SAMPLING ERRORS IN THE DETERMINATION OF BACTERIAL OR VIRUS DENSITY BY THE DILUTION METHOD

By J. B. S. HALDANE, F.R.S.
Department of Biometry, University College, London

In certain bacteriological investigations it is only practicable to obtain an estimate of density by diluting the culture until a substantial fraction of all inoculations are negative. This is still more necessary when dealing with viruses where microscopic methods are at best unreliable and at worst impossible. Parker (1938) has applied the dilution method with great success to the problem of vaccinia virus. The mathematical method was given by Halvorsen & Ziegler (1933), but while their estimate is entirely correct they do not give the standard error of the bacterial density estimated, and it seems worth doing so.

Suppose that we have made a number of inoculations with amounts of the original suspension equal to c_1, c_2, c_3, ..., c_r, ..., c_n times a standard amount. In the case summarized in Table I (from Parker, 1938) the standard amount is 0·25 c.c. of a 10^{-9} dilution of Board of Health vaccine, and the values of c_r are 2, 5, 10, 20, 50, 100, 200. Let the concentration c_r give a_r successes and b_r failures. Thus $c_3 = 10$, $a_3 = 14$, $b_3 = 58$.

We desire to obtain an estimate of the expected number of virus particles in a unit (here $2{\cdot}5 \times 10^{-10}$ ml.) of vaccine. If the c_r's are integers, x will be a small fraction.

In the concentration c_r the expected number of bacteria is xc_r. Thus the expected value of the failures is $E(b_r) = e^{-xc_r}(a_r + b_r)$ and of successes

$$a_r = (1 - e^{-xc_r})(a_r + b_r).$$

If we only work with one concentration the likelihood of a failures and b successes is

$$\frac{(a+b)!}{a!\,b!}(1 - e^{-xc})^a e^{-bcx}.$$

This is maximal when its logarithm

$$L = \text{constant} + a \log (1 - e^{-xc}) - bcx$$

is maximal.

Hence

$$\frac{dL}{dx} = \frac{ace^{-cx}}{1 - e^{-cx}} - bc = 0,$$

or

$$e^{cx} = \frac{a+b}{b}, \qquad \ldots\ldots(1)$$

as is otherwise obvious.

https://doi.org/10.1017/S002217240001192X Published online by Cambridge University Press

We can next calculate the standard deviation σ of x, or the amount of information I concerning it, from the equation

$$I=\sigma^{-2}=\frac{-d^2L}{dx^2}=\frac{ac^2e^{cx}}{(e^{cx}-1)^2}$$
$$=\frac{bc^2(a+b)}{a}. \qquad \text{.....(2)}$$

We desire that σ/x should be as small as possible, that is to say that Ix^2 should be as large as possible. If $cx=m$ (the expected number of bacteria) per sample, and if $a+b=s$, then

$$Ix^2=\frac{m^2s}{e^m-1}.$$

Hence $\frac{e^m-1}{m^2}$ must be a minimum. Differentiating, we find $e^m=\frac{2}{2-m}$. Hence $m=1{\cdot}59363$, $e^{-m}=0{\cdot}20318$.

Thus the most accurate results will be obtained with a dilution which contains on an average 1·6 particles, and gives 80 % positive results (Fisher, 1921, 1938).

The standard error of m is $\sqrt{\frac{e^m-1}{s}}$. In this case $\frac{\sigma_m}{m}=\frac{1{\cdot}24}{\sqrt{s}}$, whereas, to take two examples, when 50 % of samples are positive $e^m=2$, $m=0{\cdot}69315$, $\frac{\sigma_m}{m}=\frac{1{\cdot}44}{\sqrt{s}}$, and when 90 % are positive $e^m=10$, $m=2{\cdot}30259$, $\frac{\sigma_m}{m}=\frac{1{\cdot}30}{\sqrt{s}}$. Thus the accuracy does not vary very greatly, but the best results will be obtained when the majority of samples are positive.

Returning to the case where a number of concentrations are used, we have

$$\left.\begin{aligned} L&=\text{constant}+\Sigma a_r\log(1-e^{-c_rx})-x\Sigma b_rc_r,\\ \frac{dL}{dx}&=\Sigma\frac{a_rc_r}{e^{c_rx}-1}-\Sigma b_rc_r=0. \end{aligned}\right\} \qquad \text{.....(3)}$$

This equation was obtained by Halvorsen & Ziegler, who give tables which are sometimes useful for solving such equations.

$$\left.\begin{aligned} I&=\Sigma\frac{a_rc_r^2e^{c_rx}}{(e^{c_rx}-1)^2},\\ \text{and}\quad \sigma_x&=I^{-\frac{1}{2}}. \end{aligned}\right\} \qquad \text{.....(4)}$$

Let us apply these equations to Table I.

Equation (3) gives

$$\frac{2}{e^{2x}-1}+\frac{30}{e^{5x}-1}+\frac{140}{e^{10x}-1}+\frac{400}{e^{20x}-1}+\frac{2450}{e^{50x}-1}+\frac{5800}{e^{100x}-1}+\frac{13400}{e^{200x}-1}=4745. \qquad \text{.....(5)}$$

Such equations can be solved within an hour or so with no tables beyond an ordinary table of logarithms, by successive approximations. As a first approximation let us solve equation (1) for the bottom row but one, namely, $a_r=58$, $b_r=11$, $c_r=100$, which is likely, as appears from the above calculation, to give

https://doi.org/10.1017/S002217240001192X Published online by Cambridge University Press

a more accurate result than any of the other dilutions. We have $e^{100x} = \frac{69}{11}$, whence $e^x = 1{\cdot}01853$, $x = 2{\cdot}3026 \log_{10} 1{\cdot}01853 = 0{\cdot}01836$.

As a preliminary attempt to solve equation (5) we put, on the left-hand side, $x = 0{\cdot}0185$. Hence $\log_{10} e^x = 0{\cdot}4343x = 0{\cdot}00803$, and the denominators can be readily calculated. The sum of the left-hand side is 4977, which is too large. So a higher value of x must be tried. Putting $x = 0{\cdot}195$ we get the value 4555, which is too low. $x = 0{\cdot}190$ gives 4767·5, which exceeds 4745 by 13·5 only. Since a difference of 0·01 in x causes a difference of 422 in the total, this implies that $x = 0{\cdot}01903$. However, in view of the standard error to be found later, 0·0190 is a sufficiently accurate value.

The amount of information about x is

$$I = \frac{4e^{2x}}{(e^{2x}-1)^2} + \frac{150\,e^{5x}}{(e^{5x}-1)^2} + \frac{1400\,e^{10x}}{(e^{10x}-1)^2} + \frac{8000\,e^{20x}}{(e^{20x}-1)^2} + \frac{122{,}500e^{50x}}{(e^{50x}-1)^2} + \frac{580{,}000\,e^{100x}}{(e^{100x}-1)^2} + \frac{2{,}680{,}000\,e^{200x}}{(e^{200x}-1)^2}.$$

The values of such quantities as $e^{20x}-1$ have already been determined during the calculation of x. Hence we readily find

$$I = 421{,}315; \quad \sigma = 0{\cdot}0015.$$

Thus $$x = 0{\cdot}0190 \pm 0{\cdot}0015.$$

This is the expected number in $2{\cdot}5 \times 10^{-10}$ c.c. of the original vaccine. In other words 1 c.c. of the vaccine contains $(76 \pm 6) \times 10^6$ virus particles.

As an alternative to solving equation (5) we may solve equations (1) and (2) separately for each concentration, and take $x = \frac{\Sigma x_r I_r}{\Sigma I_r}$, $I = \Sigma I_r$. This gives $x = 0{\cdot}0182 \pm 0{\cdot}0015$.

The result is not quite so accurate. But the calculation, which is shown in Table I, is very rapid, and quite good enough for most practical purposes.

Parker took the point of 50 % success at $c = 40$. This gives $e^{-cx} = \frac{1}{2}$, or $x = 0{\cdot}0173$. This is a possible value, but distinctly on the low side.

The goodness of fit may be calculated by the χ^2 method. If a_r' and b_r' be the calculated values of a_r and b_r, and $d = a_r - a_r' = b_r' - b_r$, then

$$\chi^2 = \Sigma\left(\frac{d^2}{a_r'} + \frac{d^2}{b_r'}\right) = \Sigma \frac{d^2\,(a_r + b_r)}{a_r' b_r'}.$$

Unfortunately, Parker used a formula which in my notation is $\Sigma \frac{d^2}{a_r'}$. His value is thus too low. My estimate of x gives $\chi^2 = 3{\cdot}53$. The number of degrees of freedom is 6, for one has been lost by estimating x. Thus $P = 0{\cdot}74$. Parker's theory, when χ^2 is correctly calculated, gives $\chi^2 = 3{\cdot}59$. A value of x intermediate between Parker's and my own would give a minimum value of χ^2. However,

this particular test is not the best possible for estimating x, though it is very nearly the best for testing the consistency of the data, that is to say whether they agree with the particular theory in question, as Parker's very clearly do.

Table I. *Parker's data on "Board of Health" vaccine*

c_r	a_r	b_r	x_r	I_r	$I_r x$	b_r' (calc.)	χ^2
2	1	75	0·006625	22,800	152·2	73·2	1·23
5	6	69	0·016676	21,563	359·6	68·2	0·11
10	14	58	0·021622	29,829	645·0	59·5	0·23
20	20	51	0·016543	72,420	1198·0	48·6	0·39
50	49	23	0·022823	84,490	1894·0	27·8	1·38
100	58	11	0·018362	130,862	2402·9	10·3	0·05
200	67	2	0·017705	82,388	1457·7	1·5	0·14
				444,352	8109·4		3·53

c_r = dose of vaccine, the unit being $2{\cdot}5 \times 10^{-10}$ c.c.
a_r = number of positive inoculations.
b_r = number of negative inoculations.
x_r = deduced value of x, the expected number of virus particles in a unit dose.
I_r = amount of information concerning x_r.
b_r' is calculated from the weighted mean value of x.

We can easily see how to make the corresponding calculation if just two or more particles are needed to cause infection. Thus in the case of two particles, and a single dilution

$$E(b) = e^{-cx}(1+cx)(a+b).$$

So the estimate of x is given by

$$\frac{e^{cx}}{1+cx} = \frac{a+b}{b}. \qquad \text{.....(6)}$$

Or if $e^{cx} = m$, $\log_e m = \frac{bm}{a+b} - 1$, which can readily be solved with a table of natural logarithms.

$$L = \text{constant} + a \log [1 - e^{-cx}(1+cx)] + b \log]e^{-cx}(1+cx)].$$

Hence

$$I = \frac{b(a+b)c^4x^2}{a(1+cx)^2}. \qquad \text{.....(7)}$$

We can readily find the weighted mean $\Sigma \frac{I_r x_r}{\Sigma I_r}$ of the values of x so obtained for different dilutions. Parker's data are so treated in Table II.

$$x = 0{\cdot}0352 \pm 0{\cdot}0024.$$

(Parker's calculated values appear to correspond to $x = 0{\cdot}066$.) The trend of the different x_r values is obvious, and it is clear that this hypothesis will not fit the facts.

The χ^2 is also large. Parker's value, correctly calculated, is 30·4, which is much smaller. Where data do not agree with the hypothesis it is naturally to be expected that the value of x giving the maximum likelihood will not give the minimum χ^2.

https://doi.org/10.1017/S002217240001192X Published online by Cambridge University Press

Table II. *Parker's data on the hypothesis that two particles are needed for an infection*

c_r	a_r	b_r	x_r	I_r	$I_r x_r$	b_r'	χ^2
2	1	75	0·08381	547	45·84	75·820	3·744
5	6	69	0·09314	1,624	151·33	73·965	12·283
10	14	58	0·08090	5,966	482·66	68·459	32·528
20	20	51	0·05237	37,153	1947·70	59·975	8·649
50	49	23	0·04700	42,479	1998·64	34·186	6·977
100	58	11	0·03293	23,541	775·21	9·230	0·392
200	67	2	0·02698	58,803	1586·50	0·486	4·750
				170,113			69·321

SUMMARY

The statistical theory of virus determination by the dilution method is considered. It is shown that an algebraical method gives a solution which is rather more reliable than that given by the graphical method, and whose standard error may be calculated.

REFERENCES

FISHER, R. A. (1921). On the mathematical foundations of theoretical statistics. *Philos. Trans.* A, **222**, 309–68.

—— (1938). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

HALVORSEN, H. O. & ZIEGLER, N. R. (1933). Application of statistics to problems in bacteriology. I. A means of determining bacterial population by the dilution method. *J. Bact.* **25**, 101–21.

PARKER, R. F. (1938). Statistical studies of the nature of the infectious unit of vaccine virus. *J. exp. Med.* **67**, 725–38.

(*MS. received for publication* 25. II. 39.—Ed.)