the 0·025 points of the distribution were found to be $B : A = 9{\cdot}6$, $A : B = 4{\cdot}8$ respectively.

The choice of a level of significance is discussed.

The results of the application of some statistical tests of significance to data from a control experiment are discussed. In this experiment the members of each pair of samples were taken from the same water sample and treated by identical methods.

The methods of replacing a difference by the corresponding value of the normal abscissa and of applying the normal test of significance of a mean to the mean of such normal abscissae is explained. This method is applied in testing the significance of a mean difference in N.F.T. and in testing the significance of a mean error from zero.

The theoretical bases of tests of the significance of differences in M.P.N. and in N.F.T. are compared.

The advantage of knowledge of the exact distribution of a single difference in tests of the significance of a series of differences is explained and examples are given which demonstrate this advantage very clearly.

The combination of results taken at different times and places for testing the significance of time or locality effects is explained and a suggestion is made as to the use of replicated observations when the value of P for each observation is known.

REFERENCES

BOTTOMLEY, J. T. (1919). *Four Figure Mathematical Tables.* London: Macmillan and Co.

FISHER, R. A. (1934). *Statistical Methods for Research Workers,* 5th ed. Edinburgh: Oliver and Boyd.

—— (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

PEARSON, KARL (1930). *Tables for Statisticians and Biometricians.* Part I, 3rd ed. Cambridge University Press.

(*MS. received for publication* 24. VI. 41.—Ed.)