A Simple Method of Estimating the Segregation Ratio under Complete Ascertainment

C. C. Li[1] and Nathan Mantel[2]

The problem of estimating the segregation ratio is so familiar to human geneticists that we need only a few brief remarks to introduce the subject. Consider a pair of alleles (A, a) with A completely dominant to a, so that the genotypes AA and Aa are of one phenotype (normal) and aa is of another (affected with a disease). We are concerned here with estimating the segregation ratio in the offspring of two heterozygous parents ($Aa \times Aa$) which are phenotypically indistinguishable from nonsegregating families ($AA \times AA$ and $AA \times Aa$) unless they have produced at least one recessive (affected) child. Therefore, the observable families for study are those with at least one recessive child, while those without any recessive children are unidentifiable and missing from our recorded data.

Weinberg (1912) was the first to recognize and correct the bias in collecting sibships with at least one affected member. If the probability of producing a recessive offspring is 1/4, then the probability of having no recessives at all in a sibship of s members is $(3/4)^s$, which is the proportion of $Aa \times Aa$ families that is missing from the observations. Hence, Weinberg's method is based on the correction factor $1 - (3/4)^s$. Twenty years later, Haldane (1932) was concerned with estimating the probability (p) of producing a recessive without any a priori assumption, and, hence, the corresponding correction factor becomes $1 - q^s$, where $q = 1 - p$. He employed the maximum likelihood method of estimation and the resulting equation involving the hard-to-handle factors $1 - q^s$, where $s = 2, 3, 4, \ldots$, up to the largest family, is extremely difficult to solve. Biometricians subsequently have prepared tables to facilitate the arithmetic; high-speed computers have been used to obtain iterative solutions.

In an attempt to bypass the factor $1 - q^s$ and arrive at a simpler but somewhat less efficient estimate, Li (1964, 1965, 1966) has suggested subdividing the data into two independent parts: (1) children up to the first affected in a sibship and (2) children after the first affected. In the latter group, an estimate (p_0) may be obtained by simple counting with full efficiency. In the former group, a series of estimates ($p_1, p_2, p_3, \ldots$) may be obtained by taking certain ratios because the "first appearance time" has a truncated geometric distribution for sibships of any fixed size. These estimates are not fully efficient except for sibships of size 2. By combining the estimates p_1, p_2, etc., and p_0, it is found that the over-all loss of efficiency is small. A further discussion of certain aspects of this method will be found in the last section

Received April 18, 1967.

[1] Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.

[2] Biometry Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20014.

of Discussion. The main purpose of the present communication is to propose a new and even simpler method of estimation, which neither involves factors $1 - q^s$ nor requires the combination of several estimates. In fact, the new method amounts to little more than simple counting.

METHOD OF DISCARDING THE SINGLES

The basic argument of the method to be proposed has been given by Mantel (1951) in connection with the problem of evaluating the efficacy of diagnostic tests. Suppose that s tests have been performed on a person. These tests are assumed to be binomial trials, as in the case of genetic segregation. It is also assumed that these tests do not give false positive results, so that a single positive result among the s tests made is sufficient to indicate that the individual has infection, but a test may fail to detect the pathogen of an infected person, thus yielding a negative result. In estimating the probability of a test yielding a positive result on *infected* persons, Mantel's estimate takes the form $p' = A/B$, where A is the total number of positive test results obtained under conditions where the person was known to be infected independently of the particular test result, and B is the total number of tests made under such conditions. An individual is known to be infected independently of an examination result if there is external information that he is infected. Also such independent knowledge exists if at least one positive result has been obtained among the $s - 1$ remaining tests made on the individual, provided the conduct of subsequent tests in no way depends on the outcome of previous tests. Where such serial dependence for the conduct of tests may exist, independent knowledge would come from a positive result on an earlier examination only. Further elaboration on the rationale of the method will be given in the section on Discussion.

The consequences of the rule given above are: When the number of positive results of an individual is $r \geq 2$, this individual contributes r to numerator A and s to denominator B; when $r = 1$, he contributes 0 to numerator A and $s - 1$ to denominator B. Individuals with all negative results and individuals on whom only one test was performed, whether positive or negative in outcome, are automatically eliminated because they contribute 0 to both A and B.

Mantel's argument and method are general and applicable to a wide range of situations. However, in the text of this paper, we shall limit ourselves to genetic considerations. Before proceeding to generality, we shall illustrate the method with 175 sibships of fixed size $s = 4$, which have the binomial distribution $(q + p)^s = (3/4 + 1/4)^4$ without the first term. This is to assume "complete ascertainment" of the affected families in a community. Then, the "ideal" observed data will be as follows:

No. of Recessives in a Sibship	1	2	3	4	Totals	
					n	$t = sn$
No. of sibships	108	54	12	1	175	700
No. of recessives	108	108	36	4	$\Sigma r = 256$	

(1)

The $n = 175$ sibships, each of size $s = 4$, consist of a total of $t = ns = 175 \times 4 = 700$ children, of whom $\Sigma r = 256$ are affected. Since the missing families contribute no recessive children, the number $\Sigma r = 256$ would remain the same even if the "missing" families were included. In the latter case, however, the total number of children will no longer be 700, since they represent only a fraction, $1 - (3/4)^4$, of the theoretical total. Hence, the classical estimate of p is given by the solution of the equation:

$$\hat{p} = \Sigma r \Big/ \frac{t}{1 - q^s}, \text{ i.e., } \quad \frac{256}{p} = \frac{700}{1 - q^4}. \tag{2}$$

It will be found that $\hat{p} = 1/4$ is the solution. Now, our simplified estimate is based on a relationship exhibited in example (1) above. Note that $j = 108$ is the number of sibships with only one affected member each, and therefore it is also the number of

TABLE 1

BINOMIAL DISTRIBUTION WITH AND WITHOUT THE FIRST TERM, q^s = PROB(r=0), WHERE r IS THE NUMBER OF RECESSIVES IN A SIBSHIP OF SIZE s

No. of Recessives in a Sibship (r)	Frequency of Sibships (f)	No. of Recessives (fr)	Correspondence to Observed Numbers
0	q^s	0	
1	spq^{s-1}	spq^{s-1}	j singles
2	$\binom{s}{2}p^2q^{s-2}$	$s(s-1)p^2q^{s-2}$	
3	...	...	
.	.	.	
s	p^s	sp^s	
Complete total	1	sp	
Truncated total	$1-q^s$	sp	Σr recessives
Truncated total children		$s(1-q^s)$	t children

affected children contributed by such families. For brevity, we shall call $j = 108$ the number of "singles" (or "singletons"). Our proposed estimate is

$$p' = \frac{A}{B} = \frac{\Sigma r - j}{t - j}; \tag{3}$$

and, from our example (1) we obtain

$$p' = \frac{256 - 108}{700 - 108} = \frac{148}{592} = \frac{1}{4}.$$

Expressions (2) and (3) bring out the difference between the two methods of estimation. In a practical case, expression (2) involves the unknown p on both sides of the equation (as $q = 1 - p$) and the denominator is a polynomial of degree s. Expression (3) yields a proportion directly, there being no equation to solve.

Now we proceed to justify the proposed estimate $p' = (\Sigma r - j)/(t - j)$ algebraically for sibships of any fixed size s. For this purpose, it is simplest to refer to Table 1, which gives the complete binomial distribution and the totals for the truncated binomial. It is seen that the total number of children t is proportional to

$s(1 - q^s)$; the total number of recessives Σr is proportional to sp, and the number of "singles" j is proportional to spq^{s-1}. Hence, the expectation of (3) tends to equal

$$\frac{sp - spq^{s-1}}{s(1-q^s) - spq^{s-1}} = \frac{p(1-q^{s-1})}{1-q^{s-1}} = p. \tag{4}$$

Since equation (4) is true for all values of s, a pooled estimate of p from families of all sizes is

$$p' = \frac{A}{B} = \frac{R-J}{T-J}, \tag{5}$$

where R is the grand total of affected children from all families, T is the grand total number of children, and J is the grand total number of singles. Expression (5) is the formula given by Mantel (1951) for estimating the probability of yielding a positive result by diagnostic tests. The individuals with all tests negative are eliminated from the estimation because they are indistinguishable from individuals who have no infection in the first place.

APPLICATIONS TO DATA

In order to see how the present procedure of estimation fares with the fully efficient method, we shall apply (5) to two sets of data which have been analyzed previously by the maximum likelihood method. As the first example, the data on 27 segregating sibships are given in Table 2, in which a plus sign (+) indicates the presence and a minus sign (−) indicates the absence of the Ellis-van Creveld syndrome (McKusick *et al.*, 1964). The three grand totals are $T = 172$, $R = 48$, and $J = 14$. The estimate of the proportion of recessives is then

$$p' = \frac{R-J}{T-J} = \frac{48-14}{172-14} = \frac{34}{158} = 0.215. \tag{6}$$

Note that here we are treating sibship No. 23, which contained one child of unknown status due to stillbirth, as a reduced sibship of size 2. When Li (1965) treated this unknown as a negative, he obtained an interpolated maximum likelihood estimate of 0.215; the present formula then would yield 34/159 = 0.214.

It is seen from Table 2 that the $J = 14$ "singles" have been included in the totals T and R. If we had deleted these singles (leaving the remaining normal sibs of the families intact) from the data, we would have directly counted 158 children, of whom 34 are affected. Hence, the method amounts to simple counting after the deletion of the singles from the data; and we call this procedure "the method of discarding the singles" (or "discarding the singletons").

The data to be used for the second example are reproduced in Table 3, which was originally analyzed by Haldane (1938). The maximum likelihood estimate $\hat{p}$ requires the solution of the equation (pooled form of [2])

$$\frac{864}{p} = \frac{80}{1-q^2} + \frac{165}{1-q^3} + \ldots + \frac{56}{1-q^{14}} + \frac{15}{1-q^{15}}; \tag{7}$$

and Haldane found that $\hat{p} = 0.308$. Table 3 shows that the total number of singles among these 411 sibships is $J = 171$. Hence, our simplified estimate is

$$p' = \frac{864-171}{2{,}435-171} = \frac{693}{2{,}264} = 0.306. \tag{8}$$

TABLE 2

THE 27 SIBSHIPS WITH ELLIS-VAN CREVELD SYNDROME
(MCKUSICK *et al.*, 1964)

Family No.	Conditions of Children (+, Affected; −, Normal)	No. of Children	No. of Affected	No. of "Singles"
14	+−	2	1	1
23	?−+	2	1	1
1	+−+	3	2	.....
10	+−−	3	1	1
16	+++	3	3	.....
2	−−−+	4	1	1
5	−+−−	4	1	1
9	−−+−	4	1	1
15	++−−	4	2	.....
20	+−−−	4	1	1
21	−−+−	4	1	1
3	−++−−	6	3	.....
6	−−−−−+	6	1	1
8	−−+−−−	6	1	1
27	+−−−−−	6	1	1
7	−+−−−++	7	3	.....
17	−−−−+++	7	3	.....
13	−−−−+++−	8	3	.....
19	+−−−−−−−	8	1	1
29	+−+−−−−−	8	2	.....
30	−−+−−−−−	8	1	1
4	−−−+−−+−	9	3	.....
24	−−−+−+++−	9	4	.....
28	−−−+−−−−−+	10	2	.....
25	−−+−−−−−+−−	11	2	.....
12	−−+−−+−−−−−−	12	2	.....
11	−−−−−−−−+−−−−−	14	1	1
Total...	.....	$T=172$	$R=48$	$J=14$

NOTE.—Pedigrees 18 and 26 are omitted because each consists of a single affected child. Pedigree 22 is omitted because the father is affected.

TABLE 3

DISTRIBUTION OF SIBSHIPS BY SIBSHIP SIZE (s) AND THE NUMBER OF RECESSIVES (HUMAN ALBINISM) IN A SIBSHIP; DATA OF PEARSON *et al.* (1913), AS CITED AND ANALYZED BY HALDANE (1938)

Size (s)	Number of Recessives in a Sibship										Number of		
	1	2	3	4	5	6	7	8	9	10	Sibships	Recessives	Children
2	31	9	.....	.....	.....	.....	.....	.....	.....	.....	40	49	80
3	37	15	3	.....	.....	.....	.....	.....	.....	.....	55	76	165
4	22	21	7	0	.....	.....	.....	.....	.....	.....	50	85	200
5	25	23	10	1	1	.....	.....	.....	.....	.....	60	110	300
6	18	13	18	3	0	1	.....	.....	.....	.....	53	116	318
7	16	10	14	5	1	0	0	.....	.....	.....	46	103	322
8	4	8	7	6	1	0	1	0	.....	.....	27	77	216
9	10	4	9	4	1	0	1	0	0	.....	29	73	261
10	6	3	7	2	1	1	0	0	0	0	20	52	200
11	0	2	4	6	2	0	0	0	0	0	14	50	154
12	2	0	0	4	2	0	0	0	0	0	8	28	96
13	0	0	1	1	1	0	1	0	0	0	4	19	52
14	0	1	1	1	0	0	1	0	0	0	4	16	56
15	0	0	0	0	0	0	0	0	0	1	1	10	15
Total...	171 (J)	109	81	33	10	2	4	0	0	1	411 (N)	864 (R)	2,435 (T)

Again, it is very close to but slightly smaller than the maximum likelihood estimate. The deviation from the theoretical value 0.250 may be due to a variety of reasons. One obvious reason is that the assumption of complete ascertainment is not fulfilled; in other words, families with a small number of albinos have been underrepresented in the sample. For sibships of size $s = 15$, the probability of having three affected members is more than 330 times that of having 10 affected. And yet, one family with 10 albinos has been recorded and no family with a small number of albinos has been encountered. Another reason may be that one parent of some of the observed families was actually an albino but was incorrectly recorded as a normal due to incomplete information. Further discussion may be found in a later section on "Truncation at Any Point."

TABLE 4

DISTRIBUTION, VARIANCE, AND COVARIANCE OF $a = r - j$ AND $b = s - j$ FOR A SIBSHIP OF SIZE s

No. Recessive, r No. Single, j	1 1	2 0	··· ···	s 0	Total or Mean
Probability, f	$\frac{spq^{s-1}}{1-q^s}$	$\frac{\binom{s}{2}p^2q^{s-2}}{1-q^s}$	···	$\frac{p^s}{1-q^s}$	1 00
$a=r-j$ $b=s-j$ $a\cdot b$	0 $s-1$ 0	2 s $2s$	··· ··· ···	s s s^2	$\Sigma fa = E(a)$ $\Sigma fb = E(b)$ $\Sigma fab = E(ab)$

For brevity we write $f(1) = \text{Prob}(r=1) = spq^{s-1}/(1-q^s)$.

$$E(a) = \frac{sp}{1-q^s} - f(1), \qquad V(a) = \frac{spq+s^2p^2}{1-q^s} - f(1) - E^2(a).$$

$$E(b) = s - f(1) = E(a)/p, \qquad V(b) = f(1)[1-f(1)].$$

$$\text{Cov}(a,b) = sE(a) - E(a)E(b) = E(a)[s-E(b)] = E(a)\cdot f(1).$$

VARIANCE FOR SIBSHIPS OF THE SAME SIZE

The variance of the estimate (5) as originally given by Mantel (1951) is inapplicable for this particular instance. For an estimate based on the ratio of two variables, $p' = A/B$, its asymptotic variance for large samples is

$$V(p') \cong p^2\left\{\frac{V(A)}{E^2(A)} + \frac{V(B)}{E^2(B)} - \frac{2\,\text{cov}(A,B)}{E(A)E(B)}\right\}. \tag{9}$$

We shall first find the required expectations, variances, and covariance per sibship ($n = 1$) of size s. Although (9) is an asymptotic expression and is true only for large n, we shall for convenience use one sibship as the basic unit of data. The lowercase symbols a, b, j, r are employed for one sibship, but $t = ns = s$ for $n = 1$. The distribution of the relevant variables is shown in the upper half of Table 4, from which the various expected values, variances, and covariance are found and listed in the lower half of that table. Substituting these values in (9), we obtain the asymptotic variance

for one sibship (but recognizing that the variance formula applies only for large n):

$$V(p') \cong \frac{pq}{s} \times \frac{(1-q^s)[1-q^s+(s-2)pq^{s-1}]}{(1-q^{s-1})^2}. \tag{10}$$

Note that the factor, pq/s, is the variance for a complete binomial distribution. Hence, the second factor, which can readily be shown to exceed unity, may be regarded as the correction for truncation and discarding of singles.

The "amount of information" concerning p contributed by one sibship, or the "weight" to be attached to the estimate from one sibship, is the reciprocal of the variance (10). These weights, $w = 1/V$, have been tabulated for various values of p and s in Table 5 (for genetical use) and Table 6 (for general use). The estimate based on n sibships of the same size will have a variance V/n, or a weight nw. For maximum likelihood estimation, the asymptotic unit sibship variance takes the form:

$$V(\hat{p}) \cong \frac{pq}{s} \times \frac{(1-q^s)^2}{1-q^s-spq^{s-1}}. \tag{11}$$

Here, the correction factor on the binomial variance can also be shown to exceed unity. The reciprocal of (11) has been tabulated by various authors (e.g., Li, 1961, p. 66), and the efficiency of our present estimate p' relative to maximum likelihood may be obtained easily for sibships of any given size. The comparisons within the limits of genetic interest are given in Table 7.

It is to be observed that for sibships of size $s = 2$, the variances (10) and (11) are identical, so the present method is fully efficient for all values of p. The estimates given by the two methods are also identical, as the maximum likelihood equation

$$\frac{r}{p} = \frac{t}{1-q^2}$$

yields the solution

$$\hat{p} = \frac{2r-t}{r} = \frac{r-j}{t-j},$$

since $t = r + j$ for $s = 2$. From Table 7 we see that the efficiency gradually decreases with s to a certain point and then gradually increases again. For $p = .25$, the minimum efficiency is 96.8% at $s = 6$. Further calculation shows that the efficiency is similarly high for all other values of p and s.

VARIANCE OF POOLED ESTIMATE

The expression (5) represents a pooled estimate from sibships of all sizes and may be rewritten as

$$p' = \frac{A}{B} = \frac{\Sigma a}{\Sigma b}, \tag{12}$$

where the a's and b's are the numerator and denominator contributions by single sibships. The variance (9) of the pooled estimate can be written alternatively as

$$V(p') \cong \frac{1}{E^2(B)} \{V(A) + p^2 V(B) - 2p \operatorname{cov}(A, B)\},$$

TABLE 5

VALUES OF $w = \frac{s}{pq} \cdot \frac{(1-q^{s-1})^2}{(1-q^s)[1-q^s+(s-2)pq^{s-1}]}$

s	$p = 16$	$p = .17$	$p = .18$	$p = .19$	$p = .20$	$p = .21$	$p = .22$	$p = .23$	$p = .24$	$p = .25$
2	4.40	4.23	4.09	3.97	3.86	3.76	3.68	3.60	3.54	3.48
3	9.13	8.81	8.54	8.29	8.08	7.90	7.74	7.60	7.48	7.37
4	14.21	13.74	13.34	12.99	12.68	12.41	12.18	11.98	11.81	11.66
5	19.64	19.03	18.50	18.04	17.65	17.30	17.01	16.75	16.53	16.34
6	25.42	24.67	24.02	23.46	22.97	22.55	22.19	21.88	21.61	21.38
7	31.54	30.64	29.88	29.21	28.64	28.14	27.71	27.34	27.01	26.73
8	37.98	36.95	36.06	35.29	34.62	34.04	33.53	33.09	32.70	32.36
9	44.74	43.56	42.54	41.66	40.88	40.21	39.61	39.08	38.62	38.20
10	51.79	50.46	49.30	48.29	47.40	46.61	45.90	45.28	44.71	44.20
11	59.11	57.61	56.30	55.14	54.12	53.20	52.37	51.62	50.93	50.30
12	66.68	65.00	63.51	62.19	61.00	59.93	58.95	58.06	57.23	56.46
13	74.47	72.58	70.90	69.39	68.02	66.77	65.62	64.55	63.56	62.63
14	82.44	80.32	78.42	76.70	75.13	73.68	72.33	71.07	69.89	68.78
15	90.58	88.20	86.06	84.10	82.29	80.62	79.05	77.58	76.20	74.88
16	98.84	96.19	93.77	91.55	89.49	87.56	85.76	84.06	82.45	80.93
17	107.21	104.24	101.53	99.02	96.68	94.49	92.43	90.49	88.65	86.92
18	115.65	112.35	109.31	106.49	103.85	101.38	99.05	96.85	94.78	92.83
19	124.15	120.48	117.09	113.94	110.99	108.22	105.61	103.16	100.85	98.67
20	132.67	128.62	124.86	121.35	118.07	115.00	112.11	109.39	106.84	104.45

s	$p = .25$	$p = .26$	$p = .27$	$p = .28$	$p = .29$	$p = .30$	$p = .31$	$p = .32$	$p = 33$	$p = .34$
2	3.48	3.43	3.39	3.35	3.32	3.30	3.27	3.26	3.24	3.23
3	7.37	7.28	7.20	7.13	7.08	7.03	7.00	6.97	6.95	6.95
4	11.66	11.53	11.42	11.34	11.26	11.20	11.16	11.13	11.11	11.10
5	16.34	16.18	16.04	15.93	15.84	15.77	15.72	15.68	15.66	15.65
6	21.38	21.18	21.01	20.88	20.76	20.67	20.60	20.55	20.52	20.50
7	26.73	26.49	26.29	26.12	25.97	25.85	25.75	25.67	25.61	25.57
8	32.36	32.06	31.81	31.58	31.38	31.21	31.07	30.94	30.83	30.74
9	38.20	37 83	37.50	37.20	36.94	36.70	36.48	36.29	36.11	35.95
10	44.20	43.73	43.31	42.92	42.56	42.23	41.92	41.64	41.38	41.13
11	50.30	49.72	49.18	48.67	48.20	47.76	47.35	46.96	46.59	46.25
12	56.46	55.73	55.06	54.42	53.82	53.25	52.72	52.21	51.73	51.28
13	62.63	61.75	60.92	60.13	59.39	58.68	58.01	57.38	56.78	56.22
14	68.78	67.72	66.72	65.78	64.88	64.03	63.22	62.46	61.74	61.07
15	74.88	73.64	72.47	71.35	70.29	69.30	68.35	67.46	66.63	65.84
16	80.93	79.50	78.14	76.85	75.63	74.48	73.40	72.39	71.43	70.55
17	86.92	85.28	83.73	82.27	80.89	79.60	78.38	77.24	76.18	75.19
18	92.83	90.99	89.25	87.62	86.09	84.65	83.30	82.05	80.88	79.80
19	98.67	96.63	94.71	92.91	91.22	89.64	88.17	86.81	85.54	84.36
20	104.45	102.20	100.10	98.13	96.30	94.59	93.00	91.53	90.16	88.90

TABLE 6

VALUES OF $w = \frac{s}{pq} \cdot \frac{(1-q^{s-1})^2}{(1-q^s)[1-q^s+(s-2)pq^{s-1}]}$

s	$p=.05$	$p=.10$	$p=.15$	$p=.20$	$p=.25$	$p=.30$	$p=.35$	$p=.40$	$p=.45$	$p=.50$
2	11.07	6.16	4.58	3.86	3.48	3.30	3.23	3.26	3.36	3.56
3	22.42	12.61	9.50	8.08	7.37	7.03	6.94	7.04	7.29	7.71
4	34.05	19.38	14.76	12.68	11.66	11.20	11.11	11.28	11.68	12.30
5	45.95	26.46	20.36	17.65	16.34	15.77	15.66	15.88	16.36	17.08
6	58.15	33.86	26.30	22.97	21.38	20.67	20.51	20.69	21.16	21.86
7	70.62	41.57	32.58	28.64	26.73	25.85	25.54	25.60	25.93	26.52
8	83.39	49.60	39.18	34.62	32.36	31.21	30.67	30.49	30.61	31.02
9	96.44	57.94	46.11	40.88	38.20	36.70	35.81	35.31	35.16	35.37
10	109.78	66.60	53.33	47.40	44.20	42.23	40.91	40.03	39.58	39.61
11	123.42	75.56	60.84	54.12	50.30	47.76	45.93	44.64	43.89	43.76
12	137.35	84.83	68.61	61.00	56.46	53.25	50.86	49.14	48.12	47.86
13	151.57	94.39	76.62	68.02	62.63	58.68	55.69	53.56	52.29	51.92
14	166.09	104.24	84.83	75.13	68.78	64.03	60.44	57.91	56.42	55.95
15	180.90	114.37	93.23	82.29	74.88	69.30	65.11	62.21	60.51	59.97
16	196.01	124.76	101.79	89.49	80.93	74.48	69.72	66.47	64.59	63.98
17	211.41	135.41	110.48	96.68	86.92	79.60	74.28	70.70	68.65	67.99
18	227.09	146.30	119.27	103.85	92.83	84.65	78.79	74.91	72.70	71.99
19	243.08	157.41	128.15	110.99	98.67	89.64	83.28	79.11	76.75	76.00
20	259.35	168.73	137.08	118.07	104.45	94.59	87.74	83.29	80.80	80.00

s	$p=.50$	$p=.55$	$p=.60$	$p=.65$	$p=.70$	$p=.75$	$p=.80$	$p=.85$	$p=.90$	$p=95$
2	3.56	3.84	4.25	4.82	5.64	6.83	8.68	11.86	18.37	38.19
3	7.71	8.31	9.13	10.23	11.74	13.85	17.01	22.21	32.44	62.71
4	12.30	13.14	14.26	15.71	17.66	20.36	24.37	31.01	44.28	84.17
5	17.08	18.04	19.29	20.93	23.14	26.28	31.05	39.13	55.53	105.26
6	21.86	22.81	24.10	25.84	28.28	31.86	37.44	47.04	66.66	126.32
7	26.52	27.40	28.67	30.51	33.21	37.29	43.73	54.90	77.78	147.37
8	31.02	31.80	33.07	35.05	38.05	42.65	50.00	62.74	88.89	168.42
9	35.37	36.06	37.37	39.51	42.84	48.00	56.25	70.59	100.00	189.47
10	39.61	40.24	41.60	43.93	47.61	53.33	62.50	78.43	111.11	210.53
11	43.76	44.35	45.80	48.34	52.38	58.67	68.75	86.27	122.22	231.58
12	47.86	48.44	49.98	52.74	57.14	64.00	75.00	94.12	133.33	252.63
13	51.92	52.50	54.16	57.14	61.90	69.33	81.25	101.96	144.44	273.68
14	55.95	56.55	58.33	61.54	66.67	74.67	87.50	109.80	155.56	294.74
15	59.97	60.60	62.50	65.93	71.43	80.00	93.75	117.65	166.67	315.79
16	63.98	64.64	66.67	70.33	76.19	85.33	100.00	125.49	177.78	336.84
17	67.99	68.68	70.83	74.73	80.95	90.67	106.25	133.33	188.89	357.89
18	71.99	72.73	75.00	79.12	85.71	96.00	112.50	141.18	200.00	378.95
19	76.00	76.77	79.17	83.52	90.48	101.33	118.75	149.02	211.11	400.00
20	80.00	80.81	83.33	87.91	95.24	106.67	125.00	156.86	222.22	421.05

TABLE 7

PERCENTAGE OF RELATIVE EFFICIENCY OF THE PRESENT METHOD OF ESTIMATION TO THAT OF MAXIMUM LIKELIHOOD

s	2	3	4	5	6	7	8	10	12	15
$p=.20$	100.0	98.7	97.8	97.2	96.8	96.6	96.5	96.8	97.3	98.1
$p=.25$	100.0	98.5	97.6	97.1	96.8	96.9	97.0	97.6	98.3	99.1
$p=.30$	100.0	98.4	97.5	97.2	97.2	97.4	97.7	98.4	99.1	99.6

which becomes

$$V(p') \cong \frac{1}{[\Sigma E(b)]^2} \{\Sigma V(a) + p^2 \Sigma V(b) - 2p \Sigma \operatorname{cov}(a, b)\}, \quad (12V)$$

where the summation is over all the sibships of all sizes. The variance as given by (12V) is cumbersome, since for sibships of each size we have to obtain the quantities: $E(b)$, $V(b)$, $E(a)$, $V(a)$, cov (a, b). However, an approximate method may be used, which underestimates the variance only very slightly. Justification of the approximate method will be given in more detail in the Discussion section. In the text, we shall merely give the method and numerical illustrations.

Tables 5 and 6 give the values of w, the reciprocal of (10), and this may be taken as the "information" (relative to p) contributed by one sibship of size s. If there are n_s sibships of size s, the total information will be $n_s w_s$. The grand total information from sibships of all sizes and the approximate variance are then, respectively,

$$W = \Sigma n_s w_s, \qquad V(p') \cong \frac{1}{W}. \quad (13)$$

In order to see the difference in computation and numerical value between variance (12V) and the approximate expression (13), let us suppose that we observed three sibships as follows:

$$s = 4(- - + -), \qquad s = 5(+ - - - +),$$

$$s = 12(- - - + - - - - + - + -).$$

The combined estimate from the first two sibships (s = 4, 5), which do not differ greatly in size, is $p' = (3 - 1)/(9 - 1) = 2/8 = 0.25$. Combining all three sibships (s = 4, 5, 12), which differ widely in size, yields the same contrived estimate, $p' = (6 - 1)/(21 - 1) = 5/20 = 0.25$. The arithmetic involved in calculating the variance (12V) for the two pooled estimates is shown in Table 8, in which only four places of decimals are retained, although the original calculation involves six places. In using the approximate expression (13), we merely refer to Table 5 and obtain the following weights for $p = 0.25$.

s	w	Total Weight	Approximate Variance (13)	Variance (12V)
4	11.66	$W = 28.00$	$V(p') = 0.03571$	$V(p') = 0.03574$
5	16.34			
12	56.46	$W = 84.46$	$V(p') = 0.01184$	$V(p') = 0.01208$

Comparing these approximate variances with those obtained from Table 8, we see that, when sibships not differing greatly in size are combined, the difference between the two results is negligible (.03571 versus .03574). Even in the extreme case in which sibships varying from very small to very large in size are combined, the difference is still below 2% (.01184 versus .01208). In any practical situation, the approximate variance (13) will be very close to (12V).

Now, we shall apply the method (13) to the data of Table 3, for which our estimate

is $p' = .306$. The values of n_s are 40, 55, 50, etc. Table 5 lists the weights $w_s = 3.30$, 7.03, 11.20, . . . for $p = .30$ and $w_s = 3.27$, 7.00, 11.16, . . . for $p = .31$. The grand total weights are:

For $p = 0.30$, $W = \Sigma n_s w_s = 8{,}715.8$

For $p = 0.31$, $W = \Sigma n_s w_s = 8{,}666.4$

For $p = 0.306$, $W = 8{,}686.2$ by interpolation,

and the standard error of the estimate is $1/\sqrt{8{,}686.20} = 1/93.2 = .0107$, which is the value obtained by Haldane (1938) for his estimate, $\hat{p} = .3082$. Thus, we see that both the estimate and its variance are very close to those yielded by the maximum likelihood method.

TABLE 8

EXPECTED VALUES, VARIANCES, AND COVARIANCES OF a AND b AT $p = 0.25$
EXPRESSIONS FOR $f(1)$, $E(a)$, ETC., HAVE BEEN GIVEN IN TABLE 5

s	$f(1)$	$E(b)$	$V(b)$	$E(a)$	$V(a)$	Cov (a, b)
4	.6171	3.3829	.2363	.8457	1.2276	.5219
5	.5186	4.4814	.2496	1.1204	1.5041	.5810
Total	.....	7.8643	.4859	1.9661	2.7317	1.1029
12	.1308	11.8691	.1137	2.9673	2.6824	.3883
Grand total	.....	19.7334	.5997	4.9334	5.4141	1.4912

For the combined estimate from the first two sibships:

$$V(p') = \frac{2.7317 + .0625(.4859) - .50(1.1029)}{(7.8643)^2} = 0.03574.$$

For the combined estimate from all three sibships:

$$V(p') = \frac{5.4141 + .0625(.5997) - .50(1.4912)}{(19.7334)^2} = 0.01208.$$

TRUNCATION AT ANY POINT

The principle and procedure underlying the method of discarding the singles may be extended to truncation of a binomial distribution at any point. To illustrate, let us consider the complete binomial distribution $(q + p)^8 = (3/4 + 1/4)^5$ for 1,024 sibships, as given in the top portion of Table 9. The successive portions in that table show the successive stages of truncation. For instance, when the first three terms are deleted from the distribution, the estimate of p will be given by

$$p' = \frac{335 - 270}{530 - 270} = \frac{65}{260} = \frac{1}{4},$$

where 270 is the number of affected children who constitute "triples," and so on. Further comments on this property will be found in the Discussion.

An immediate application of the truncation property described above is that a series of estimates of the segregation ratio may be obtained at various points of truncation from the comparatively large families. Under true complete ascertainment, these successive estimates should be approximately the same. In discussing the data of Table 3, which yields $p' = .306$, we mentioned the possibility that families with a small number of recessives may be underrepresented in the sample. If so, the omission of sibships with one or two recessives from the data should yield an estimate lower than .306. The data thus truncated are shown in Table 10, in which sibships of

TABLE 9

SUCCESSIVE TRUNCATIONS OF A BINOMIAL DISTRIBUTION WITH $q = 3/4$, $p = 1/4$, AND $s = 5$, AND THE CORRESPONDING ESTIMATE OF p

r	0	1	2	3	4	5	n t	Estimate
f	243	405	270	90	15	1	1,024 5,120	$p' = \frac{1,280-0}{5,120-0} = \frac{1}{4}$
fr	0	405	540	270	60	5	$r = 1,280$	
f	.....	405	270	90	15	1	781 3,905	$p' = \frac{1,280-405}{3,905-405} = \frac{1}{4}$
fr	.....	405	540	270	60	5	$r = 1,280$	
f	.....	.....	270	90	15	1	376 1,880	$p' = \frac{875-540}{1,880-540} = \frac{1}{4}$
fr	.....	.....	540	270	60	5	$r = 875$	
f	.....	.....	.....	90	15	1	106 530	$p' = \frac{335-270}{530-270} = \frac{1}{4}$
fr	.....	.....	.....	270	60	5	$r = 335$	
f	.....	.....	.....	.....	15	1	16 80	$p' = \frac{65-60}{80-60} = \frac{1}{4}$
fr	.....	.....	.....	.....	60	5	$r = 65$	

TABLE 10

DATA OF TABLE 3 WITH THE FIRST TWO COLUMNS DELETED (SIBSHIPS OF SIZE 2 AND 3 ARE ALSO OMITTED BECAUSE THEY PLAY NO PART IN ESTIMATING p. THE SINGLE SIBSHIP OF SIZE 15 WITH 10 RECESSIVES IS OMITTED TO SHORTEN THE TABLE)

Size s	No. Recessives in a Sibship					n	r	t
	3	4	5	6	7			
4	7	0	.....	.....	.....	7	21	28
5	10	1	1	.....	.....	12	39	60
6	18	3	0	1	.....	22	72	132
7	14	5	1	0	0	20	67	140
8	7	6	1	0	1	15	57	120
9	9	4	1	0	1	15	55	135
10	7	2	1	1	0	11	40	110
11	4	6	2	0	0	12	46	132
12	0	4	2	0	0	6	26	72
13	1	1	1	0	1	4	19	52
14	1	1	0	0	1	3	14	42
Total	78 ($J = 234$)	33	10	2	4	127 (N)	456 (R)	1,023 (T)

size 2 and 3 have been omitted as they contribute nothing to the estimate. To shorten the table further, the single sibship of size 15 with 10 recessives has also been omitted. From Table 10, we obtain the new estimate

$$p' = \frac{R-J}{T-J} = \frac{456-234}{1{,}023-234} = \frac{222}{789} = .281,$$

which is somewhat lower than .306. Similarly, if we further omit the first column of Table 10 (and thus its first row), we would obtain the estimate

$$p' = \frac{222-132}{461-132} = \frac{90}{329} = .274.$$

These two estimates will become .291 and .289, respectively, if the sibship of size 15 were included. While the evidence for incomplete ascertainment is not strong, the possibility for using the $p' = A/B$ estimator in this way is illustrated.

DISCUSSION

There are several aspects of the estimation problem that we would like to discuss more in detail. This may best be done separately in the following subsections.

1. *The Rationale of the A/B Estimates*

The A/B estimation principle was first proposed by Mantel (1951). It is similar in form to the usual binomial estimator, r/n, the ratio of successes to the number of trials; the necessary change is that the denominator is the number of *effective* trials, while the numerator is the number of successes under conditions of effective trial. The modification arises from the circumstances that without external information we cannot in certain problems clearly interpret a negative result. In diagnostic testing, a negative result may indicate that either the person is uninfected or that, although infected, by chance the examination was negative. In the problems considered in this paper a similar duality of possibilities occurs when a birth is of an unaffected child.

For purposes of the A/B estimator, a trial is considered to be effective if information were available independent of the results of the particular trial that this was a true trial, that is, that the person was infected or that the parents had the proper genetic structure. Such information can arise in a number of ways. For example, we might already know for some individuals that they were infected or that for some couples both husband and wife were heterozygous.

Suppose we had no such external information. Consider that we conducted a fixed number of tests, s, on each of various individuals. We assume, which may be unrealistic in some situations, that each of the unknown infected individuals has the same probability, p, that a particular test will be positive, the various test results being independent, while for uninfected individuals $p = 0$. In this circumstance, there is an effective trial in a particular test if on any other test the result was positive. Where an individual shows no positive results, he consequently has no effective tests, and both his A and his B contributions are 0. Once an individual has more than one positive result, all his tests are effective; his A contribution is his number of positive results, his B contribution s.

Cumulating all the A and B contributions, the estimator would turn out to be

"total number of positive tests minus number of individuals with only one positive" divided by "total number of tests in persons with any positive results minus number of individuals with only one positive." It is this form which leads to the descriptive term in the text above, "method of casting out the singles." In fact, the verbal form of the estimator remains the same even when s is allowed to vary, provided s is in no way influenced by how any particular test results. The formula should, of course, be modified for instances where some known infected individuals were included in the testing program. For such individuals, the B contribution would be all the tests conducted on him, while the A contribution would be the total number of positive results, whether zero, one, or more than one.

Mantel (1951) has considered other ways in which an effective trial may exist. Suppose that, at each occasion of testing for amebiasis, a person's stool was examined by a battery of procedures. For any occasion of testing to be effective for a given procedure, it will be necessary that, on at least one other occasion, a positive result was obtained, whether by the specified procedure or by any other. Curiously, this approach permits evaluating each separate test of the battery, as well as the battery as a whole. A more formal evaluation would be difficult in the absence of a mathematical law interrelating the results of the tests by the different procedures.

Use of the method of "casting out the singles" could be inappropriate under certain circumstances, circumstances under which the conduct of subsequent tests was related in some way to the results of past tests. Reasons for the conduct of additional tests may vary. An investigator may disbelieve a negative result and so continue testing until he gets a positive; in other circumstances, a positive result may lead him to test further just to see if it continues. This is paralleled in the genetics problem. Normal offspring may encourage some families to have more children, or put a stop to family formation in others. In other circumstances, the appearance of an affected child may put a stop to family formation, while leading other families to have just one more child or just one more normal child.

Suppose that, in ignorance of the circumstances, we used the method of "casting out the singles," concentrating on the effect in situations where family formation was stopped by the appearance of the first affected child. Such a family would have an A contribution of 0, and a B contribution of $s - 1$; thus, if our data consisted only of such families, and our probability model did not take this into account, our estimate of p would incorrectly be zero. But, we note that the maximum likelihood estimate would also incorrectly be zero.

Fortunately, the A/B principle permits us to handle this difficulty. It is not necessary to know the individual family rules for controlling family size. If we need concern ourselves with the possibility that a particular test outcome may have determined in some unknown way whether subsequent tests should be performed or that the status of a particular child may have determined whether or not additional children will be born, we need only alter what we conceive of as an effective trial. Under the A/B principle, a particular test or birth is an effective trial only if some preceding test or birth was positive. The A contribution for each sequence becomes the number of positives after the first positive in each sequence, and the B contribution is the num-

ber of births or trials after the first positive. We may note that Li (1964, 1965, 1966) suggested just such an estimator:

$$p_0 = \frac{\text{Total no. of recessives after the first recessive}}{\text{Total no. of children after the first recessive}} = \frac{R}{G},$$

although he did not have the problem of family limitation in mind.

In the case of invariable family limitation at the first positive, this would take the indefinite form 0/0. This may not be of much help, but we are not misled into thinking that p is zero; instead we may take this as an indication for making an estimate appropriate for this kind of limitation.

It may be noted that the A/B estimator is not invariably an unbiased estimator of p—whether it is depends on the kind of situation in which it is employed. What is generally true, however, is that the ratio of expected A to expected B (not the expected ratio) is p. In the genetic and diagnostic test problems considered here, neither the maximum likelihood estimator nor the "casting out of singles" estimator is unbiased, but, with sufficient data, biases can disappear for both estimators.

2. *Justification for Variance Approximation*

In introducing the A/B estimator, Mantel (1951) suggests as an approximation to its variance $p(1-p)/B$. But only under particular circumstances would this be correct—for instance, when we are dealing with only known infected cases or known heterozygous couples. Also the $p(1-p)/B$ formula is appropriate for setting limits based on the proportion of successes after the first success, that is, Li's $p_0 = R/G$ (1965). It is not appropriate for the "casting out of singles" estimator we have suggested in the present genetic or diagnostic test problems.

A more general approach would be to consider A/B a ratio estimator with both A and B subject to chance variation. Using this approach, we can readily establish from Table 4 the variance formula (10). This, together with (11), yields the asymptotic relative efficiency (A.R.E.) of our proposed estimator:

$$\text{A.R.E.} = \frac{V(\hat{p})}{V(\hat{p}')} = \frac{(1-q^{s-1})^2(1-q^s)}{(1-q^s-spq^{s-1})[1-q^s+(s-2)pq^{s-1}]}.$$

For $s = 2$, the A.R.E. is unity for all q, and for no combination of s and q is it materially less than unity (see Table 7).

Two matters must yet be considered. First, does the high efficiency for $\hat{p}'$ when all sibships are of the same size imply a similar high efficiency when sibships are of varying size? Second, what happens to the variance formula when sibships are of varying sizes?

To the second question, the proper answer would be that one need only cumulate the expectations, variances, and covariances over all the sibship sizes, taking into account the frequency of each, and substitute these into the general formula (12V) given for the asymptotic variance of a ratio. Rather than provide tables which will facilitate doing this, we are leaving this in a formula stage only for those who should wish to employ it. Instead, we have provided a simpler procedure (13), together with necessary tables, which yields a slight underestimate of the variance of $\hat{p}'$, as will now be justified.

To begin, we must assert that a procedure which is highly efficient for fixed sibship size need not be so when sibship size varies. It could be that the procedure weights the results for varying sibship sizes in a way grossly at odds with the information each size provides about the parameter being estimated. But, if the procedure weights strictly according to information, then the efficiency of the procedure when applied to mixed sibship sizes will be a weighted average of the separate sibship size efficiencies—and if all the separate efficiencies are high, so too will be the average. And the closer the weighting system is to information weighting, the closer will the efficiency be to average efficiency.

We could get a combined estimate across sibship size using information weighting, but this would be at the cost of losing the simplicity of the $p' = A/B$ estimator. But, we may next inquire, how close to information weighting does the A/B estimator come anyway?

We may note that the combined A/B estimator may be viewed as weighting each separate sibship estimate (a/b) by b. The reciprocal of the variance formula (10) yields the information w_s in the A/B value for a single sibship of size s. But the expected value of b for a given s, $E(b \mid s)$, is effectively the weighting system employed. The ratio of $E(b \mid s)/w_s$ is given by

$$pq\left[\frac{1 - q^s + (s-2)pq^{s-1}}{1 - q^{s-1}}\right].$$

To the extent that this ratio remains constant with s, the A/B estimator weights by information. In the above in any particular circumstance, the factor outside the brackets, pq, is a constant. The following table shows how the bracketed expression varies with s for selected values of p.

s	$p=0$	$p=.20$	$p=.25$	$p=.75$	$p=1$
2.....	2	1.80	1.75	1.250	1
3.....	2	1.71	1.64	1.100	1
4.....	2	1.63	1.55	1.036	1
5.....	2	1.56	1.46	1.012	1
6.....	2	1.49	1.39	1.004	1
7.....	2	1.43	1.32	1.001	1
∞.....	1–2[a]	1	1	1	1

[a] Limit 1 for $p > 0$, $s = \infty$; limit 2 for s finite, $p = 0$.

Where p is large (near unity), the bracketed expression is perforce in a sharply restricted region between 1 and $2 - p$, so that the weighting by the A/B formula is effectively optimal. For small p, in which the 1 to $2 - p$ range is widened, the transition downward in the bracketed expression from a value of $2 - p$ at $s = 2$ is gradual so that for limited s values, weighting is still close to optimal. For the case of two sibships of respective sizes 4 and 5, with $p = 0.25$, the two effective weightings show a 6% range from 1.46 to 1.55, instead of being constant, with the smaller sibship getting 6% more weight than it should relative to the larger sibship. But this results in a loss of only 1/12 of 1% in efficiency, the variance of the combined A/B estimator being 0.03574, while the variance of the optimally weighted A/B estimator would be

0.03571. For the three sibships of sizes 4, 5, and 12, the corresponding loss in efficiency is almost 2%.

It is clear that use of the combined A/B estimator arising in the "casting out of singles" procedure never gives any sibship size more than twice the weight it should have relative to any other sibship size. This circumstance can be approached only when the data contain sibships of size 2 together with sibships of extremely large size, with p extremely small. Consideration of this circumstance permits us to put a limit on the loss of efficiency by the combined estimator.

Suppose that Y_1 and Y_2 are two estimates which should get optimal weights W_1 and W_2, but that instead we use $2W_1$ and W_2. The efficiency of this nonoptimal weighting procedure is readily determined as

$$1 - \frac{W_1 W_2}{4W_1^2 + 5W_1 W_2 + W_2^2}.$$

If either of W_1 or W_2 is extremely large relative to the other, there is negligible loss in efficiency. Simply put, if one observation should get virtually no weight, doubling or halving a negligible weight makes little difference. The greatest loss in efficiency arises when $W_1 = W_2$, and then it is 10%.

For the combined A/B estimator, this 10% loss would not arise if, with p near zero, we had one sibship of size 2 and the other sibship of extremely large size, for then W_1 and W_2 would be grossly unequal. What would be required is that for each sibship of very large size there should be a large number of sibships of size 2 containing the same total amount of information.

We may take this to indicate that the efficiency loss of the combined A/B estimator is limited to 10% and can achieve such loss only under extreme hypothetical circumstances. In practical situations, the loss is likely to be much more limited and is unlikely to exceed 1%–2%.

We can now answer the two questions posed previously in this discussion.

First, use of the proposed combined A/B estimator loses only limited information relative to the optimally weighted A/B estimator. But the efficiency of the optimally weighted A/B estimator is a weighted average of the fixed sibship size efficiencies. Consequently, the efficiency of the proposed estimator is high. If, for example, the weighted average is 98% efficient, and an additional 1% information is lost by not fully efficient weighting, the estimator would still be 97% efficient.

Second, we may propose using the variance of the optimally weighted estimator as a near approximation to the variance of the proposed estimator. We may on occasions then underestimate the variance by about 2%, or the standard error by about 1%, but this would be satisfactory for most purposes. But this device greatly simplifies our computations and the necessary associated tabulations. The correct determination of $V(p')$ would have required tabulating $E(a)$, $E(b)$, var (a), var (b), and cov (a, b) for each combination of p and s, so that by cumulating these over all the sibships in our data, we would obtain $E(A)$, $E(B)$, var (A), var (B), and cov (A, B) for substituting in (12V). Instead, now we need tabulate only w_s for each p, s combination. Cumulating these across all sibships gives the total information in the sample

about p by the optimally weighted A/B procedure; we can now approximate $V(p')$ by $1/W$, although it will represent a slight underestimate.

We now have a simple—in fact, immediate—estimator of p combined with a relatively simple procedure for gauging its significance. Suppose instead that we had made A/B estimates separately for each sibship size and then combined them optimally. The problem would be that the optimal weights depend on p. We would have to determine that estimate p' of p such that the weights implied by p', when applied to the data, yielded p', a kind of circular procedure—much the same as for $\hat{p}$, the maximum likelihood estimator.

In a way, a problem something like this still exists. It arises because $V(p')$ depends not on p' but on p. If we employ p' instead of p in entering the tables to estimate $V(p')$, adding and subtracting the necessary multiples of the standard error to p' would not set limits on p—rather, it would indicate the range in which the statistic p' would fall if the parameter p happened to equal the observed p'.

The proper device for setting limits on p is well understood. To set a lower limit on p, we attempt to find that p_L which, when used in entering the information table (interpolating when necessary), yields a standard error such that

$$p_L + t \text{ S.E. } (p' \text{ based on } p_L) = p' = A/B ,$$

where

$$\text{S.E. } (p' \text{ based on } p_L) = 1/\sqrt{W(p_L)} ,$$

and t corresponds to the significance level employed. Similarly, the upper limit on p is given by

$$p_U - t \text{ S.E. } (p' \text{ based on } p_U) = p' = A/B .$$

As a simple illustration of this principle, consider that we have a Poisson observation of 2. While an observation of 0 can arise with 13.5% probability when the Poisson parameter is 2, the limits on the parameter should not include 0. For with a Poisson parameter of 0, the observation 2 cannot occur, and it has only extremely small probability for other parameter values near 0.

3. *Truncated Binomial and Truncated Poisson Distribution*

One can readily recognize the relationship between the genetic problem of interest here and the problem of evaluating truncated binomial data. Were sibships all of the same size, we would know from the data the number of sibships having a specified number of affected children, but we would not know the number of sibships with no affected children. Only sibships with at least one affected child are brought to our attention.

Suppose it were the case that only sibships with at least two affected children were brought to our attention. We can apply the effective trial principle described in the rationale given above for the A/B estimator. For sibships with $r > 2$ affected children, each of the s children would represent an effective trial, with the total number of successes on such effective trials equal to r. But for sibships with $r = 2$, there would be $s - 2$ effective trials, and no successes achieved under conditions of effective trial. For n sibships of fixed size s, the A/B estimator would be

$$p' = \frac{\Sigma r - 2f(2)}{ns - 2f(2)} ,$$

where $f(2)$ is the frequency of sibships with the minimum observable number of two affected children. Where sibship size varies, the estimate becomes

$$p' = \frac{\text{Total no. of affected children} - 2f(2)}{\text{Total no. of children} - 2f(2)}.$$

And if, in general, only families with k or more positives are observable, the same rationale yields

$$p' = \frac{\Sigma r - kf(k)}{ns - kf(k)} = \frac{\text{Total no. affected children} - kf(k)}{\text{Total no. of children} - kf(k)}.$$

A similar kind of truncation problem can arise where one has made n observations on a Poisson process with mean λ subject to the restriction that zero values or, more generally, values less than k are not observable. We can treat the Poisson as a limiting case of a binomial with infinitesimal p, unlimited size s, such that $\lambda = sp$. If we were using the general truncation result above for estimating sp we would get

$$s\hat{p} = \frac{\Sigma r - kf(k)}{n - (k/s)f(k)},$$

where the r's now represent the truncated Poisson observations. If we now stipulated that s increased without limit, the estimate of λ is given by

$$\lambda = \frac{\Sigma r - kf(k)}{n}.$$

We will not develop further here the properties of the A/B estimator when applied to the general left-truncated binomial or Poisson. We may note that in the truncated Poisson case, the estimator is unbiased.

4. *Some Alternative Estimators*

We have referred above to Li's (1965) partial use of the estimator $p_0 = R/G$, the proportion of affected children after the first affected child. Although Li did not suggest it for that purpose, we have shown that this is an A/B estimator when one wishes to consider the possibility that subsequent family formation is influenced by the status of earlier births. In the absence of this possibility, we must consider p_0 an inefficient estimator, perhaps grossly so. If a sibship of size s has r affected children, p_0 is influenced by the sequence in which the affected and unaffected children arise, while the total information in the data is contained solely in r and s. But, by use of a simple principle, we can convert the inefficient estimator into a somewhat efficient one, as will now be illustrated.

Suppose $s = 5$, $r = 1$; the sibship contribution to R is always 1 less than r, in this case 0, while the contribution to G depends on whether the single affected child is the first-born, second-born, etc. If all permutations are equally likely, the single affected child will on the average have birth order 3, so that on the average there are two subsequent births after the appearance of the affected child. Suppose, in this instance, we take the denominator contribution as the average value 2, rather than as the observed number of subsequent births; this now depends only on r and s and not on the observed sequence.

If $s = 5, r = 2$, we get the following permutational probabilities for the birth order of the first appearing affected child: birth order 1, $2/5 = 0.4$; birth order 2, $3/5 \times 2/4 = 0.3$; birth order 3, $3/5 \times 2/4 \times 2/3 = 0.2$; birth order 4, $3/5 \times 2/4 \times 1/3 \times 2/2 = 0.1$; birth order 5, 0. The average birth order is 2, and the average number of subsequent births is 3. Thus the R contribution is $r - 1 = 1$, the average G contribution is 3.

We may in general determine that, in a sibship of size s with r affected children, the permutational average of the number of subsequent births after the first affected child is $\bar{G}$ contribution $= (rs - 1)/(r + 1)$. If we have n sibships each of size s and use the permutational average of G in p_0, we get

$$p_0' = \frac{R}{\bar{G}} = \frac{\Sigma r - n}{\Sigma(rs - 1)/(r + 1)},$$

where summation is over individual sibships.

It can be seen that the expression for p_0' will continue to apply even when there is variation in sibship size. If we consider each sibship to provide an estimate of p equal to $(r^2 - 1)/(rs - 1)$, we can see that in p_0' each individual sibship is weighted by $(rs - 1)/(r + 1)$. For varying s, this gives reasonably larger weights to larger sibships, suggesting that the estimator across sibship sizes is a proper extension of the fixed sibship size estimator.

We may note that in obtaining p_0' we were effectively obtaining the harmonic average of the permutational distribution of p_0, that is, $p_0' = 1/\overline{(1/p_0)}$. This seemed proper in view of the fact that each permutation would influence only the denominator of $p_0 = R/G$ but would leave the numerator unaltered. Use of the average $\bar{G}$ resulted in a simple formulation covering multiple sibships and multiple sibship sizes; the general formulation for average p_0 would be much more complex. In point of fact, the permutational arithmetic average of p_0 for a single sibship is the same as the A/B estimate for a single sibship; that is, it equals 0 when $r = 1$, and equals r/s when $r \geq 2$.

Li (1965) proposes yet another independent inefficient estimator which, in the symbolism he employs, is given by $p_1 = (A - B)/(N - B)$. In this "first appearance time" estimator, A is the number of sibships in which the first-born child is affected, B is the number of sibships in which only the last-born child is affected, and N is the number of sibships. We can adopt a permutational approach here also, keeping fixed for the while the sibship size, s. For a sibship with r affected children, the A contribution will average r/s. Where $r \geq 2$, the B contribution will of necessity equal zero; while where $r = 1$, the B contribution will average $1/s$, being unity with probability $1/s$. Inserting the averages, we get

$$p_1' = \frac{\Sigma(r/s) - [f(1)/s]}{N - [f(1)/s]} = \frac{\Sigma r - f(1)}{Ns - f(1)}.$$

We note that the permutational principle has yielded an estimator identical to our highly efficient A/B estimator in the case of fixed sibship size. What happens to p_1' following permutation when s is allowed to vary? If we cumulated the numerators and denominators over varying s in the simplified expression for p_1' (after multiplying numerator and denominator by s), we would remain with the efficient A/B estimator.

But Li's expression for p_1 does not call for this simplification, so that the corresponding p_1' would require cumulating the numerators and denominators before simplification. The effective result is to give substantially the same weight to small sibships as to large sibships, and, in consequence, there is a gross loss in efficiency.

This last may be taken as an example of a possibility raised elsewhere in this paper. The estimator p_1' is highly efficient where the data involve a fixed sibship size but becomes inefficient when sibship size is allowed to vary. One method of modifying p_1' to correct for this deficiency would have resulted in obtaining the proposed $p' = A/B$ estimator.

We have in this section illustrated a permutation principle by which inefficient estimators can be improved. This has been applied to two estimators suggested by Li (1965). Li's $p_0 = R/G$ estimator is readily converted to $p_0' = R/\tilde{G}$. For a single sibship $(\overline{R/G})$ is equivalent to use of the presently proposed $p' = A/B$ estimator. Li's proposed $p_1 = (A - B)/(N - B)$ [symbolism of that paper] is converted to the $p' = A/B$ estimator for the case of a single sibship size. Where sibship size may vary, the converted estimator remains inefficient because of improper weighting—by modifying the weightings (multiplying numerator and denominator contributions by s), the converted estimate again becomes $p' = A/B$.

SUMMARY

A nearly fully efficient but extremely simple method of estimating the human segregation ratio under complete ascertainment has been described and illustrated by numerical examples. The method amounts to discarding the recessive member from families having only one such member and then counting the remaining children, both normals and recessives. A table has been provided to facilitate calculation of the approximate variance. The method may be extended to binomial and Poisson distributions with truncation at any point and applicable to problems other than genetical ones. A permutational principle for improving other inefficient estimates has been discussed.

REFERENCES

HALDANE, J. B. S. 1932. A method for investigating recessive characters in man. *J. Genet.* **25**:251–255.

HALDANE, J. B. S. 1938. The estimation of the frequencies of recessive conditions in man. *Ann. Eugen.* (London) **8**:255–262.

LI, C. C. 1961. *Human genetics: principles and methods.* McGraw-Hill, New York, chap. 5.

LI, C. C. 1964. Estimate of recessive proportion by first appearance time. *Ann. Hum. Genet.* (London) **28**:177–180.

LI, C. C. 1965. Segregation of the Ellis-van Creveld syndrome as analyzed by the first appearance method. *Amer. J. Hum. Genet.* **17**:343–351.

LI, C. C. 1966. A new method of studying Mendelian segregation in man. *Proceedings of the symposium on mutation in population.* Publishing House, Czechoslovak Academy of Sciences, Prague. Pp. 155–166.

MCKUSICK, V. A., EGELAND, J. A., ELDRIDGE, R., and KRUSEN, D. E. 1964. Dwarfism in the Amish. Part 1. The Ellis-van Creveld syndrome. *Bull. Johns Hopkins Hosp.* **115**:306–336.

MANTEL, N. 1951. Evaluation of a class of diagnostic tests. *Biometrics* **7**:240–246.

WEINBERG, W. 1912. Methode und Fehlerquellen der Untersuchung auf Mendelsche Zahlen beim Menschen. *Arch. Rass. Ges. Biol.* **9**:165–174.