A NOTE ON THE CHOICE OF SAMPLE VOLUME IN A BACTERIOLOGICAL STUDY OF WATER

By J. H. DARWIN

Department of Scientific and Industrial Research, New Zealand

1. INTRODUCTION

In bacteriological investigations of stretches of water an estimate of the number of a certain type of bacterium in a unit of volume of the water is required so that the pattern of the pollution of the area can be discerned. Usually it is known or guessed that the density of bacteria, δ say, lies in a range δ_L up to δ_H. Several samples of water are taken and each is tested for a positive reaction which would indicate the presence of at least one bacterium of the type being searched for. When it is assumed that bacteria are spread in a Poisson manner through the water in each sample an estimate of their density can be made from the number of positively reacting samples. If the range δ_L to δ_H is large it is unlikely that one volume of water for each sample will give a very efficient estimate of the density; e.g. if this volume is in an average position with respect to (δ_L to δ_H), say, $2/(\delta_L+\delta_H)$, and the true density is near an end of the range, all the reactions are likely to have the same sign and the sample gives no discrimination. Even worse estimation for some densities will occur if a very small or very large volume is used. The setting of reasonably sized confidence intervals for the density requires a series of volumes designed to locate the density with about equal accuracy no matter where it is in (δ_L to δ_H). Hence the usual set of volumes is the dilution series discussed by several writers, for example recently by Cochran (1950).

In this series, which for uniformly good estimation should range from $1/\delta_H$ to $1/\delta_L$ (Cochran, 1950), the ratio of one volume to its lower neighbour should be a constant, the dilution ratio. This ratio should be low and the number of samples at each volume preferably the same. Even with this latter proviso the equation of estimation of the most probable number of organisms per unit of volume is not particularly easy to solve, although good methods have been devised (Finney, 1951). The usual provision of confidence intervals for δ depends on the assumption that the estimated density (or most probable number), d, is log normally distributed. However, Cochran's table, showing the approximate constancy of the variance of the log of this estimate, indicates that this is a very reasonable assumption. Swaroop (1951) gives confidence intervals by means of a normalizing transformation.

There may occasionally arise a simplified problem in which these two difficulties of solution of an equation and of normality do not arise; namely when good estimation over the whole possible range (δ_L to δ_H) of densities is not required. For instance, in routine testing of water supplies to see if pollution is at a satisfactorily low level it is perhaps not so important to know what the level is provided it is demonstrably low enough. If the density is getting near a dangerous level it will be more important to locate it. If it is much worse than it should be steps will be taken to lower it

https://doi.org/10.1017/S0022172400000711 Published online by Cambridge University Press

and the more accurate estimate will be needed when these have almost succeeded. Thus we may require a test giving good estimation in the range of densities hovering between acceptable and unacceptable and speedily indicating if the density is quite acceptable or quite unacceptable. The result of the test should be a decision to accept the water or to reject it. Because of the fairly high standard error of log d in dilution series estimations, water is sometimes accepted, sometimes rejected when its density is in the doubtful region. By concentrating the best estimation in the doubtful region we can cut down the seriousness of this sort of test behaviour.

The density range being considered has been cut down from (δ_L to δ_H) to (δ'_L to δ'_H), say, where δ'_L is the acceptable density level and δ'_H the unacceptable density level. Generally the possible density level may go far beyond the unacceptable level so that δ'_H is much less than δ_H, and it may go almost to zero so that δ_L is much less than δ'_L. Whereas at least 3 sample volumes are required to give efficient estimation over the complete range (δ_L to δ_H) it is possible that only 1 may be needed to give good estimation over the range (δ'_L to δ'_H) and at the same time give us stated risks of rejecting the water when in fact the density is less than δ'_L and of accepting it when in fact the density is greater than δ'_H. It is desirable that the risks α and β, say, of making these two mistakes should be controllable.

2. SEQUENTIAL TESTING OF SAMPLES

Suppose the samples are examined in turn, i.e. sequentially, and the number of samples reacting positively is plotted in a running graph against the number of samples examined. Then for such tests these risks α and β can be fixed. The samples are examined till the proportion that react has become so high that the water is rejected or so low that it is accepted. Sequential tests of this sort for given risks α and β require a lower average number of samples than ordinary tests in which all the samples are examined at once.

Ordinarily in sequential testing the items are examined on the spot to see how they react. In pollution sampling the samples would be numbered in order as they were drawn from the source and tested later. The assumption is that all samples are randomly drawn from the same population.

2.1. *The test region*

The equations of rejection and acceptance boundaries the crossing of which by the plotted line produces a decision are

No of positively reacting samples $= h_2 + s$ (no. of samples tested).

No. of positively reacting samples $= -h_1 + s$ (no. of samples tested), where

$$h_1 = \frac{\log[(1-\alpha)/\beta]}{\log[(p_2/p_1).(1-p_1)/(1-p_2)]}, \quad h_2 = \frac{\log[(1-\beta)/\alpha]}{\log[(p_2/p_1).(1-p_1)/(1-p_2)]},$$

$$s = \frac{\log[(1-p_1)/(1-p_2)]}{\log[(p_2/p_1).(1-p_1)/(1-p_2)]}.$$

p_1 and p_2 are the probabilities of a sample reacting positively when the volume being used is v and densities are δ'_L and δ'_H respectively. The equations are given,

https://doi.org/10.1017/S0022172400000711 Published online by Cambridge University Press

for example, by Wald (1947, p. 94). By the Poisson theory of random dispersal of the bacteria through the water being tested, these are $p_1 = 1 - \exp(-\delta'_L v)$ and $p_2 = 1 - \exp(-\delta'_H v)$.

2.2. *Choice of α and β*

The usual statistical levels of significance are 0·05 and 0·01. The mistake of accepting water that should be rejected may be thought more serious than that of rejecting water that should be accepted. That is, β may be made less than α. The cases considered are (1) $\alpha = \beta = 0{\cdot}05$; (2) $\alpha = \beta = 0{\cdot}01$; (3) $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}01$. It is important to remember that risks corresponding to α and β are already present in dilution series tests. These series can be so planned that these risks have a given size as here. The work would be approximate because of the assumption of normality of $\log d$, whereas here there is no approximation in the tests apart from the initial serious Poisson assumption (also used in the dilution series). The number of samples drawn from the water for inspection may not be sufficient to give a decision when all samples have been tested. If the policy is always to come to a decision, the safest one would be that all water not accepted should be rejected. This rule would tend to raise α and lower β, and if possible one should take a number of samples big enough for this contingency rarely to arise. If no decision has been reached it is probable that the density is in the doubtful range δ'_L to δ'_H, and estimation of it is as important as a decision whether or not to reject the water.

3. CHOICE OF THE VOLUME v

Wald (1947, pp. 99, 100) gives an approximate formula for the average number of samples needed to give a decision when v, δ'_L, δ'_H, α and β are given and the actual density is δ. Suppose when the probability of a positive reaction is $p = 1 - \exp(-\delta v)$ that this average number is called $\bar{n}_p$. Then it contains the known quantities α, β, δ'_L, δ'_H and the unknown v, δ. Most efficient sampling in the sense of least average work, can be achieved by minimizing $\bar{n}_p$ with respect to v for some pertinent value of δ. If δ is thought to be fairly low it might be best to minimize $\bar{n}_p$ for $\delta = \delta'_L$; or if it is thought to be high, for $\delta = \delta'_H$. An objective decision would be to minimize $\bar{n}_p$ when this is at its highest. $\bar{n}_p$ has its maximum for a density usually between δ'_L and δ'_H. Best estimation in the doubtful region will follow if $\bar{n}_p$ is minimized for a value of δ between δ'_L and δ'_H. Approximately the highest $\bar{n}_p$ occurs when $p = s$, i.e. for a density δ_s lying in (δ'_L to δ'_H) and satisfying $1 - \exp(-\delta_s v) = s$.

The three $\bar{n}_p$ named are

$$\bar{n}_s = \frac{\log[(1-\alpha)/\beta]\log[(1-\beta)/\alpha]}{\log(p_2/p_1)\log[(1-p_1)/(1-p_2)]};$$

$$\bar{n}_{p_1} = \frac{(1-\alpha)\log[(1-\alpha)/\beta] - \alpha\log[(1-\beta)/\alpha]}{(1-p_1)\log[(1-p_1)/(1-p_2)] - p_1\log(p_2/p_1)}; \quad p_1 = 1-\exp(-\delta'_L v);$$

$$\bar{n}_{p_2} = \frac{(1-\beta)\log[(1-\beta)/\alpha] - \beta\log[(1-\alpha)/\beta]}{p_2\log(p_2/p_1) - (1-p_2)\log[(1-p_1)/(1-p_2)]}; \quad p_2 = 1-\exp(-\delta'_H v).$$

https://doi.org/10.1017/S0022172400000711 Published online by Cambridge University Press

Let $\delta'_L v = x$ and $\delta'_H/\delta'_L = r$. Then the equations to find v obtained by minimizing $\overline{n}_s$, $\overline{n}_{p_1}$, $\overline{n}_{p_2}$ are independent of α and β and contain only r and x. They are—

For $\overline{n}_s$, $$\log_e\left[\frac{e^{rx}-1}{e^x-1}\right]-(r-1)\,x+\frac{r^x}{e^{rx}-1}-\frac{x}{e^x-1}=0.$$

For $\overline{n}_{p_1}$, $$r-r\,\frac{e^x-1}{e^{rx}-1}-\log_e\left[\frac{e^{rx}-1}{e^x-1}\right]=0.$$

For $\overline{n}_{p_2}$, $$1-\frac{e^{rx}-1}{e^x-1}+r\log_e\left[\frac{e^{rx}-1}{e^x-1}\right]=0.$$

3.1

Table 1 gives solutions, x, of these equations for various values of $r=\delta'_H/\delta'_L$. This table also includes the actual values of the $\overline{n}_p$'s when v is taken as x/δ'_L for each of three sets of values of α and β. Because of the approximation in the formula for $\overline{n}_p$ some of these $\overline{n}_p$'s are less than 1. In these cases it is usually impossible, because the boundaries are very close together, for a decision *not* to be reached before a certain low sample size; e.g. for $r=10$, $x=0{\cdot}702$, a decision is inevitable before 7 samples have been taken. While it is advisable therefore to treat the lowest figures with caution, it is generally unlikely that a decision will not have been reached by the time $3\overline{n}_p$ samples have been examined.

The table shows that in some of the extreme cases when r is large, $\overline{n}_s$ is not necessarily near the maximum $\overline{n}_p$ and may be less than $\overline{n}_{p_1}$ or $\overline{n}_{p_2}$. Minimizing n_s has therefore its best validity for low r. It appears, however, from the figures that if $\overline{n}_s$ is minimized and δ'_H/δ'_L is 4 or more, 15 sample volumes of size x/δ'_L will nearly always produce a decision.

3.2. *Example*

In an actual case δ'_L might be legally defined in that water must not contain more than a given number of bacteria per unit volume. It is of course unrealistic to make $\delta'_L=0$, or even to say that water must *never* contain more than this given number, as no test can find out if either of these is so. (Hence the risks α and β must be defined.) One might set δ'_L as less than or equal to this legal minimum. The choice of δ'_H might be more difficult. Clearly one wants to be almost certain no bad water is accepted, but it is apparent from the table that $\overline{n}_s$ is smallest when the ratio of δ'_H to δ'_L is high. For example, if $\delta'_L=5/100$ c.c. and $\delta'_H=20/100$ c.c., $r=4$, $\alpha=\beta=0{\cdot}05$, $x=0{\cdot}861$, and the volume to be taken is $x/\delta'_L=(0{\cdot}861)100/5$ c.c. $=17{\cdot}22$ c.c. It appears that water with an actual density of δ'_L will seldom require more than $3\times 3{\cdot}3\simeq 10$ samples at this volume for a decision to be made. On the average 19 times out of 20 (i.e. $20(1-\alpha)$) the water will be accepted, and once it will be rejected. Water with density less than δ'_L will need fewer samples on the average to produce a decision and there will be a smaller risk than α that it is rejected. Again water with density δ'_H will seldom require more than $6{\cdot}3\times 3\simeq 19$ samples for a decision to be reached. If the density is between δ'_L and δ'_H hardly any more samples than if the density is elsewhere will be needed before

Table 1

	$\bar{n}_s$ minimized				$\bar{n}_{p_1}$ minimized				$\bar{n}_{p_2}$ minimized			
$r=\delta'_H/\delta'_L$	$x=\delta'_L v$	$\bar{n}_s$	$\bar{n}_{p_1}$	$\bar{n}_{p_2}$	x	$\bar{n}_s$	$\bar{n}_{p_1}$	$\bar{n}_{p_2}$	x	$\bar{n}_s$	$\bar{n}_{p_1}$	$\bar{n}_{p_2}$
1·5	1·308	81·0	46·2	54·4	1·457	81·6	45·9	55·7	1·159	81·6	47·3	53·9
		197·2	78·4	92·4		198·7	78·0	94·6		198·8	80·3	91·5
		127·0	72·8	59·7		127·9	72·3	61·1		128·0	74·5	59·2
2·0	1·145	27·4	15·1	20·0	1·367	27·9	14·8	21·5	0·922	28·1	16·1	19·5
		66·7	25·6	34·0		68·0	25·1	36·5		68·4	27·3	33·1
		43·0	23·7	22·0		43·8	23·3	23·6		44·0	25·3	21·4
2·5	1·037	15·5	8·3	12·2	1·302	16·0	8·1	13·8	0·770	16·2	9·3	11·6
		37·7	14·1	20·8		38·9	13·7	23·5		39·4	15·8	19·8
		24·3	13·1	13·4		25·0	12·7	15·2		25·4	14·6	12·8
3·0	0·961	10·6	5·6	9·0	1·255	11·1	5·4	10·7	0·664	11·3	6·6	8·4
		25·9	9·5	15·4		27·0	9·1	18·3		27·6	11·1	14·2
		16·7	8·8	9·9		17·4	8·5	11·8		17·8	10·3	9·2
3·5	0·905	8·1	4·2	7·4	1·218	8·5	4·0	9·2	0·585	8·8	5·1	6·6
		19·6	7·1	12·5		20·6	6·8	15·6		21·3	8·7	11·3
		12·6	6·6	8·1		13·3	6·3	10·1		13·7	8·1	7·3
4·0	0·861	6·5	3·3	6·3	1·189	6·9	3·2	8·3	0·521	8·3	4·3	5·5
		15·8	5·7	10·8		16·7	5·4	14·0		20·2	7·3	9·4
		10·2	5·3	1·1		10·8	5·0	9·1		13·0	6·8	6·1
4·5	0·827	5·4	2·8	5·7	1·165	5·8	2·6	7·7	0·475	6·2	3·7	4·8
		13·2	4·7	9·6		14·0	4·4	13·1		15·0	6·3	8·2
		8·5	4·4	6·2		9·0	4·1	8·4		9·7	5·9	5·3
5·0	0·800	4·7	2·4	5·2	1·146	5·0	2·2	8·1	0·435	5·4	3·3	4·3
		11·4	4·0	8·8		12·1	3·8	13·7		13·2	5·6	7·3
		7·3	3·7	5·7		7·8	3·5	8·9		8·5	5·2	4·7
5·5	0·778	4·1	2·1	4·9	1·131	4·4	1·9	7·0	0·401	4·8	3·0	3·9
		10·0	3·5	8·3		10·7	3·3	12·0		11·8	5·1	6·6
		6·5	3·3	5·4		6·9	3·0	7·7		7·6	4·7	4·3
6·0	0·761	3·7	1·8	4·6	1·118	3·9	1·7	6·8	0·373	4·4	2·8	3·6
		8·9	3·1	7·8		9·6	2·9	11·6		10·7	4·7	6·1
		5·8	2·9	5·1		6·2	2·7	7·5		6·9	4·4	3·9
7·0	0·735	3·0	1·5	4·3	1·098	3·2	1·4	6·6	0·327	3·8	2·4	3·1
		7·4	2·5	7·3		7·9	2·3	11·2		9·2	4·1	5·3
		4·8	2·4	4·7		5·1	2·2	7·2		5·9	3·8	3·4
8·0	0·719	2·6	1·3	4·1	1·084	2·8	1·2	6·4	0·292	3·3	2·2	2·8
		6·3	2·1	7·0		6·7	2·0	11·0		8·1	3·7	4·7
		4·1	2·0	4·5		4·3	1·8	7·1		5·2	3·5	3·1
9·0	0·708	2·3	1·1	4·0	1·073	2·4	1·0	6·3	0·264	3·0	2·0	2·5
		5·5	1·8	6·7		5·9	1·7	10·8		7·3	3·5	4·3
		3·5	1·7	4·4		3·8	1·6	7·0		4·7	3·2	2·8
10·0	0·702	2·0	1·0	3·9	1·064	2·1	0·9	6·3	0·241	2·8	1·9	2·4
		4·9	1·6	6·6		5·2	1·5	10·6		6·7	3·2	4·0
		3·1	1·5	4·3		3·4	1·4	6·9		4·3	3·0	2·6
20·0	0·693	1·0	0·4	3·8	1·029	1·0	0·4	6·0	0·131	1·7	1·4	1·6
		2·3	0·7	6·5		2·4	0·7	10·2		4·2	2·3	2·7
		1·5	6·7	4·2		1·6	6·2	6·6		2·7	2·2	1·7
	0·693	—	—	—	1·000	—	—	—	0·000	—	—	—

For each value of r the first row corresponds to $\alpha=\beta=0{\cdot}05$, the second row to $\alpha=\beta=0{\cdot}01$, the third to $\alpha=0{\cdot}05$ $\beta=0{\cdot}01$.

https://doi.org/10.1017/S0022172400000711 Published online by Cambridge University Press

the water is accepted or rejected. Suppose 15 samples are taken. If the danger of accepting water whose true density is in this region is regarded as serious it might be best to take δ'_L lower than, and δ'_H about equal to, the legal minimum so that the doubtful region is actually almost an acceptable region. If no decision has been reached when about 10 samples have been examined it will be profitable to consider the actual estimate of the density based on the results of the 15 samples.

4. ESTIMATION

Suppose the sample size on the basis of Table 1 has been fixed at n and that r of these samples of volume v have reacted positively. Then the estimation equation for δ is

$$r = n(1 - \exp(-\delta v)),$$

or

$$\delta = (-1/v).\log_e[(n-r)/n].$$

The curves of Clopper & Pearson (1934) give confidence intervals for $\exp(-\delta v)$ (and so for δ since v is known) for a fixed sample size. These will be exact here only if estimation is always made on all the n results. Since sequential tests are suggested, partly for the reduction of work in reaching a decision, an estimate of δ will usually only be made when no decision has been reached before all n samples have been examined. The accuracy of the estimate must then be computed for this conditional situation, and the Clopper-Pearson confidence intervals are then only approximate.

4.1

As an example of their use, suppose as previously that $n = 15$, $x = 0{\cdot}861$. Suppose all n samples are examined and 15 react positively. Then from the Clopper-Pearson curves we say that $\exp(-\delta v)$ lies between 0·84 and 0·34 with confidence coefficient approximately 0·95; then δ lies between 0·01 and 0·06, the ratio of these two extreme limits being 6·19. This compares with Cochran's ratios of 3·46, 5·81, 6·65 and 10·89 for 5 samples of each dilution and dilution ratios of 2, 4, 5 and 10 respectively, when δ is estimated by the most probable number. Cochran's figures are independent of the number of dilutions used. However, the number of dilutions needed to cover a given range (δ_L to δ_H) will go up as the dilution ratio decreases. In fact the number will go up approximately as the inverse of the log of the dilution ratio. Thus the number of samples needed for dilution ratios of 2, 4, 5 and 10, in a series intended to give an estimate of δ when this is somewhere in (δ_L to δ_H), will go down as 3·32 : 1·66 : 1·43 : 1·00 if the same number of samples per volume is used for each dilution ratio.

The smallest number of different volumes used in a series will probably be 3, so that for a dilution ratio of 10 with 5 samples a volume, 15 samples in all will be needed. If the dilution ratio were 2 and the series were used for estimating over the same range, about $15 \times 3{\cdot}32$ or nearly 50 samples would be needed to achieve the greater accuracy of 3·46 as against 10·89 for a dilution ratio of 10, and 6·19 for our single volume sampling, both these latter cases requiring 15 samples.

As an example of a greater number of samples, suppose we have 50 samples of one volume with, say, 20 reacting positively. Then the ratio of the end-points of the

https://doi.org/10.1017/S0022172400000711 Published online by Cambridge University Press

approximate 95 % confidence interval is 2·54. We compare this with, say, a dilution series with ratio 10, and 5 volumes with 10 samples a volume, or with a series covering the same range of δ, with dilution ratio 5, 10 samples a volume and $5 \times 1{\cdot}43 \simeq 7$ volumes; and so on. The ratios of the end-points of the 95 % confidence intervals for dilution ratios 2, 4, 5 and 10 are then 2·40, 3·46, 3·80 and 5·38, while the number of samples used varies roughly as 165 : 80 : 70 : 50. The comparison of our accuracy of 2·54 for 50 samples, with these figures, is favourable as of course it comes from an estimation process catering especially for the middle range of δ, whereas these long series estimate with the same precision over a much wider range.

REFERENCES

CLOPPER, C. J. & PEARSON, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–13.

COCHRAN, W. G. (1950). Estimation of bacterial densities by means of the 'Most Probable Number'. *Biometrics*, **6**, 105–16.

FINNEY, D. J. (1951). The estimation of bacterial densities from dilution series. *J. Hyg., Camb.*, **49**, 26–35.

SWAROOP, S. (1951). The range of variation of the most probable number of organisms estimated by the dilution method. *Indian J. med. Res.* **39**, 107–134.

WALD, A. (1947). *Sequential Analysis*. Wiley.

(*MS. received for publication* 25. I. 55)