J. Hyg., Camb. (1968), **66**, 531
Printed in Great Britain

The estimation of carrier rate from amoebic surveys

By J. ROBINSON
University of Sydney

(*Received* 1 *March* 1968)

1. INTRODUCTION

The purpose of surveys of amoebae or other organisms is the estimation of the carrier rate in a population, that is, of the proportion of persons who carry a given parasite. As a rule, the tests do not infallibly detect the presence of the organism in a known carrier. Thus simple estimation from a single examination of each member of a sample is not applicable.

Some progress has been made by Lancaster (1950), who defined a measure, the demonstrability, as the probability of a carrier being detected in one examination. His results indicated that this demonstrability varied between organisms and between persons, but remained constant over a period of time. Lancaster (1950) showed that the results of surveys could be misinterpreted if the demonstrability was assumed constant over the population of carriers. As examples he constructed mathematical models in which the demonstrability had the beta distribution. It was clear from such models that other authors had made inappropriate inferences from data collected from surveys. Lancaster (1950) reconsidered a number of surveys and showed that the estimates of carrier rates should be increased.

Here, procedures for the estimation of carrier rate are proposed, based on the assumption that the demonstrability is constant for any person between examinations, but is a random variable with some beta distribution over the population of carriers. The methods are applied to the published data of Boeck & Stiles (1923), McCoy (1936) and Andrews (1934). The expected numbers of positives in each examination and a chi-square goodness of fit statistic are calculated. Although no test of significance is available for the increase in goodness of fit due to assuming variability in the demonstrability, it is clear in many of the examples that an increase has been achieved. This indicates that better estimation of the carrier rate has been achieved by the methods proposed for the model with varying demonstrability. Finally, a simulation experiment has been performed under the models assumed, to determine the order of errors inherent in the estimation procedure.

2. METHOD OF ESTIMATION

Consider a large population with a proportion p of carriers of some parasite. A sample of N is taken from the population and each member of the sample is examined. If they are found positive they are not re-examined. Of those not found positive there may be a number of withdrawals and the remainder are re-examined and some found positive at this second examination. This process is continued until some predetermined number of examinations has been carried out.

We will consider some simpler cases which are useful preliminaries to the discussion of the solution of this general problem.

2·1. *No withdrawals*

In this section it will be assumed that each individual of the sample is examined either until an examination is positive or until he has been examined a given number of times.

2·1·1. *Equal demonstrability*

Suppose all carriers have an equal probability P of giving a positive result in any one examination. Let X_i $(i = 1, \ldots, t)$ be the number of positives in the ith examination. Then the probability of an individual being found positive in the ith examination is pPQ^{i-1}, where $Q = 1-P$, since he must be a carrier, he must obtain negative results in the first $i-1$ examinations and then a positive result in the ith examination. The probability of an individual not being found positive in the first t examinations is $q+pQ^t$. The results of a survey and the expectations given by the model to be fitted under the assumptions of this section may be easily seen in tabular form (Table 1).

Table 1

Examination	No. examined	No. of positives	Expected no. of positives
1	N	X_1	NpP
2	$N-X_1$	X_2	$NpPQ$
...	...	...	...
t	$N-\sum_{i=1}^{t-1} X_i$	X_t	$NpPQ^{t-1}$

The number with no positives is $N-\sum_{i=1}^{t} X_i$ and its expectation is $N[1-p(1-Q^t)]$.

Approximations to the maximum likelihood estimates of p and P may be obtained using an iterative procedure. The formulae used for this calculation and a brief description of the method are given in the appendix. The procedure is a simple one using a computer but it would be somewhat laborious with a desk calculator.

A test of goodness of fit may be obtained by using the statistic

$$X^2 = \sum_{i=1}^{t} \frac{(X_i - NpPQ^{i-1})^2}{NpPQ^{i-1}} \tag{1}$$

as a chi-square variate with $t-2$ degrees of freedom, where we have used the same notation for the parameters p and P, and their estimates. No contribution is obtained from the frequency of negatives since the maximum likelihood solution equates this to its expectation.

2·1·2. *Unequal demonstrability*

Now assume that each carrier in the population has a fixed probability P of giving a positive result in any examination, but that P varies between carriers. Further, assume that the frequency function of P is of the form

$$f(P) = P^{r-1}(1-P)^{s-1}/B(r,s). \tag{2}$$

https://doi.org/10.1017/S0022172400028278 Published online by Cambridge University Press

It was suggested by Bailey (1956), who considered a similar problem with a chain binomial model, that the parameters r and s be replaced by the parameters $\bar{P} = r/(r+s)$ and $Z = 1/(r+s)$. Then Table 1 may be replaced by Table 2.

Table 2

Examination	No. examined	No. of positives	Expected no. of positives
1	N	X_1	$Np\bar{P}$
2	$N - X_1$	X_2	$Np\bar{P}\bar{Q}/(1+Z)$
...	...	...	...
t	$N - \sum_{i=1}^{t-1} X_i$	X_t	$\frac{Np\bar{P}\bar{Q}\ldots(\bar{Q}+(t-2)Z)}{(1+Z)\ldots(1+(t-1)Z)}$

The number with no positives after t examinations is $N - \sum_{i=1}^{t} X_i$ and its expectation is

$$N\{1-p[1-\bar{Q}(\bar{Q}+Z)\ldots(\bar{Q}+(t-1)Z)/(1+Z)\ldots(1+(t-1)Z)]\}.$$

Here the likelihood equations become intractable so it is proposed that another iterative procedure be used. This procedure consists of using as an estimate of p the maximum likelihood estimate of p under the assumption that $\bar{P}$ and Z are known. This is given by the formula

$$p = \sum_{i=1}^{t} X_i / N[1-\bar{Q}(\bar{Q}+Z)\ldots(\bar{Q}+(t-1)Z)/(1+Z)\ldots(1+(t-1)Z)]. \tag{3}$$

The estimates of $\bar{P}$ and Z are obtained by minimizing the chi-square goodness of fit statistic

$$X_0^2 = \sum_{i=1}^{t} \frac{[X_i - Np\bar{P}\bar{Q}\ldots(\bar{Q}+(i-2)Z)/(1+Z)\ldots(1+(i-1)Z)]^2}{Np\bar{P}\bar{Q}\ldots(\bar{Q}+(i-2)Z)/(1+Z)\ldots(1+(i-1)Z)}. \tag{4}$$

The iterative procedure is as follows: Take $Z_1 = 0$ as a first approximation for Z, and estimate p and $\bar{P}$ by the methods of section 2·1·1. Then take these values of p, $\bar{P}$ as p_1, $\bar{P}_1$, the first approximations for p, $\bar{P}$. Now consider a grid of 25 points $(\bar{P}, Z)$, where $\bar{P}$ takes the values $\bar{P}_1$, $\bar{P}_1 \pm \delta_1$, $\bar{P}_1 \pm 2\delta_1$ and Z takes the values 0, Δ_1, $2\Delta_1$, $3\Delta_1$, $4\Delta_1$. δ_1 and Δ_1 are chosen so that the grid covers a suitable range of values of $(\bar{P}, Z)$. For every point of the grid calculate first p using the formula (3) and then X_0^2 using the formula (4), then take, as the next approximation $(p_2, \bar{P}_2, Z_2)$, those points which minimized X_0^2 among the 25 points of the grid. Now take $\delta_2 = \frac{1}{2}\delta_1$ and $\Delta_2 = \frac{1}{2}\Delta_1$, and take as the next refinement of the grid the 25 points $(\bar{P}, Z)$, where $\bar{P}$ takes the values $\bar{P}_2$, $\bar{P}_2 \pm \delta_2$, $\bar{P}_2 \pm 2\delta_2$ and Z takes the values Z_2, $Z_2 \pm \Delta_2$, $Z_2 \pm 2\Delta_2$, if $Z_2 \neq 0$, and the values 0, Δ_2, $2\Delta_2$, $3\Delta_2$, $4\Delta_2$ if $Z_2 = 0$. Now the procedure above is repeated and the next approximation, $(p_3, \bar{P}_3, Z_3)$ is obtained. The iterative procedure is continued until δ_i and Δ_i are of a specified size. Here δ_1 and Δ_1 were chosen to be 0·02 and 0·25 respectively.

2·2. *Withdrawals—treatment by life tables*

In practice in most surveys there are withdrawals resulting in incomplete data. However, if it is assumed that withdrawals are independent of the probability of

https://doi.org/10.1017/S0022172400028278 Published online by Cambridge University Press

detecting a parasite in any examination, then life tables may be constructed in the way shown in Table 3.

An illustration of a constructed life table is given in Table 4 from the data of Boeck & Stiles (1923, p. 20).

Approximate estimates in the case of equal demonstrability can be obtained as in section 2·1·1 by using the life tables in the place of the data. However, the chi-square goodness of fit statistics in this case should be obtained from Table 5.

Table 3

	Data		Life table	
Examination	No. examined	No. positive	No. examined	No. positive
1	n_1	x_1	N	$X_1 = x_1 N/n_1$
2	n_2	x_2	$N - X_1$	$X_2 = x_2(N - X_1)/n_2$
...	...	...	...	...
t	n_t	x_t	$N - \sum_{i=1}^{t-1} X_i$	$X_t = x_t \left(N - \sum_{i=1}^{t-1} X_i\right) \Big/ n_t$

Table 4. *The data of Boeck & Stiles* (1923, *p.* 20) *treated by life-table methods.* Entamoeba coli

	Data		Life table	
Examination	No. examined	No. positive	No. examined	No. positive
1	8,029	1,269	100,000	15,805
2	1,441	155	84,195	9,056
3	1,050	73	75,138	5,224
4	912	44	69,915	3,373
5	791	27	66,541	2,271
6	623	13	64,270	1,341

Table 5

No. positive	Expected no. positive with constant P	Expected no. positive with varying P
x_1	$e_1 = \frac{n_1}{N} NpP$	$e_1^* = \frac{n_1}{N} Np\bar{P}$
x_2	$e_2 = \frac{n_2}{N - X_1} NpPQ$	$e_2^* = \frac{n_2}{N - X_1} Np \frac{\bar{P}\bar{Q}}{(1+Z)}$
...	...	...
x_t	$e_t = \frac{n_t}{N - \sum_{i=1}^{t-1} X_i} NpPQ^{t-1}$	$e_i^* = \frac{n_t}{N - \sum_{i=1}^{t-1} X_i} Np\bar{P} \frac{\bar{Q} \ldots (\bar{Q} + (t-2)Z)}{(1+Z) \ldots (1+(t-1)Z)}$

Then the goodness of fit statistics are

$$X^2 = \sum_{i=1}^{t} \frac{(x_i - e_i)^2}{e_i},$$

when the demonstrability is constant, and

$$X_0^2 = \sum_{i=1}^{t} \frac{(x_i - e_i^*)^2}{e_i^*}.$$

https://doi.org/10.1017/S0022172400028278 Published online by Cambridge University Press

Table 6

Parasite	Constant demonstrability			Varying demonstrability					
	p	P	X^2	p	$\bar{P}$	Z	X_0^2	N	C
E. coli†	0·39	0·39	1·8	0·41	0·38	0·09	0·2	8029	3313
Endolimax nana†	0·30	0·31	25·1**	0·46	0·23	0·38	5·6	8029	3683
E. histolytica†	0·18	0·16	2·0	0·18	0·16	0	2·0	8029	1463
Endolimax williamsi†	0·10	0·35	32·6**	0·16	0·27	0·68	9·8*	8029	1278
Unencysted amoebae†	0·11	0·43	60·6**	0·18	0·36	1·23	7·4	8029	1433
Chilomastix†	0·15	0·17	21·4**	0·15	0·17	0	21·4**	8029	1231
Giardia lamblia†	0·14	0·31	51·0**	0·24	0·23	0·68	11·7**	8029	1926
Blastocystis†	0·52	0·49	161·0**	0·75	0·41	1·02	11·0*	8029	6026
Phycomycete†	0·36	0·15	1·3	0·36	0·15	0	1·3	8029	2928
E. coli‡	0·63	0·46	13·2*	0·84	0·38	0·64	1·6	505	426
Endolimax nana‡	0·33	0·23	4·8	0·34	0·22	0·02	4·8	505	173
E. histolytica‡	0·19	0·24	4·5	0·20	0·22	0·03	4·5	505	103
Giardia lamblia‡	0·17	0·35	17·1**	0·27	0·27	0·70	8·3	505	136
Blastocystis‡	0·68	0·25	1·5	0·68	0·25	0	1·5	505	343
E. histolytica§	0·49	0·34	7·7	0·58	0·30	0·16	6·0	1176	680
G. intestinalis‖	0·22	0·44	18·9**	0·32	0·36	0·79	1·7	1713	541
E. coli‖	0·33	0·44	6·3	0·38	0·40	0·23	4·2	1713	647
E. histolytica‖	0·16	0·32	7·3	0·23	0·25	0·35	4·5	1713	401

† Data of Boeck & Stiles (1923, p. 20).
‡ Data of Boeck & Stiles (1923, p. 25).
§ Data of McCoy—Hotel X (1936).
‖ Data of Andrews (1934).
* Significant at $P < 0·05$
** Significant at $P < 0·01$

for unequal demonstrability. Estimates in the latter case are obtained by using the same procedure as in section 2·1·2 to minimize X_0^2. A test of goodness of fit may be obtained by using the statistics X_0^2 as a chi-square variate with $t-3$ degrees of freedom when P is assumed to vary and X^2 as a chi-square variate with $t-2$ degrees of freedom when P is assumed constant.

3. APPLICATION TO PUBLISHED DATA

The data of Boeck & Stiles (1923, pp. 20 and 25), McCoy (1936) and Andrews (1934) have been put in appropriate form and estimates of the proportion of carriers have been made by the methods described in section 2. In all these cases, except for that of Boeck & Stiles (1923, p. 25), the data suffered from withdrawals. The data of Boeck & Stiles (1923, p. 25) were not subject to withdrawals since their surveys were of persons in institutional life. In fact here six examinations were made on each person. However, the data were in a form such that details of the number of positive examinations per person were not available and so estimation of the proportion of carriers has been done using only information on first positive examinations.

Table 6 gives the final results for all cases where estimates were made and it is clear that in some cases a considerable improvement in goodness of fit has resulted from using the model assuming variability of the demonstrability. An appropriate test for the existence of variability of the demonstrability would be a test of the hypothesis that $Z = 0$. However, no estimates of the variance of Z are available, so it is not possible to perform this test. Even so, if Z is large and X_0^2 is considerably less than X^2, it is clear that the demonstrability is not constant. Estimates of p based on the assumption that $Z = 0$ are biased down, so estimates of carrier rate based on this assumption will be, in general, too low.

Table 7. *The data of Boeck & Stiles* (1923, p. 20). Endolimax nana

No. examined	No. positive	Estimated no. positive assuming constant demonstrability		Estimated no. positive assuming variable demonstrability	
8029	855		744·1		847·6
1441	82		103·3		94·8
1050	43		55·2		47·8
912	40		34·5		30·9
791	25		21·6		21·3
623	13		12·2		13·7
		p	0·30	p	0·46
		P	0·31	$\bar{P}$	0·23
				Z	0·38
		X^2	25·1**	X_0^2	5·6

It is interesting to notice that estimates of $\bar{P}$, the mean demonstrability, vary quite markedly for the different parasites and surveys.

It seems worth while for the purpose of illustration to consider some cases in detail. Four have been chosen and these are set out in detail in Tables 7–10. In Tables 7 and 8, Z is reasonably large and an improvement in goodness of fit, due to fitting the model with the assumption of variability of the demonstrability, is

evident, not only in comparisons of X^2 and X_0^2 but also in a direct consideration of the tables themselves. In Table 9 the results are not so clear. There seems to be little increase in goodness of fit, but Z is not small and the increase in the estimate of p is still marked. In table 10, Z is small, there is little difference in goodness of fit and the estimates of p are not markedly different. Consideration of the two tables, 9 and 10, indicates that the appropriate measure of improvement in estimation is Z and not a comparison of X^2 and X_0^2.

Table 8. *The data of Andrews* (1934). G. intestinalis

No. examined	No. positive	Estimated no. positive assuming constant demonstrability		Estimated no. positive assuming variable demonstrability	
1713	194	164·1		195·4	
1560	71	94·5		71·7	
1093	29	38·9		29·2	
459	6	9·4		8·3	
301	4	3·5		4·0	
236	4	1·6		2·4	
		p	0·22	p	0·32
		P	0·44	$\bar{P}$	0·36
				Z	0·79
		X^2	18·9**	X_2^0	1·7

Table 9. *The data of McCoy* (1936)—*Hotel X*. E. histolytica

No. examined	No. positive	Estimated no. positive assuming constant demonstrability		Estimated no. positive assuming variable demonstrability	
1176	203	197·2		206·0	
876	123	116·2		110·0	
670	56	69·2		64·2	
558	33	41·6		40·1	
454	32	23·8		24·9	
202	8	7·5		8·9	
		p	0·49	p	0·58
		P	0·34	$\bar{P}$	0·30
				Z	0·16
		X^2	7·7	X_0^2	6·0

Table 10. *The data of Boeck & Stiles* (1923, p. 20). E. coli

No. examined	No. positive	Estimated no. positive assuming constant demonstrability		Estimated no. positive assuming variable demonstrability	
8029	1269	1234·1		1268·5	
1441	155	159·4		152·6	
1050	73	78·8		74·6	
912	44	44·6		43·7	
791	27	24·6		26·0	
623	13	12·2		14·3	
		p	0·39	p	0·41
		P	0·39	$\bar{P}$	0·38
				Z	0·09
		X^2	1·8	X_0^2	0·2

https://doi.org/10.1017/S0022172400028278 Published online by Cambridge University Press

In many cases the models have enabled a very good fit to the data to be made. This suggests, at least, that the model and estimates are realistic. It is worth noting that in all cases the estimates of carrier rate are higher than those suggested in the past. A comparison of two of the estimates of the carrier rates of *E. histolytica*, 0·57 for McCoy (1936) and 0·18 for Boeck & Stiles (1923, p. 20), with the minimum estimates of the carrier rates proposed by Lancaster (1950) for these two surveys, 0·52 and 0·12 respectively, shows agreement of these estimates with his conditions.

4. SIMULATION EXPERIMENTS

The methods used in the estimation procedures do not enable us to evaluate the standard errors of the estimates, so the accuracy of the methods is uncertain. To ascertain the order of accuracy involved, two simulation experiments were performed and empirical means and standard errors of the estimates were calculated. The first experiment was performed assuming the model of equal demonstrability and in the second experiment the demonstrability was supposed to be a random variable with a specified beta distribution. In each of these cases the expected values and the variances and covariances of the number of positives at each examination were calculated. Random normal variables with zero expectations and variances and covariances equal to the calculated values were added to the calculated expected values. These values were then used as the simulated data of a survey and analysed in the same way as actual survey data. This process was repeated 20 times for each experiment and empirical means and standard errors were calculated.

In the first experiment the parameters were taken to be $N = 10{,}000$, $p = 0{\cdot}3$ and $P = 0{\cdot}3$. The empirical means of the estimates of the two parameters p and P were 0·3011 and 0·3008 respectively and the empirical standard errors per experiment of the estimates of these two parameters were 0·0069 and 0·0087 respectively. These standard errors are quite small, as could be expected from the large numbers in the simulated samples.

In the second experiment the parameters were taken to be $N = 10{,}000$, $p = 0{\cdot}3$, $\bar{P} = 0{\cdot}3$ and $Z = 0{\cdot}4$. The empirical means of the estimates of the parameters p, $\bar{P}$ and Z were 0·2943, 0·3075 and 0·3856 respectively and the empirical standard errors per experiment of the estimates of these three parameters were 0·0195, 0·0211 and 0·0593 respectively. The methods are still fairly efficient in this case, but the efficiency has been considerably reduced from that of the case of equal demonstrability. This drop in efficiency is to be expected, since there is, in this case, a large number of carriers with very low demonstrability and it is clearly difficult to separate these carriers from persons not infected.

These experiments have demonstrated that the estimates given by these methods are unbiased but that, when the demonstrability has large variation between carriers, the efficiency of the estimation procedure is lowered. Thus quite large sample numbers are necessary to ensure accurate estimation of the carrier rate when the demonstrability varies between carriers.

I wish to thank Prof. H. O. Lancaster for his suggestion of this problem and for his help in the preparation of this paper.

REFERENCES

ANDREWS, J. (1934). The diagnosis of intestinal protozoa from purged and normally passed stools. *J. Parasit.* **20**, 252.

BAILEY, N. T. J. (1956). Significance tests for a variable chance of infection in chain-binomial theory. *Biometrika* **43**, 332.

BOECK, W. C. & STILES, C. W. (1923). Studies on various intestinal parasites (especially amoebae) of man. *Bull. hyg. Lab., Wash.* no. 133.

LANCASTER, H. O. (1950). The theory of amoebic surveys. *J. Hyg., Camb.* **48**, 257.

MCCOY, G. (1936). Epidemic amoebic dysentery. The Chicago outbreak of 1933. *Natn Inst. Hlth Bull.* no. 166.

RAO, C. R. (1966). *Linear Statistical Inference and its Applications.* New York: John Wiley Inc.

APPENDIX

The likelihood function for the numbers of positives at each examination is

$$P(X_1, \ldots, X_t; N) = \frac{N!(pP)^{X_1}(pPQ)^{X_2}\ldots\{1-p(1-Q^t)\}^{N-\sum_{i=1}^{t} X_i}}{X_1!\ldots X_t!(N-\sum_{i=1}^{t} X_i)!}.$$

Now using the logarithm of this likelihood function we may use the so-called 'method of scoring' (see, for example, Rao (1966, pp. 302–309)) to obtain approximations to the maximum likelihood estimates by an iterative procedure.

Formulae are required for the values of the first and second derivatives of the logarithm of the likelihood function, $L = \log P$, with respect to p and P, and these are set out below for $p = p_j$ and $P = P_j$.

$$S_p^j = \left(\frac{\partial L}{\partial p}\right)_{p_j, P_j} = \frac{\Sigma X_i}{p_j} - \frac{(N-\Sigma X_i)(1-Q_j^t)}{1-p_j(1-Q_j^t)},$$

$$S_P^j = \left(\frac{\partial L}{\partial P}\right)_{p_j, P_j} = \frac{\Sigma X_i}{P_j} - \frac{\Sigma(i-1)X_i}{Q_j} - \frac{(N-\Sigma X_i)\, tp_jQ_j^{t-1}}{1-p_j(1-Q_j^t)}$$

$$-I_{pp}^j = \left(\frac{\partial^2 L}{\partial p^2}\right)_{p_j, P_j} = -\frac{\Sigma X_i}{p_j^2} - \frac{(N-\Sigma X_i)(1-Q_j^t)^2}{\{1-p_j(1-Q_j^t)\}^2}$$

$$-I_{pP}^j = \left(\frac{\partial^2 L}{\partial p\, \partial P}\right)_{p_j, P_j} = -\frac{(N-\Sigma X_i)tQ_j^{t-1}}{\{1-p(1-Q_j^{t-1})\}^2}$$

$$-I_{PP}^j = \left(\frac{\partial^2 L}{\partial P^2}\right)_{p_j, P_j} = -\frac{\Sigma X_i}{P_j^2} - \frac{\Sigma(i-1)X_i}{Q_j^2} + \frac{(N-\Sigma X_i)t(t-1)p_jQ_j^{t-2}}{1-p_j(1-Q_j^t)} - \frac{(N-\Sigma X_i)t^2p_j^2Q_j^{2(t-1)}}{\{1-p_j(1-Q_j^t)\}^2}$$

First estimates of p and P were taken to be $p_1 = \Sigma X_i/N$ and $P_1 = X_1/\Sigma X_i$. The iterative procedure consists of calculating δp_j and δP_j from the equations

$$S_p^j = I_{pp}^j\, \delta p_j + I_{pP}^j\, \delta P_j,$$

$$S_P^j = I_{pP}^j \delta p_j + I_{PP}^j \delta P_j,$$

and putting $P_{j+1} = P_j + \delta P_j$ and $p_{j+1} = p_j + \delta p_j$. The iteration may be stopped when a predetermined level of approximation is obtained.