Sequential Tests for the Detection of Linkage[1]

NEWTON E. MORTON

University of Wisconsin

INFORMATION ON LINKAGE in man is accumulated as a succession of samples, each of which is typically small relative to the amount of data required to detect even moderately close linkage. The best method of analysis for such sequential samples, in the sense of requiring the least number of observations consistent with a given risk of error, has been found to be a sequential probability ratio test (Wald, 1947). It will now be shown that this test, in addition to minimizing the number of observations, is in other respects a useful method for the detection of linkage in man.

1. THE ASSUMPTIONS

Consider two gene loci, G and T, not necessarily on the same chromosome. An individual of genotype GG′ TT′ may be of either of two possible phases, GT/G′T′ or G′T/GT′, corresponding to his formation by the union of GT and G′T′ gametes, or of G′T and GT′ gametes. If the G and T loci happen to be on the same chromosome, these two phases correspond to the usual meanings of coupling and repulsion. In any case, the frequencies of the four types of gametes produced by this individual, if he is GT/G′T′, will be

$$\tfrac{1}{2}(1-\theta)\mathrm{GT},\quad \tfrac{1}{2}(1-\theta)\mathrm{G'T'},\quad \tfrac{1}{2}\theta\mathrm{GT'},\quad \tfrac{1}{2}\theta\mathrm{G'T},$$

whereas, if he is G′T/GT′, they will be

$$\tfrac{1}{2}\theta\mathrm{GT},\quad \tfrac{1}{2}\theta\mathrm{G'T'},\quad \tfrac{1}{2}(1-\theta)\mathrm{GT'},\quad \tfrac{1}{2}(1-\theta)\mathrm{G'T},$$

where θ is the probability of recombination between the two loci ($0 \leq \theta \leq 1$; nearly always, $\theta \leq 1/2$).

Now, a sufficient set of assumptions for a "linkage" test is the following:

1. The parental genotypes are known with certainty, except for phase.
2. The segregation ratios are not disturbed by incomplete penetrance or differential viability.
3. The method of ascertainment and selection of families is properly allowed for.

With this postulational basis, the null hypothesis to be tested is that "the three assumptions are correct and the recombination fraction in the population equals 1/2". Some of the alternative hypotheses are:

1. Incomplete penetrance or differential viability.
2. Biased ascertainment or selection of families.
3. Nonrandom segregation of nonhomologous chromosomes.
4. Co-existence of the two loci on the same chromosome (linkage).

Received May 28, 1955.

[1] Paper number 576 from the Department of Genetics.

Although a distinction between nonrandom chromosome segregation and linkage (which is presumably much the commoner of the two phenomena) will not be possible until the human linkage groups are better known, it should not be difficult to recognize the other disturbing factors in data that have been carefully collected and reported.

The above assumptions are rather stringent and must be examined in detail. Cases to be treated in this paper include incomplete ascertainment, uncertain parental genotypes, and incomplete penetrance.

No attempt will be made to treat "linkage" tests in which the basis of either character is not a single Mendelian factor. If the basis of one or both conditions is multifactorial or unknown, "linkage" is at best ambiguous and generally cannot be distinguished from any other phenotypic correlation which varies among families. The exploration of these complicated situations may be of some interest, but to include such characters on fancied "linkage" maps, as some authors have done, is to depreciate the linkage maps that have been determined with some precision in other organisms.

Since even the most conservative set of assumptions confounds linkage with other phenomena, the burden of proof is on the investigator who asserts that a particular example of linkage-like effects is evidence of true linkage. When two genes satisfy regular Mendelian ratios, however, it is convenient to denote such effects as linkage, with the assurance that this designation is rather precise, and that its precision will increase as the human linkage map is developed.

2. CURRENT TEST PROCEDURES

The three methods most commonly used to detect human linkage are the method of efficient scores (u scores), the Penrose sib-pair method, and the probability ratio method of Haldane and Smith (1947). Smith (1953) has recently shown that they are all really different forms of the nonsequential probability ratio test.

Valid scoring procedures were first applied to human linkage by Bernstein (1931), who showed that each family can be assigned a score whose sum, expected value, and variance provide a test of the null hypothesis in any body of data that is sufficiently large for the distribution of the total score to be nearly normal. Bernstein's scores were further developed by Hogben (1934) and Haldane (1934), but the evolution by Fisher of a maximum likelihood scoring procedure made these methods obsolete. Fisher (1935) was able to show that his u scores are more efficient than Bernstein's scores for all linkage intensities and are, in fact, fully efficient in the limit for loose linkage. Finney (1940 et seq.) has treated a great variety of cases by u scores, which are now commonly considered to be the method of choice whenever the amount of data is large and the families are not grouped into large pedigrees. However, u scores have certain disadvantages, some of which Smith (1953) has summarized as follows:

1. Although u scores are very easy to use when the parental genotype is completely known (except for phase), the calculation of the variance may be intractable when the parental genotypes are unknown. In large samples this can be circumvented by the use of a simple approximation (Smith, 1953).

2. The u scores are fully efficient only in the limit for loose linkage, which it is not practicable to detect. An ideal test would be efficient for moderate rather than loose linkage.

3. Information about linkage can be greatly increased by using data involving 3 or more generations. It is not feasible to extract this information by u scores.

4. The assumption of normality for the total score may be far from true for moderate sample sizes. Haldane (1946) has developed a normalizing transformation for such cases, and shown that in one instance an exact test fails to confirm the significance of a u score test.

The sib-pair method of Penrose has sometimes been recommended as an alternative to u scores when the parental genotypes are unknown. The investigations of Finney (1942) do not support this recommendation, since in his data the sib-pair method extracted only a small fraction of the information that could be obtained by u scores. However, when one of the test characters is a rare recessive trait, the sib-pair method fares somewhat better (Penrose, 1953). A serious disadvantage of the method is that it may be quite inexact when, as the current procedure requires, a family of size $s > 2$ is partitioned into all $s(s - 1)/2$ possible pairs (Penrose, 1953; Smith, 1953). Smith (1953) has shown how a large-sample correction for non-independence of sib pairs may be applied, but its use destroys the principal advantage of the method, that of arithmetical simplicity. Finney (1941a) has pointed out that the Penrose sib-pair method is particularly sensitive to heterogeneity in gene frequencies when different populations are pooled. The sib-pair method can be applied to traits whose mode of inheritance is unknown, but then the term "linkage" is scarcely appropriate.

The probability ratio test of Haldane and Smith (1947) was devised to extract information from families and pedigrees without making the assumption of normality that is required by the maximum likelihood method. Their test depends on the theorem that the expected value of a probability ratio is 1 on the null hypothesis, regardless of the alternative hypothesis (Wald, 1947). Since this is true for any simple hypothesis, it must be true for any composite hypothesis, which is merely a weighted average of simple hypotheses such that the sum of the weights is 1. Let Λ be a probability ratio for the test of the null hypothesis that $\theta = 1/2$ against some alternative hypothesis. Then, on the null hypothesis, the inequality

$$\Lambda > A, \qquad (A > 1)$$

cannot occur with probability greater than $1/A$, since if it did, this in itself would be enough to raise the mean value $E(\Lambda)$ to 1, and therefore the occurrence of a value of Λ greater than A is at least as strong evidence against the null hypothesis as a significance level of $1/A$. Clearly this method of analysis has several advantages, among them its reliability in small as well as large samples, its dependence solely on elementary laws of probability, and the ease with which all kinds of families and pedigrees may be combined. However, the method is conservative, and a recent modification (average backward odds) is less efficient (Smith, 1953).

The three common methods of linkage detection in man do not exhaust the procedures that have been proposed, but of the current tests, the u statistics of Fisher and Finney and the probability ratio method of Haldane and Smith are the best alternatives to sequential tests.

3. SEQUENTIAL TEST PROCEDURES

Let $f(y; \theta)$ denote the distribution of a random variable y, where θ is the recombination fraction and successive observations on y are indicated by $y_1, y_2, \ldots,$ etc. The observation $y = 1$ signifies that $f(y; \theta)$ is of the form $f(1; \theta)$, and so on. For example, double backcross families of size 2 have two possible forms of the function $f(y; \theta)$, which may arbitrarily be specified by $y = 1$ and $y = 2$. Under the conditions of Section 8 below,

$$f(1; \theta) = \theta^2 + (1 - \theta)^2$$

$$f(2; \theta) = 2\theta(1 - \theta).$$

Thus, a particular sample of 3 independent sib pairs might be $y_1, y_2, y_3 = 2, 1, 2$, and the probability of this sample is $f(2; \theta)f(1; \theta)f(2; \theta)$.

Let H_0 be the null hypothesis that $\theta = 1/2$ and H_1 be the alternative hypothesis that $\theta = \theta_1$. The probability that a sample $y_1, y_2, \cdots, y_m$ is obtained is given by

$$p_{1m} = f(y_1 ; \theta_1) \cdots f(y_m ; \theta_1)$$

when H_1 is true, and by

$$p_{0m} = f(y_1 ; 1/2) \cdots f(y_m ; 1/2)$$

when H_0 is true. The sequential test (Wald, 1947) employs the probability ratio p_{1m}/p_{0m} and two positive numbers A and B, with $A > 1$ and $B < 1$. For purposes of practical computation it is much more convenient to work with the logarithm of this ratio rather than the ratio itself, since

$$\log \frac{p_{1m}}{p_{0m}} = \log \frac{f(y_1; \theta_1)}{f(y_1; 1/2)} + \cdots + \log \frac{f(y_m ; \theta_1)}{f(y_m ; 1/2)}.$$

Let z_i denote the i^{th} term in this sum, viz.,

$$z_i = \log \frac{f(y_i ; \theta_1)}{f(y_i ;1/2)}.$$

The test procedure is carried out as follows, the quantities z_i $(i = 1, 2, \cdots)$ being used: with each accession of data (consisting of one or more families or pedigrees), the cumulative sum $z_1 + \cdots + z_m$ is computed. If

$$\log B < z_1 + \cdots + z_m < \log A$$

the evidence on linkage is not decisive, and judgment with the preassigned significance level and power must be suspended until more data can be collected. If

$$z_1 + \cdots + z_m \geq \log A$$

there is significant evidence for linkage under the assumptions of the test. If

$$z_1 + \cdots + z_m \leq \log B$$

the recombination fraction is significantly greater than θ_1.

More data can always be used following a sequential test, either to estimate a significant linkage or to detect or exclude linkage in the range $\theta_1 < \theta \leq 1/2$, but this latter enterprise may be unprofitable if a stringent choice was made for θ_1.

The constants A and B are related to α, the probability of rejecting H_0 when H_0 is true (a Type I error), and β, the probability of rejecting H_1 when H_1 is true (a Type II error). In practice, two simple approximations are used to determine A and B:

$$A \cong \frac{1-\beta}{\alpha}$$

$$B \cong \frac{\beta}{1-\alpha}$$

Wald (1947) has shown that these approximations cannot result in any appreciable increase in the value of either α or β, and that they may be used to obtain expressions for the power function $P(\theta)$ and the average sample number function $E(n)$ of a sequential test. These two functions determine the best sequential test for a particular purpose and the extent of its superiority over nonsequential procedures. Requirements to impose on these functions are suggested by the probability distribution of θ.

4. THE PROBABILITY DISTRIBUTION OF THE RECOMBINATION FRACTION θ

Haldane and Smith (1947) have suggested "chiefly from a comparison with the known linkage values of *Drosophila*" that it may not be a bad approximation to assume that the recombination fraction for linked genes has a uniform distribution from 0 to 1/2. The distribution may also be arrived at more pedantically.

Consider a chromosome with genetic map length of L morgans, along which gene loci are distributed uniformly. We need not assume that the genes are distributed uniformly along the physical chromosome, only that their locations on the linkage map are so distributed. Choose two loci at random with locations C_1 and C_2, where C_1 is the first locus chosen. The quantity $w = |C_1 - C_2|$ is called the *map distance* between the two loci ($0 < w < L$). The cumulative density function of w may be represented on $(C_1/L, C_2/L)$ coordinates by the area within a unit square between the lines $w = C_2 - C_1$ and $w = C_1 - C_2$, or

$$F(w) = \frac{2}{L^2}\{\tfrac{1}{2}L^2 - \tfrac{1}{2}(L-w)^2\} = \frac{2Lw - w^2}{L^2}.$$

Kosambi (1944) has shown that the map distance w is related to the recombination fraction θ as

$$w = \tfrac{1}{4}\log\frac{1+2\theta}{1-2\theta}, \qquad 0 < \theta < \tfrac{1}{2}$$

assuming that the coincidence is 2θ. By this approximation

$$F(\theta) = \frac{\log \frac{1+2\theta}{1-2\theta}}{2L} - \frac{\left\{\log \frac{1+2\theta}{1-2\theta}\right\}^2}{16L^2}$$

and the probability distribution of θ for linked genes, gotten by differentiating $F(\theta)$, is

$$f(\theta) = \frac{2L - \frac{1}{2}\log \frac{1+2\theta}{1-2\theta}}{L^2(1-4\theta^2)}, \qquad 0 < \theta < \theta' < 1/2$$

$$= 0 \text{ elsewhere.}$$

The critical point θ' beyond which $f(\theta) = 0$ is determined by the equation

$$L = \tfrac{1}{4}\log \frac{1+2\theta'}{1-2\theta'} = \tfrac{1}{2}\tanh^{-1} 2\theta'$$

$$\therefore \theta' = \tfrac{1}{2}\,\frac{1-e^{-4L}}{1+e^{-4L}}.$$

We may verify that $f(\theta)$ is a density function over the interval 0 to θ';

$$F(\theta') = \frac{4L}{2L} - \frac{16L^2}{16L^2} = 1$$

since

$$\log \frac{1+2\theta'}{1-2\theta'} = 4L.$$

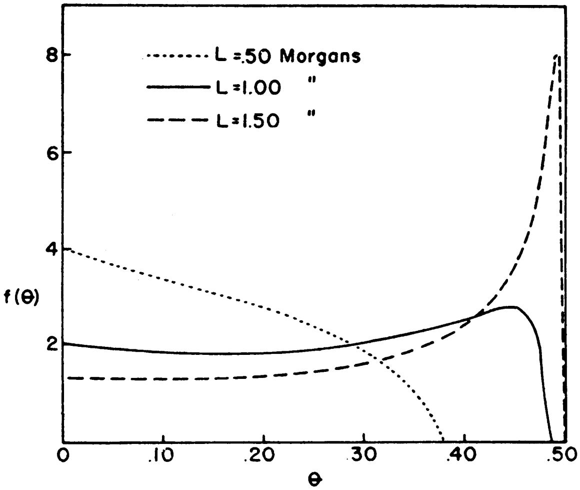

FIG. 1. The distribution of the recombination fraction θ for chromosomes of length L

TABLE 1.—THE DISTRIBUTION OF GENETIC MAP LENGTHS (L) IN DIFFERENT ORGANISMS

Source	L = .25	L = .50	L = .75	L = 1.00	L = 1.25	L = 1.50	L = 2.00	L = 2.50	L = 3.00	$\overline{L^2}/n(\bar{L})^2$
Drosophila[1]	—	—	1	2	—	—	—	—	—	.345
Corn (Zea)[2]	1	1	3	2	2	1	—	—	—	.117
Mouse[3]	—	15	—	48	—	46	13	3	2	.058

[1] Linkage map, neglecting the dot-like IVth chromosome, $L_{IV} = .002$ (Bridges and Brehme, 1944).

[2] Linkage map (Rhoades, 1950).

[3] Based on chiasma frequency in random chromosomes, assuming $L = \frac{\text{chiasma frequency}}{2}$ (Crew and Koller, 1932).

Recent data (Carter, 1955) suggest that the average value of L in the mouse is nearer to unity than here indicated, hence the distribution $g(\theta)$ in Figure 2 should presumably be even closer to uniformity.

Figure 1 shows $f(\theta)$ corresponding to different values of L. For chromosomes of length near unity (100 centimorgans) the distribution of θ is almost uniform. In fact, the recombination fraction has an exactly uniform distribution for chromosomes of unit genetic length according to the simple mapping function $\theta = w - \frac{1}{2} w^2$ $(0 < w < 1)$, for since $F(w) = 2w - w^2$, the distribution of θ is

$$F(\theta) = 2\theta, \qquad 0 < \theta < 1/2$$

$$f(\theta) = 2.$$

Actually chromosomes of unit length are nearly modal in the few higher organisms whose genetic maps are known. Table 1 gives the distribution of L for Drosophila, corn, and (very approximately) for the mouse. On the assumption of a uniform density of loci on the chromosome map, the probability distribution of the recombination fraction between two randomly chosen loci is

$$g(\theta) = \frac{\sum_L L^2 f(\theta)}{\sum_L L^2}.$$

Figure 2 shows that in all three species $g(\theta)$ is closely approximated by a uniform distribution, and that the greatest departure from this approximation is for values of θ close to 1/2, which in practice could seldom be distinguished from independent assortment. The distribution $g(\theta)$ is probably much the same in man, where the average genetic length, based on mean chiasma frequency, may be close to unity (Schultz, unpublished; cited by Neel, 1949).

Table 1 may also be used to compute the probability ϕ that two randomly chosen loci be on the same chromosome. If the number of loci per chromosome is proportional to L,

$$\phi = \frac{\sum L^2}{(\sum L)^2} = \frac{\overline{L^2}}{n(\bar{L})^2}$$

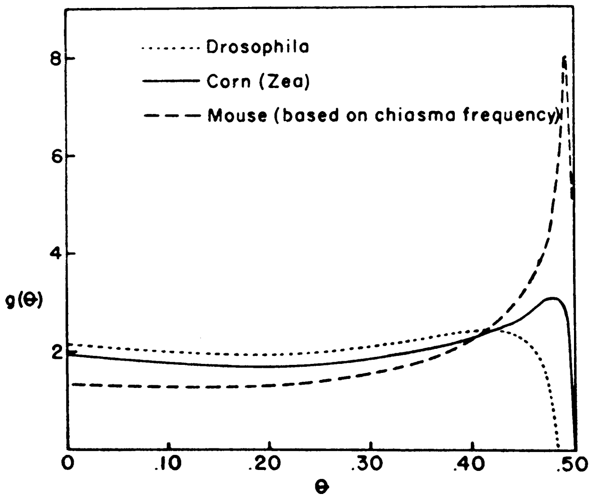

FIG. 2. The distribution of the recombination fraction θ for linked genes in three different species

where n is the haploid number of chromosomes. If all chromosomes are of equal length, $\phi = 1/n$, and for the organisms tabulated this turns out to be a good approximation. In Drosophila, neglecting the dot-like IVth chromosome, $n = 3$, $\phi = .345$; in corn, $n = 10$, $\phi = .117$; in the mouse, $n = 20$, $\phi = .058$, or $\phi = .064$ if pachytene length is proportional to L (Slizynski, 1949). In man, with 23 autosomes, the frequency of autosomal linkage may reasonably be taken as $\phi = .05$, so that the distribution of recombination values in man may be approximated as follows:

$$\begin{aligned} g(\theta) &= 2\phi = .10 \qquad 0 < \theta < 1/2 \\ &= 1 - \phi = .95 \qquad \theta = 1/2 \\ &= 0 \qquad \text{elsewhere.} \end{aligned}$$

5. THE CHOICE OF A SEQUENTIAL TEST

The validity of a sequential test does not depend on the accuracy of these approximations, but they do suggest criteria by which a suitable sequential test may be selected. We are especially anxious to avoid the assertion that two genes are linked when in fact they are not, since a misleading linkage map is worse than no linkage map at all. One source of linkage-like effects can be nearly eliminated by considering only pairs of loci which satisfy our assumption that the expected segregation ratios for both loci are realized in the population sampled. However, cases of apparent linkage will still be made up in part of true linkages, in part of Type I errors. If the prior probability of linkage is $\phi = .05$, then the posterior probability that a case of apparent linkage be a Type I error is

$$\rho = \frac{\alpha(1 - \phi)}{\alpha(1 - \phi) + \phi\bar{P}} = \frac{19\alpha}{19\alpha + \bar{P}}.$$

where P is the average power of the test, or the probability of detecting linkage when it is present. R. S. Krooth (personal communication) has termed ρ the *reliability* and $\bar{P}$ the *sensitivity* of a linkage test. Calculations of ρ for different values of α and $\bar{P}$ show that the usual values of α are inadequate in this problem, and that for the posterior probability of a Type I error to be less than .05, α must be about .002 when $\bar{P} = .95$, .001 when $\bar{P} = .60$ and .0005 when $\bar{P} = .20$ (cp. Haldane, 1934).

Having placed the requirement on α that it be small enough to reduce the posterior probability of a Type I error to .05, we impose a second condition on the power function of the test. To be at all useful, the test must have a power close to unity for values of θ near zero. We are at liberty to choose θ_1, the formal alternative to $\theta_0 = 1/2$, as near to 1/2 as we please, and the only adverse effect of this choice is to increase the average sample number. On this reasoning it seems appropriate to let θ_1 take the largest value which is likely to give a significant result in a practicably large body of data, and to consider the average sample number function a basis for the selection of a sequential test.

As an application of this argument, consider four sequential test procedures defined by the relations

(1) $\theta_1 = .05, \quad A = 2000, \quad B = .01, \quad \theta_0 = 1/2$

(2) $\theta_1 = .10, \quad A = 1000, \quad B = .01, \quad \theta_0 = 1/2$

(3) $\theta_1 = .20, \quad A = 1000, \quad B = .01, \quad \theta_0 = 1/2$

(4) $\theta_1 = .30, \quad A = 1000, \quad B = .01, \quad \theta_0 = 1/2$

and assume that the data consist entirely of double backcross sibships of size 2, sampled under the conditions of §8 below. The probability can take only the value $f(1; \theta) = \theta^2 + (1 - \theta)^2$, corresponding to a sib pair that is either concordant in both traits or discordant in both, and $f(2; \theta) = 2\theta(1 - \theta)$, which corresponds to a sib pair that is concordant in one trait and discordant in the other. Following Wald (1947) and assuming that the excess over the boundaries at the termination of the test can be neglected, we obtain a good approximation to the power function $P(\theta)$ by solving two equations for various values of h

$$P(\theta) = \frac{1 - B^h}{A^h - B^h}$$

and

$$\sum_y f(y; \theta) \left[\frac{f(y; \theta_1)}{f(y; 1/2)}\right]^h = 1.$$

From the power function, again neglecting the excess over the boundaries, we obtain the average sample number function as

$$E_\theta(n) = \frac{P(\theta) \log A + [1 - P(\theta)] \log B}{E_\theta(z)}$$

where

$$E_\theta(z) = \sum_y f(y; \theta) \log \left[\frac{f(y; \theta_1)}{f(y; 1/2)}\right].$$

In particular,

$$E_{\theta_1}(n) = \frac{(1 - \beta) \log A + \beta \log B}{E_{\theta_1}(z)}$$

and

$$E_{\theta_0}(n) = \frac{\alpha \log A + (1 - \alpha) \log B}{E_{\theta_0}(z)} \qquad \text{(Wald, 1947).}$$

The power functions and average sample number functions for the four test procedures are plotted in figures 3 and 4, the information from which is summarized in table 2. All four tests have power greater than .99 for values of θ less than .05 and power less than .03 for values of θ greater than .40. In the intervening range, the first test has good power at $\theta = .10$, the second is moderately good at $\theta = .20$, the third has appreciable power at $\theta = .30$, and the fourth is good for all values of θ less than $\theta = .35$. The value of α has been taken so as to keep the posterior probability of a Type I error (ρ) nearly constant and less than .05, provided that the assumptions of the previous sections are satisfied. The average power $\overline{P}$ increases from .28 to .71, and the average sample number, which represents the cost of this gain in power, increases from 10 to 355.

The investigator will probably seldom have need for sequential tests outside the above range. A test so insensitive as not to detect virtually all cases of close linkage ($\theta < .05$) is of little use, while an increase in sensitivity much beyond $\theta_1 = .30$ requires a prohibitively large average sample number: for example, when $\theta = 1/2$, the test $\theta_1 = .40$, A = 1000, B = .01 requires an average sample number of 5700 double backcross sib pairs.

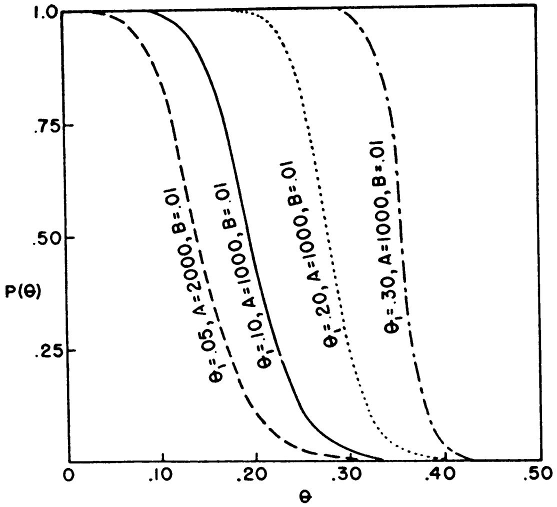

FIG. 3. The power function $P(\theta)$ for different values of θ_1. Double backcross sibships of size 2

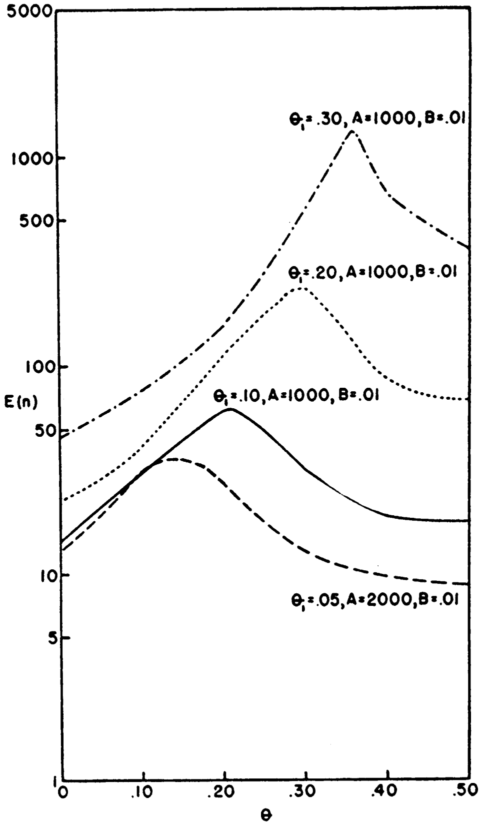

FIG. 4. The average sample number E(n) for different values of θ_1. Double backcross sibships of size 2

6. THE NUMBERS OF OBSERVATIONS REQUIRED BY FIXED-SAMPLE-SIZE TESTS AND SEQUENTIAL TESTS

The exposition so far has considered criteria by which a sequential test may be chosen, and has suggested a battery of four tests which should be adequate for most purposes. We still require, however, to select among these procedures and, more immediately, to determine whether a sequential test is so superior to current fixed-sample-size tests in efficiency, computational simplicity, or exactness that the choice of a sequential test has more than academic interest.

For a start, we may calculate the number of independent double backcross sib pairs required by current tests of strength (α, β). In the case of u statistics there are two possible scores, 1 and -1, with frequencies $\theta^2 + (1 - \theta)^2$ and $2\theta(1 - \theta)$ respectively (Finney, 1940). The expected value of the score is $\mu_\theta = (1 - 2\theta)^2$, with variance $\sigma_\theta^2 = (1 - \mu_\theta)(1 + \mu_\theta)$. (Note that these symbols designate the expected value and variance of the score, not of θ.) If the sample size is small, it may be estimated by trial and error from a table of the cumulative binomial distribution, using the parameters $p_1 = 2\theta_1(1 - \theta_1)$ and $p_0 = 2\theta_0(1 - \theta_0) = 1/2$. If the sample

TABLE 2.—CHARACTERISTICS OF FOUR SEQUENTIAL TESTS

θ_1 = the formal alternative to the null hypothesis that $\theta = \frac{1}{2}$.
α = the probability of rejecting the null hypothesis when $\theta = \frac{1}{2}$.
β = the probability of accepting the null hypothesis when $\theta = \theta_1$.
$P(\theta)$ = the probability of detecting linkage when the true recombination fraction is θ.
$\overline{P}$ = the probability of detecting linkage when θ is uniformly distributed between 0 and $\frac{1}{2}$.
ρ = the probability that a significant "linkage" be a Type I error.
$\overline{E(n)}$ = the average number of double backcross sibships of size 2 required to terminate the test.

θ_1	α	β	$P(\theta)$				$\overline{P}$	ρ	$\overline{E(n)}$
			$\theta = .10$	$\theta = .20$	$\theta = .30$	$\theta = .35$			
.05	.0005	.01	.86	.10	.006	.002	.28	.032	10
.10	.001	.01	.99	.46	.02	.006	.39	.046	19
.20	.001	.01	>.999	.99	.23	.025	.56	.032	68
.30	.001	.01	>.999	>.999	.99	.64	.71	.026	355

$$\overline{P} = 2\int_0^{1/2} P(\theta)\, d\theta$$

$$\rho \simeq \frac{19\alpha}{19\alpha + \overline{P}}$$

$$\overline{E(n)} \simeq .10\int_0^{1/2} E_\theta(n)\, d\theta + .95E_{1/2}(n)$$

is sufficiently large, the distribution of the sample mean will be nearly normal, and the following conditions will determine $n(\alpha, \beta)$, the required sample number:

$$G\left[\frac{d - \mu_{\theta_0}}{\sigma_{\theta_0}/\sqrt{n}}\right] = 1 - \alpha$$

$$G\left[\frac{d - \mu_{\theta_1}}{\sigma_{\theta_1}/\sqrt{n}}\right] = \beta$$

where d is a preassigned constant defining the critical region of the test and

$$G(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\, dx.$$

If we let t_0 be the value for which $G(t_0) = 1 - \alpha$, and t_1 be the value for which $G(t_1) = \beta$, and observe that $\mu_{\theta_0} = 0$ and $\sigma_{\theta_0} = 1$, then the two conditions may be written as

$$\sqrt{n}\, d = t_0$$

$$\sqrt{n}(d - \mu_{\theta_1}) = t_1\sqrt{(1 - \mu_{\theta_1})(1 + \mu_{\theta_1})}.$$

Solving the above equations, we obtain

$$n = n(\alpha, \beta) = \left[\frac{t_0 - t_1\sqrt{(1 - \mu_{\theta_1})(1 + \mu_{\theta_1})}}{\mu_{\theta_1}}\right]^2.$$

If this expression is not an integer, then, as in all formulae determining fixed sample size, $n(\alpha, \beta)$ is the smallest integer in excess (Wald, 1947).

In the case of the probability ratio test of Haldane and Smith (1947), there are two possible values of the logarithm of the probability ratio, namely

$$z' = \log\left[\frac{f(1; \theta_1)}{f(1; 1/2)}\right] = \log(2 - 4\theta_1 + 4\theta_1^2)$$

and

$$z'' = \log\left[\frac{f(2; \theta_1)}{f(2; 1/2)}\right] = \log[4\theta_1(1 - \theta_1)].$$

The expected value of z_θ is $\mu_\theta = z' - 2\theta(1 - \theta)(z' - z'')$, with variance

$$\sigma_\theta^2 = (z')^2 - 2\theta(1 - \theta)(z' - z'')(z' + z'') - (\mu_\theta)^2.$$

The first condition determining the sample size is

$$\sum z = \log(1/\alpha),$$

and if n is sufficiently large, the second condition becomes

$$\frac{\sum z - n\mu_{\theta_1}}{\sqrt{n}\,\sigma_{\theta_1}} = t_1.$$

Solving for n, we obtain

$$n = n^*(\hat{\alpha}, \beta) = \left[\frac{\sqrt{t_1^2\sigma_{\theta_1}^2 + 4\mu_{\theta_1}\log(1/\alpha)} - t_1\sigma_{\theta_1}}{2\mu_{\theta_1}}\right]^2.$$

For the Haldane-Smith test the true significance level $\hat{\alpha}$ is less by a varying amount than the nominal level α, so that in this respect the test is conservative. Smith (1953) calculated that the median $\hat{\alpha}$ is approximately $\alpha/10$ for $\hat{\alpha} = .001$. The error of the normal approximation in determining $n(\alpha, \beta)$ and $n^*(\hat{\alpha}, \beta)$ is in the opposite direction, since the alternative distribution is skewed toward $\theta_0 = 1/2$, and therefore β and n tend to be underestimated. This error is negligible unless n is very small, and in table 3, which gives the results of these calculations, the smallest value of $n(\alpha, \beta)$ is in close agreement with an exact determination from the cumulative binomial distribution.

TABLE 3.—THE AVERAGE SAMPLE NUMBER E(n) FOR A SEQUENTIAL TEST, COMPARED WITH THE FIXED SAMPLE NUMBERS REQUIRED BY THE FISHER-FINNEY U SCORE TEST, $n(\alpha, \beta)$, AND THE HALDANE-SMITH PROBABILITY RATIO TEST, $n^*(\hat{\alpha}, \beta)$

n = the required number of double backcross sibships of size 2.

θ_1	α	β	E(n)		$n(\alpha, \beta)$	$n^*(\hat{\alpha}, \beta)$
			$\theta = ½$	$\theta = \theta_1$		
.05	.0005	.01	9	20	34	49
.10	.001	.01	18	31	59	89
.20	.001	.01	67	103	214	328
.30	.001	.01	355	529	1,134	1,740
.40	.001	.01	5,700	8,546	18,324	28,420

The conclusions from table 3 are quite simple and consistent. Of the fixed-sample-size tests, u statistics require only about 2/3 as many observations for a given risk of error as the Haldane-Smith probability ratio test. If, in view of the conservatism of the latter test, a value of α ten times as large is used, the number of observations required by the test is intermediate between n(α, β) and n*($\hat{\alpha}$, β), and is still appreciably in excess of the sample size required by the u score test.

Although the superiority of the u score test over the Haldane-Smith probability ratio test is marked, the superiority of the sequential test is even more striking. When the alternative hypothesis is true, the sequential test requires only about 1/2 as many observations as a u score test of the same strength, and when the null hypothesis is true (as it usually will be), the sequential test requires less than 1/3 as many observations as the u score test. Similar savings in the number of observations have been found for other distributions by Wald (1947) and Bross (1952).

For the detection of linkage we have knowledge that the user of a sequential test does not ordinarily have, in that the approximate parameter distribution is known, and we may calculate a mean sequential sample number $\overline{\text{E(n)}}$ averaged over this distribution (table 2). Over the range of tests considered, the mean sample number required by a sequential test of strength (α, β) is less than 1/3 the number required by a u score test of the same strength.

7. CLASSIFICATION OF FACTORS, MATINGS, AND METHODS OF SAMPLING

In view of the considerable saving in observations indicated in the last section, sequential tests would seem to be the method of choice for the detection of linkage. For practical use, the determination of probabilities must be extended to families of different types and sizes. We first require a few definitions.

Consider two loci, G and T, which are to be tested for linkage. The genetic characters which are determined by these loci may be divided into four classes. These are:

1. Recessive abnormalities, such as albinism. The symbols G,g or T,t will be used for factors of this class.

2. Common recessives, such as the gene for the inability to taste phenylthiocarbamide. Symbols G,g or T,t will also be used here.

3. Factors without dominance, the heterozygote being distinguishable from both homozygotes. Sicklemia and the MN blood groups are examples of this class. The letters G_1, G_2 or T_1, T_2 will be used for such factors.

4. "Dominant" abnormalities, such as ovalocytosis. The normal homozygote is exceedingly rare (in most cases never having been observed), and all abnormal persons are therefore assumed to be heterozygous. The symbol G_1 or T_1 will be used for the normal allele, G_2 or T_2 for the abnormal factor.

For a family to give information on linkage, neither parent may be GG or TT and at least one parent must be doubly heterozygous. An informative mating is termed a double backcross, a single backcross, or a double intercross according to whether the other parent is doubly homozygous, singly heterozygous, or doubly heterozygous. Since the phase of linkage is unknown, the probability for a double or single backcross will consist of two terms, one for each possible phase of the doubly heterozygous parent, and the probability for a double intercross will consist of three terms, corresponding to the possibilities that both parents are in coupling, both in repulsion, or that one is in coupling and the other in repulsion. We shall assume that the two phases are at equilibrium in the population, a condition that should nearly always be closely approximated, except perhaps after recent hybridization. On the null hypothesis this assumption is of course supererogatory.

It rarely happens that families selected for a linkage study are effectively a random sample from the general population. Usually families are selected first on the basis of the character determined by the "main" locus and are tested afterwards for the character determined by the "test" locus. There are three methods of selecting families on the basis of the main character (Bailey, 1951):

1. Selection through the parents or grandparents, without consideration of the children. The sampling of families is effectively random, and in families of a given mating type and size, the distribution of the number of children manifesting the main character is a complete binomial series (*complete* selection).

2. Selection through the children themselves, with complete selection of affected individuals. In families of a given mating type and size, the distribution of the number of children manifesting the main character is a truncated binomial series, with the first term missing (*truncate* selection).

3. Selection through the children, with incomplete selection of affected individuals. The distribution of affected individuals in sibships of a given mating type and size is not a truncated binomial, since families with large numbers of affected children are more likely to be ascertained than families with a smaller number of abnormals (*arbitrary* selection). This is the usual method of selection for recessive abnormalities and a not uncommon method of selection for "dominant" abnormalities and rare factors without dominance.

Except in cases of gross ascertainment bias, the test character is never subject to incomplete selection of affected individuals (method 3).

It should be noted that these three methods of selecting families for analysis subsume the rejection of some classes of ascertained families. The fundamental attribute of each type of selection is the distribution to which it gives rise, regardless of how the families were detected. For example, with recessive genes the propositus is sometimes an affected parent mated to a normal dominant, who may be either homozygous or heterozygous. A mating of a dominant parent is called "certain" if there is at least one recessive child (in which case the dominant parent must be heterozygous), and is called "doubtful" otherwise. Sampling is by method 1 or 2, according to whether doubtful families are included or rejected. The method of ascertainment is the same in both cases, but the method of selection is different, and determines the proper method of analysis.

8. BOTH CHARACTERS SELECTED THROUGH THE PARENTS (COMPLETE SELECTION). PARENTAL GENOTYPES KNOWN, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

Unless there is no dominance for either character, some of the families will usually be of uncertain parental genotype. If these doubtful families are analysed separately

TABLE 4.—MATINGS SCORED WITH z_1. DOUBLE BACKCROSSES AND SINGLE BACKCROSSES WITH NO DOMINANCE IN THE INTERCROSS FACTOR

$s = a + b + c + d$

Parental genotype	Mating Type	Progeny Phenotype				Uninformative Progeny
		a	b	c	d	
$Gg\ Tt \times gg\ tt$	1	$G\ T$	$G\ t$	$g\ T$	$g\ t$	—
$Gg\ T_1T_2 \times gg\ T_1T_1$	2	$G\ T_1$	$G\ T_1T_2$	$g\ T_1$	$g\ T_1T_2$	—
$G_1G_2\ Tt \times G_1G_1\ tt$	3	$G_1\ T$	$G_1\ t$	$G_1G_2\ T$	$G_1G_2\ t$	—
$Gg\ T_1T_2 \times gg\ T_1T_2$	4	$G\ T_1$	$G\ T_2$	$g\ T_1$	$g\ T_2$	T_1T_2
$G_1G_2\ Tt \times G_1G_2\ tt$	5	$G_1\ T$	$G_1\ t$	$G_2\ T$	$G_2\ t$	G_1G_2
$G_1G_2\ T_1T_2 \times G_1G_1\ T_1T_1$	6	$G_1\ T_1$	$G_1\ T_1T_2$	$G_1G_2\ T_1$	$G_1G_2\ T_1T_2$	—
$G_1G_2\ T_1T_2 \times G_1G_1\ T_1T_2$	7	$G_1\ T_1$	$G_1\ T_2$	$G_1G_2\ T_1$	$G_1G_2\ T_2$	T_1T_2
$G_1G_2\ T_1T_2 \times G_1G_2\ T_1T_1$	8	$G_1\ T_1$	$G_1\ T_1T_2$	$G_2\ T_1$	$G_2\ T_1T_2$	G_1G_2

Frequency	a	b	c	d	Total
Coupling 1	$1 - \theta$	θ	θ	$1 - \theta$	2
Repulsion 1	θ	$1 - \theta$	$1 - \theta$	θ	2
Total	1	1	1	1	4

$$z_1 = \log \frac{f(y; \theta_1)}{f(y; \tfrac{1}{2})} = \log 2^{s-1} \left[\theta_1^{a+d}(1 - \theta_1)^{b+c} + \theta_1^{b+c}(1 - \theta_1)^{a+d}\right]$$

(see §12), then the methods of this section are appropriate to the certain families. If the doubtful families are rejected, the certain families should be analysed by the methods of §§9–10.

Neglecting multiple allelism, the possible kinds of certain families may be grouped into 5 classes, which by the method of u scores have 3 essentially different scores and 2 derived scores (Finney, 1940). In sequential tests the same classes exist. The scores in a sequential test are "lods", or logarithms of the probability ratio, the five functional forms of which may be denoted by z_1, z_2, z_3, z_4, and z_5, in exact correspondence with the u_{11}, u_{31}, u_{33}, $2u_{31}$, and $2u_{11}$ scoring types of Finney.

Tables 4–8 give the possible certain matings and the lod scores appropriate to them. Matings scored with z_1 (table 4) comprise double backcrosses and those single backcrosses in which there is no dominance for the intercross factor. There is thus a one-to-one correspondence between progeny genotype and phenotype for both loci. Note that some progeny have probabilities that are independent of the recombination fraction and phase, and therefore give no information on linkage. Matings scored with z_2 (table 5) are single backcrosses with dominance in the intercross factor. Matings scored with z_3 (table 6) are double intercrosses with dominance in both factors. Most matings of common occurrence are scored with the z_1, z_2, or z_3 lods, of which the z_1 type is much the most informative.

The two remaining scoring types are of particular interest because the u score method omits progeny from which information is extracted by the lod scores. Matings scored with z_4 (table 7) are double intercrosses with dominance in only one factor. There are six progeny phenotypes, the last two of which have probabilities that are

Table 5.—Matings scored with z_2. Single backcrosses with dominance in the intercross factor

Parental genotype	Mating Type	Progeny phenotype				Uninformative Progeny
		a	b	c	d	
Gg Tt × Gg tt	9	G T	g T	G t	g t	—
Gg Tt × gg Tt	10	G T	G t	g T	g t	—
Gg T_1T_2 × Gg T_1T_1	11	G T_1	g T_1	G T_1T_2	g T_1T_2	—
G_1G_2 Tt × G_1G_1 Tt	12	G_1 T	G_1 t	G_1G_2 T	G_1G_2 t	—

Frequency	a	b	c	d	Total
Coupling 1	$2-\theta$	θ	$1+\theta$	$1-\theta$	4
Repulsion 1	$1+\theta$	$1-\theta$	$2-\theta$	θ	4
Total	3	1	3	1	8

$$z_2 = \log\frac{f(y;\theta_1)}{f(y;\tfrac{1}{2})} = \log\frac{2^{s-1}}{3^{a+c}}\left[(2-\theta_1)^a\theta_1^b(1+\theta_1)^c(1-\theta_1)^d + (1+\theta_1)^a(1-\theta_1)^b(2-\theta_1)^c\theta_1^d\right]$$

Table 6.—Matings scored with z_3. Double intercrosses with dominance in both factors

Parental genotype	Mating Type	Progeny phenotype				Uninformative Progeny
		a	b	c	d	
Gg Tt × Gg Tt	13	G T	G t	g T	g t	—

Frequency	a	b	c	d	Total
G T/g t × G T/g t 1	$3-2\theta+\theta^2$	$\theta(2-\theta)$	$\theta(2-\theta)$	$(1-\theta)^2$	4
G T/g t × G t/g T 2	$2+\theta-\theta^2$	$1-\theta+\theta^2$	$1-\theta+\theta^2$	$\theta(1-\theta)$	8
G t/g T × G t/g T 1	$2+\theta^2$	$1-\theta^2$	$1-\theta^2$	θ^2	4
Total...............	9	3	3	1	16

$$z_3 = \log\frac{f(y;\theta_1)}{f(y;\tfrac{1}{2})} = \log\frac{4^{s-1}}{9^a3^{b+c}}\Big[(3-2\theta_1+\theta_1^2)^a\,\theta_1^{b+c}\,(2-\theta_1)^{b+c}\,(1-\theta_1)^{2d} + 2\,(2+\theta_1-\theta_1^2)^a$$
$$\cdot\,(1-\theta_1+\theta_1^2)^{b+c}\,\theta_1^d\,(1-\theta_1)^d + (2+\theta_1^2)^a\,(1-\theta_1^2)^{b+c}\,\theta_1^{2d}\Big]$$

linear functions of $\theta(1-\theta)$, whereas the other four types include terms which are not linear in $\theta(1-\theta)$, like θ^2. When $\theta \to 1/2$, the deviation of $\theta(1-\theta)$ from 1/4 is vanishingly small compared with the deviation of θ^2 from 1/4, and the last two classes contribute almost no information on linkage. It is not surprising, therefore, that when the probability is expanded in powers of $1-2\theta$, and the cubic and higher terms neglected, the appropriate u score is a function of only the first four classes (Finney, 1940). Since loose linkage ($\theta \to 1/2$) is never in practice distinguished from non-linkage ($\theta = 1/2$), the important consideration is that the information contributed by the neglected progeny (which constitute 1/2 of the total children) is not negligible when θ is small.

TABLE 7. MATINGS SCORED WITH z_4. DOUBLE INTERCROSSES WITH DOMINANCE IN ONE FACTOR

$s = a + b + c + d + e + f$

Parental genotype	Mating Type	Progeny phenotype						Uninformative Progeny
		a	b	c	d	e	f	
Gg T_1T_2 x Gg T_1T_2	14	G T_1	g T_1	G T_2	g T_2	G T_1T_2	g T_1T_2	—
G_1G_2 Tt x G_1G_2 Tt	15	G_1 T	G_1 t	G_2 T	G_2 t	G_1G_2 T	G_1G_2 t	—

Frequency		a	b	c	d	e	f	Total
Coupling x coupling	1	$1 - \theta^2$	θ^2	$\theta(2 - \theta)$	$(1 - \theta)^2$	$2(1 - \theta + \theta^2)$	$2\theta(1 - \theta)$	4
Coupling x repulsion	2	$1 - \theta + \theta^2$	$\theta(1 - \theta)$	$1 - \theta + \theta^2$	$\theta(1 - \theta)$	$1 + 2\theta - 2\theta^2$	$1 - 2\theta + 2\theta^2$	8
Repulsion x repulsion	1	$\theta(2 - \theta)$	$(1 - \theta)^2$	$1 - \theta^2$	θ^2	$2(1 - \theta + \theta^2)$	$2\theta(1 - \theta)$	4
Total....................		3	1	3	1	6	2	16

$$z_4 = \log \frac{f(y;\,\theta_1)}{f(y;\,\tfrac{1}{2})} = \log \frac{4^{s-1}}{3^{a+c+e}2^{e+f}} \{2^{e+f}\theta_1{}^{2b+c+f}(1 - \theta_1)^{a+2d+f}(1 + \theta_1)^{a}(2 - \theta_1)^{c}(1 - \theta_1 + \theta_1{}^2)^{e} + 2\theta_1{}^{b+d}(1 - \theta_1)^{b+d}(1 - \theta_1 + \theta_1{}^2)^{a+c}(1 + 2\theta_1 - 2\theta_1{}^2)^{e}(1 - 2\theta_1 + 2\theta_1{}^2)^{f} + 2^{e+f}\theta_1{}^{a+2d+f}(1 - \theta_1)^{2b+c+f}(1 + \theta_1)^{c}(2 - \theta_1)^{a}(1 - \theta_1 + \theta_1{}^2)^{e}\}$$

TABLE 8. MATINGS SCORED WITH z_5. DOUBLE INTERCROSSES WITH NO DOMINANCE IN EITHER FACTOR

$s = a + b + c + d + e + f + g + h + i$

Parental genotype	Mating Type	Progeny phenotype									Uninformative Progeny
		a	b	c	d	e	f	g	h	i	
G_1G_2 T_1T_2 x G_1G_2 T_1T_2	16	G_1 T_1	G_1 T_2	G_2 T_1	G_2 T_2	G_1G_2 T_1	G_1G_2 T_2	G_1 T_1T_2	G_2 T_1T_2	G_1G_2 T_1T_2	—

Frequency		a	b	c	d	e	f	g	h	i	Total
Coupling x coupling	1	$(1 - \theta)^2$	θ^2	θ^2	$(1 - \theta)^2$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2(1 - 2\theta + 2\theta^2)$	4
Coupling x repulsion	2	$\theta(1 - \theta)$	$\theta(1 - \theta)$	$\theta(1 - \theta)$	$\theta(1 - \theta)$	$1 - 2\theta + 2\theta^2$	$1 - 2\theta + 2\theta^2$	$1 - 2\theta + 2\theta^2$	$1 - 2\theta + 2\theta^2$	$4\theta(1 - \theta)$	8
Repulsion x repulsion	1	θ^2	$(1 - \theta)^2$	$(1 - \theta)^2$	θ^2	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2\theta(1 - \theta)$	$2(1 - 2\theta + 2\theta^2)$	4
Total..................		1	1	1	1	2	2	2	2	4	16

$$z_5 = \log \frac{f(y;\,\theta_1)}{f(y;\,1/2)} = \log \frac{4^{s-1}}{2^{w+i}} \left\{2^{w}\theta_1{}^{2v+w}(1 - \theta_1)^{2u+w}(1 - 2\theta_1 + 2\theta_1{}^2)^{i} + 2^{i+1}\theta_1{}^{u+v+i}(1 - \theta_1)^{u+v+i}(1 - 2\theta_1 + 2\theta_1{}^2)^{w} + 2^{w}\theta_1{}^{2u+w}(1 - \theta_1)^{2v+w}(1 - 2\theta_1 + 2\theta_1{}^2)^{i}\right\}$$

$$u = a + d$$
$$v = b + c$$
$$w = e + f + g + h$$

Matings scored with z_5 (table 8) are double intercrosses with no dominance in either factor. The lod score is based on 9 distinguishable progeny classes, the last 5 of which contribute no information when $\theta \rightarrow 1/2$, and are therefore neglected in computing the u scores (Finney, 1940). When θ is small, however, the information contained in these children (which constitute 3/4 of the progeny) is no longer negligible.

9. ONE CHARACTER SELECTED THROUGH THE PARENTS (COMPLETE SELECTION), THE OTHER THROUGH THE CHILDREN (INCOMPLETE SELECTION). PARENTAL GENOTYPES KNOWN, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

For convenience we may denote the factor that is selected through the children by G,g, G_1, or G_2, and the factor selected through the parents by T, t, T_1, or T_2. The method of this section is appropriate only if families of doubtful parental genotype with regard to the T locus are not rejected (section 12); the selection of the G factor is arbitrary.

In a family of size s let there be s_1 children of one G type, say G, and s_2 of the other ($s_1 + s_2 = s$). The prior probability of the family will be designated by $f(y;\theta)$ and the conditional probability by $f(y;\theta \mid s_1)$. Then

$$f(y;\theta \mid s_1) = \frac{f(y;\theta)}{P(s_1,s_2)}$$

where $P(s_1,s_2)$ is the probability measure of the selected class of families. Since the two characters are selected independently, and the probabilities which are pooled in $P(s_1,s_2)$ are complementary, $P(s_1,s_2)$ is independent of θ and of the phase of linkage and cancels when the probability ratio is formed. Thus the probability ratio and the lod score derived from it have the convenient property of being invariant with respect to biased sampling of one character only, and families selected in this way are scored just as if both characters had been ascertained through the parents (Smith, 1953).

10. BOTH CHARACTERS SELECTED THROUGH THE CHILDREN, COMPLETE SELECTION OF AFFECTED INDIVIDUALS (TRUNCATE SELECTION). PARENTAL GENOTYPES KNOWN, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

Families in which the parental genotype is unknown for either factor are rejected. The condition on both factors makes the marginal distribution of the selected families a function of θ, and the methods of the previous sections require modification. There are three types to be considered, corresponding to the z_1, z_2, and z_3 scoring types. We shall suppose that the selected factors are g and t, since only matings in which both characters are common recessives are likely to be selected in this way.

(1) The z_1 scoring type (Mating 1)

The distribution of the selected families is

$$f(y;\theta \mid g, t) = \frac{f(y;\theta)}{P(g,t)}$$

where P(g,t) is the probability that a mating of this type have at least one g and one t child. To satisfy this condition, it is sufficient that $c + d \neq 0$ and $b + d \neq 0$. Therefore,

$$P(g,t) = 1 - P(c + d = 0) - P(b + d = 0) + P(b + c + d = 0).$$

But $P(c + d = 0) = \sum_{a=0}^{s} \binom{s}{a} \left\{ \frac{1}{2} \left(\frac{1-\theta}{2} \right)^{a} \left(\frac{\theta}{2} \right)^{s-a} + \frac{1}{2} \left(\frac{1-\theta}{2} \right)^{s-a} \left(\frac{\theta}{2} \right)^{a} \right\}$

$$= (1/2)^s = P(b + d = 0)$$

and $P(b + c + d = 0) = P(a = s) = \frac{1}{2} \left\{ \left(\frac{1-\theta}{2} \right)^{s} + \left(\frac{\theta}{2} \right)^{s} \right\}$, and so

$$P(g,t) = \frac{2^s - 2 + \frac{1}{2}\theta^s + \frac{1}{2}(1-\theta)^s}{2^s}.$$

It follows that

$$\log \frac{f(y;\theta_1 \mid g,t)}{f(y;1/2 \mid g,t)} = \log \frac{f(y;\theta_1)}{f(y;1/2)} + \log \frac{P(g,t;1/2)}{P(g,t;\theta_1)}$$

$$= z_1 + c_1$$

where $c_1 = \log \frac{2^s - 2 + (1/2)^s}{2^s - 2 + \frac{1}{2}\theta_1^s + \frac{1}{2}(1-\theta_1)^s}$.

Thus the lod score in this case, and in general, is simply the score appropriate to random sampling plus a correction factor which is determined by the method of selection. The factor c_1 is exactly analogous to $-\epsilon_5$ in the theory of u scores (Finney, 1940).

(2) The z_2 scoring type (Matings 9 and 10)

Using the same notation as before, we find that

$$\log \frac{f(y;\theta_1 \mid g,t)}{f(y;1/2 \mid g,t)} = z_2 + c_2$$

where $c_2 = \log \frac{4^s - 2^s - 3^s + (3/2)^s}{4^s - 2^s - 3^s + \frac{1}{2}(2-\theta_1)^s + \frac{1}{2}(1+\theta_1)^s}$.

(3) The z_3 scoring type (Mating 13)

$$\log \frac{f(y;\theta_1 \mid g,t)}{f(y;1/2 \mid g,t)} = z_3 + c_3.$$

$$c_3 = \log \frac{4^s - 2(3)^s + (9/4)^s}{4^s - 2(3)^s + \frac{1}{4}(3 - 2\theta_1 + \theta_1^2)^s + \frac{1}{2}(2 + \theta_1 - \theta_1^2)^s + \frac{1}{4}(2 + \theta_1^2)^s}.$$

11. BOTH CHARACTERS SELECTED THROUGH THE CHILDREN, ONE COMPLETELY (TRUNCATE SELECTION), THE OTHER INCOMPLETELY (ARBITRARY SELECTION). PARENTAL GENOTYPES KNOWN, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

Let the character with arbitrary selection be denoted by g or G_2, and let t denote the character with truncate selection. The family is ascertained through the

G factor and then tested for the T factor, with rejection of families in which there is not at least one t child. (If these families are not rejected, or if there is no dominance in the T factor, see §9.) Occasionally the method of incomplete ascertainment of the G factor may be known exactly, but the simplest and most reliable procedure is to consider the distribution of the families with the G factor fixed, so that the method of selection does not enter into the argument (Finney, 1940).

A. Dominance in the G factor (G,g type)

Let there be s_1 children of type G and s_2 of type g ($s_1 + s_2 = s$). The distribution of selected families is

$$f(y;\theta \mid s_1,s_2,t) = \frac{f(y;\theta)}{P(s_1,s_2,t)}$$

where $P(s_1,s_2,t)$ is the probability measure of selected families of this class. Note that $s_2 = 0$ implies ascertainment of the G factor through the parents or uninformative children, hence the s_1,s_2 method of scoring is not appropriate unless $s_2 > 0$ or the viability of the G,g types is abnormal.

(1A) The z_1 scoring type (Mating 1)

$$P(s_1,s_2,t) = P(s_1,s_2) - P(s_1,s_2\,,\, b + d = 0)$$

$$P(s_1,s_2) = k\binom{s}{s_1}(1/2)^{s_1}(1/2)^{s_2}$$

$$P(s_1,s_2, b + d = 0) = P(a = s_1\,, c = s_2) = k\binom{s}{s_1}\left\{\frac{1}{2}\left(\frac{\theta}{2}\right)^{s_1}\left(\frac{1-\theta}{2}\right)^{s_2} + \frac{1}{2}\left(\frac{\theta}{2}\right)^{s_2}\left(\frac{1-\theta}{2}\right)^{s_1}\right\}.$$

Therefore,

$$P(s_1,s_2,t) = k\binom{s}{s_1}(1/2)^{s}\{1 - \tfrac{1}{2}\,\theta^{s_1}(1-\theta)^{s_2} - \tfrac{1}{2}\,\theta^{s_2}(1-\theta)^{s_1}\},$$

where k is a selection factor dependent only on s_1 and s_2 and

$$\log\frac{f(y;\theta_1 \mid s_1, s_2, t)}{f(y;1/2 \mid s_1,s_2,t)} = z_1 + e_1$$

where

$$e_1 = \log\frac{1 - (1/2)^s}{1 - \frac{1}{2}\,\theta_1^{s_1}(1-\theta_1)^{s_2} - \frac{1}{2}\,\theta_1^{s_2}(1-\theta_1)^{s_1}}\,.$$

(2A) The z_2 scoring type (Mating 9)

$$\log\frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_2 + e_2$$

$$e_2 = \log\frac{3^{s_1}[1 - (1/2)^s]}{3^{s_1} - \frac{1}{2}\,(2-\theta_1)^{s_1}\theta_1^{s_2} - \frac{1}{2}\,(1+\theta_1)^{s_1}(1-\theta_1)^{s_2}}\,.$$

(3A) The z_2 scoring type (Mating 10)

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_2 + d_2$$

$$d_2 = \log \frac{2^s - (3/2)^s}{2^s - \frac{1}{2}(2-\theta_1)^{s_1}(1+\theta_1)^{s_2} - \frac{1}{2}(1+\theta_1)^{s_1}(2-\theta_1)^{s_2}}\,.$$

(4A) The z_3 scoring type (Mating 13)

$$\log \frac{f(y;\theta \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_3 + e_3$$

$$e_3 = \log \frac{3^{s_1}[1-(3/4)^s]}{3^{s_1} - \frac{1}{4}(3-2\theta_1+\theta_1^2)^{s_1}\theta_1^{s_2}(2-\theta_1)^{s_2} - \frac{1}{2}(2+\theta_1-\theta_1^2)^{s_1}(1-\theta_1+\theta_1^2)^{s_2} - \frac{1}{4}(2+\theta_1^2)^{s_1}(1-\theta_1^2)^{s_2}}.$$

B. Incomplete dominance in the G factor (G_1,G_2 type)

Rare "dominants" and a few characters lacking dominance (sicklemia, thalassemia) are sometimes selected incompletely in this way. This situation was not considered by Finney (1940).

(1B) The z_1 scoring type (Mating 3)

Let s_1 be the number of G_1 children, and s_2 be the number of G_1G_2 children. Then the probability ratio is the same as for type 1A above, and

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_1 + e_1\,.$$

(2B) The z_1 scoring type (Mating 5)

If the family is selected through a G_1G_2 child, then there is random sampling for the informative progeny, and the method of section 9 applies. If selection is through an informative G_1 or G_2 child, then

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_1 + e_1\,,$$

where s_1 is the number of G_1 children and s_2 the number of G_2 children.

(3B) The z_2 scoring type (Mating 12)

Let there be s_1 children of type G_1 and s_2 children of type G_1G_2. The probability ratio is the same as for 3A above, and

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,t)}{f(y;1/2 \mid s_1,s_2,t)} = z_2 + d_2\,.$$

(4B) The z_4 scoring type (Mating 15)

Let there be s_1 children of type G_1, s_2 of type G_1G_2, and s_3 of type G_2 ($s_1 + s_2 + s_3 = s$). Then

$$\log \frac{f(y;\theta_1 \mid s_1,s_2,s_3,t)}{f(y;1/2 \mid s_1,s_2,s_3,t)} = z_4 + e_4$$

$$e_4 = \log \frac{1 - (3/4)^s}{1 - \frac{1}{4}(1 - \theta_1^2)^{s_1}(1 - \theta_1 + \theta_1^2)^{s_2}[\theta_1(2 - \theta_1)]^{s_3} - (1/2)^{s_2+1}(1 - \theta_1 + \theta_1^2)^{s_1+s_3} \cdot (1 + 2\theta_1 - 2\theta_1^2)^{s_2} - \frac{1}{4}[\theta_1(2 - \theta_1)]^{s_1}(1 - \theta_1 + \theta_1^2)^{s_2}(1 - \theta_1^2)^{s_3}}.$$

This completes the analysis of the matings in tables 4–8. These include all the scoring types of Finney (1940), who used 3 essentially different scores, 2 derived scores, 7 score corrections, and 12 essentially different information functions. For the same matings, the probability ratio method requires only 5 scores and 7 correction factors. The development of the probability ratio scores is extremely simple and may easily be extended to more complex cases, such as multiple allelism, uncertain parental genotypes, and pedigree data. To facilitate numerical analysis of the matings that have been treated so far, the scores for small families are given in tables 10–18.

12. PARENTS OF UNKNOWN GENOTYPE, BOTH PARENTS TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

Parental heterozygosity for recessive factors can be established by the observation of recessive children, in the absence of which a family without pedigree information is termed "doubtful". Information may still be extracted from these families, provided that the population gene frequencies are known and that mating is at random with respect to the doubtful locus. We have seen in §9 that when families are selected through the parents for the test factor, and doubtful families are not rejected, then no score correction is needed for families of known parental genotype regardless of how the main character is selected. Matings doubtful for the main character may also be analysed.

In connection with the doubtful families it will be convenient to introduce a few new symbols. Let p_t denote the frequency of the t gene and p_g the frequency of the g gene. Occasionally children will not be scorable for linkage, either because they are uninformative or because they are incompletely tested. If these children are tested for the doubtful character, they give information about the parental genotypes and should enter into the present calculations. Let S be the number of scored and unscored children which are tested for the doubtful character, in contradistinction to s, the number of children which are scored for linkage. As an example of the general procedure, we shall develop scores for the "doubtful" analogues of the z_1 scoring type.

(*1*) *Families doubtful for the t factor* (*Matings 1, 3, 5*)

All children are of type T. The prior probabilities for homozygosity and heterozygosity of the T parent are $(1 - p_t)^2$ and $2p_t(1 - p_t)$, and the conditional probabilities for the children are

$$(1/2)^s \text{ and } \tfrac{1}{2}\{\theta^a(1 - \theta)^c + \theta^c(1 - \theta)^a\}(1/2)^S$$

respectively. Therefore,

$$\log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{2^{S-s} - p_t\{2^{S-s} - \theta_1^a(1 - \theta_1)^c - \theta_1^c(1 - \theta_1)^a\}}{2^{S-s} - p_t\{2^{S-s} - (1/2)^{s-1}\}}.$$

(2) Families doubtful for the g factor (Matings 1, 2, 4)

All children of type G. The probability ratio is the same as for the previous type, except for the substitution of p_g for p_t and b for c.

$$\log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{2^{S-s} - p_g\{2^{S-s} - \theta_1^a(1-\theta_1)^b - \theta_1^b(1-\theta_1)^a\}}{2^{S-s} - p_g\{2^{S-s} - (1/2)^{s-1}\}} .$$

(3) Families doubtful for the g and t factors (Mating 1)

All children of type GT. The GT parent may be GGTT, GgTT, GGTt, or GgTt, only the last of which is informative. The lod score is

$$\log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{2^{s-1} - (2^{s-1}-1)(p_g+p_t) + p_g p_t\{2^{s-1} - 2 + \theta_1^s + (1-\theta_1)^s\}}{2^{s-1} - (2^{s-1}-1)(p_g+p_t) + p_g p_t\{2^{s-1} - 2 + (1/2)^{s-1}\}} .$$

The scoring system for the doubtful families may easily be extended to the analogues of the z_2, z_3, and z_4 scoring types. However, the application of these scores is quite tedious in the absence of ancillary tables for each of the common test factors and, more important, the doubtful families have in practice been found to contribute relatively little information on linkage. Finney found in one example that scoring doubtful families for the ABO locus increased the available amount of information by only 5%, and he advised that "for a preliminary investigation of a linkage, scoring may well be confined to the certain families" (Finney, 1940). This policy, besides reducing the labor in linkage detection, has the further advantage of making linkage tests independent of the mating system and the population gene frequencies. Unless the data are extremely valuable, it seems best to score only the certain families, using where necessary the correction factors of §§10–11.

13. ONE OR BOTH PARENTS NOT DIRECTLY TESTED. COMPLETE PENETRANCE, NO NATURAL SELECTION

The extraction of information from untested parents by the method of u scores involves considerable algebraic manipulation and heavy arithmetic. Finney (1941b) has treated a few special cases and Smith (1953) has suggested an approximation for use in large samples. Fortunately the probability ratio method is so simple that *ad hoc* computation is always feasible, although the calculations are still tedious.

Suppose first that all ascertained families with untested parents are to be analysed, subject to the condition that families are sampled through the parents for both characters or that they are sampled through the parents for one character and the parental genotypes for the other character are known. On these assumptions the method of ascertainment does not affect the calculation, which consists in enumerating all parental genotypes which could give rise to F, the family in question, and then computing from the population gene frequencies and the assumption of random mating the prior probabilities of the mating types, say $P(M_1)$, $P(M_2)$, ⋯ etc. The conditional probabilities, $P(F \mid M_1)$, $P(F \mid M_2)$, ⋯ etc. are then calculated. Finally, the score for linkage is computed as

$$\log \frac{f(y;\theta_1)}{f(y;1/2)} = \log \frac{\sum_i P(M_i)P(F \mid M_i,\theta_1)}{\sum_i P(M_i)P(F \mid M_i,1/2)}$$

which of course is zero if none of the conditional probabilities is a function of θ.

These calculations are straightforward but time-consuming, and the investigator of human linkage would be well-advised to test both parents whenever possible. Full information cannot be recovered from incomplete records, although large families, whose scores are dominated by the conditional probabilities, are nearly as informative as if both parents had been tested. If instances of incomplete parental testing are not too common, no great amount of information will be lost by rejecting families with incomplete parental records. Alternatively, the scoring of incomplete records may be restricted to families whose parental genotypes can be inferred with certainty. In this case the linkage test is independent of gene frequencies and the mating structure of the population, considerable labor is saved, and at least some large families with only one tested parent will be included in the analysis. The score for the families whose parental genotypes are inferred is $z + C$, where z is the score appropriate to complete selection with both parents tested and C is a correction factor dependent on the method of sampling and inference. There are many special cases for C, all of which are easily treated *ad hoc* by the elementary methods used in §§10–11.

14. NATURAL SELECTION AND INCOMPLETE PENETRANCE

Genetic main factors with incomplete penetrance or low viability may still be used for linkage studies if we assume that the test factor is fully penetrant, viable, sampled at random through the parents or through complete selection of affected children, and that the viability and penetrance of the main factor are independent of the test factor.

For example, suppose the main factor is fully penetrant but so subvital that many affected progeny die before examination. On the above assumptions, it is still proper to test linkage by the methods of §§9 and 11, and the probabilities of Type I and Type II errors remain unaltered. Notice that no assumption need be made about the constancy of viability among families, either in the detection or estimation of linkage.

Again, suppose that the main gene is incompletely penetrant, with no assumptions made about viability or ascertainment. We shall assume that the main factor is so rare that all matings will be backcrosses if the main factor is a rare "dominant" or intercrosses if the main factor is a rare recessive. Given the above conditions on the test factor, the probability of a Type I error when the methods of §§9 and 11 are used will not be changed, regardless of whether penetrance is variable or not, but the power of the test will decrease very greatly when penetrance is low. In this case estimation of the penetrance will improve the power of the test, without affecting the probability of a Type I error.

In practice, the distinction between loose linkage to the main factor and linkage to viability or penetrance modifiers may be difficult to make, and therefore only tests of close linkage have much value when viability or penetrance is irregular. Even with such tests the rigorous justification of the assumption that the test factor does not influence the viability or penetrance of the main factor is extremely difficult, and may well be attempted only for tests which indicate a significant "linkage". Proof that the main and test factors are distributed independently in the general population, the absence of a correlation between the test phenotype of affected parents and affected progeny, constant penetrance, and homogeneity of the linkage value give supporting evidence for the hypothesis of linkage, while contrary observations suggest alternative explanations. Knowledge of the exact method of ascertainment is helpful in detecting irregularities, especially with rare recessive factors. All these problems are particularly acute when the test factor is extremely complex, and great difficulties have been encountered in attempts to distinguish linkage when sex is used as the test factor (Harris, 1948; Mohr, 1954). Even with less fundamental test traits, a significant "linkage" effect requires special scrutiny when the penetrance or viability of the main factor is low. If the test factor also behaves irregularly, the difficulties in linkage detection are vastly increased.

15. THE COMBINATION OF DATA

In §§5–6 the properties of the sequential probability ratio test were illustrated on the simplifying assumption that the data consist entirely of double backcross sibships of size 2, and it was shown that for this case the sequential test is very much superior to alternative procedures. In practice, linkage data in man comprise a mixture of family sizes and mating types, the frequencies of which vary among pairs of loci and are usually unspecified. We shall now show that this ignorance does not affect the important properties of the sequential test.

Let $k = 1, 2, \cdots$, denote a particular mating type and family size, $f_k(y;\theta)$ be the conditional distribution for the k^{th} type of data, and p_k be the prior probability of this type of data. Consider only sampling procedures for which p_k and $f_k(y;\theta)$ are independent of the stage of sampling. Then clearly the distribution $p_k f_k(y;\theta)$ is of the stationary type treated by Wald and all the important results of his sequential theory apply. In particular, it has been shown that of all tests with the same risk of error (α, β), the sequential probability ratio test requires on the average fewest observations, and that the Type I and Type II risks are approximately

$$\alpha = \frac{1 - B}{A - B}$$

$$\beta = \frac{B(A - 1)}{A - B},$$

these approximations being very good when the excess of $\sum z$ over the boundary log A or log B is negligible. This condition is satisfied if $| E(z) |$ and the standard deviation σ_z of z are sufficiently small, as in practice they usually will be. In any case the optimum character of the sequential test holds exactly (Wald and Wolfowitz, 1948).

Although the existence of a stationary distribution $p_k f_k(y;\theta)$ is sufficient for the proof of the above remarks, it is not necessary that the p_k be known to carry out the test. For the p_k are independent of θ, and therefore the probability ratio

$$\prod \frac{p_k f_k(y;\theta_1)}{p_k f_k(y;\theta_0)}$$

is identical with the ratio

$$\prod \frac{f_k(y;\theta_1)}{f_k(y;\theta_0)}.$$

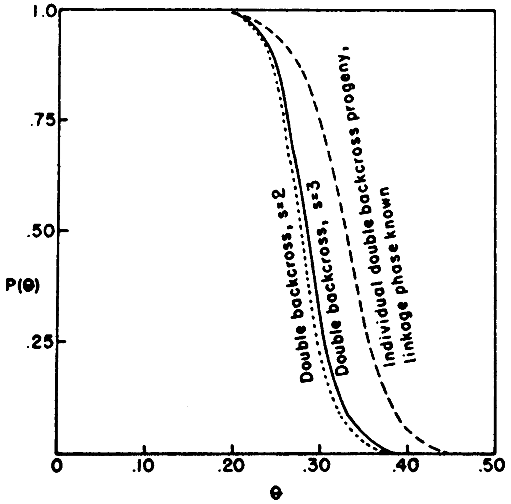

FIG. 5. The power function P(θ) for different types of data. A = 1000, B = .01, θ_1 = .20.

Determination of the p_k is necessary only if it is desired to find the power function and average sample number function of a sequential test, but this is of secondary importance so long as there is some basis for the choice of a particular test and we know that the sequential test on the average leads to a saving in the number of observations.

To choose a sequential test, it is convenient to have a rough notion of the average power of alternative tests. The power function depends on the distribution p_k, but the risks (α, β) do not, and this limits the possible fluctuation of the power function. Figure 5 shows a typical power function for three different types of data. The power function and the average power do not seem to be so highly variable as to jeopardize the control over Type I errors demanded for the idealized case in §5. In particular, it still seems appropriate to choose an unusually small value of α, of the order of .001.

The choice of θ_1 for a sequential test is largely determined by the average sample number on the null hypothesis, since (1) for randomly chosen loci the null hypothesis will usually be true and (2) the number of observations that can be tolerated is not narrowly bounded, so that random excesses over the expected number will usually not be a serious annoyance. A rough correspondence between expected sample number and amount of information may be established as follows.

Let n be the number of families required to terminate the test in mixed data and n_k be the number of families required for the test in data entirely of the k^{th} type. Let E(z) denote the expected value of z in mixed data and $E(z_k)$ the expected value of z in the k^{th} type of data. Also let c be a fixed value of k. Then on the null hypothesis

$$E(n)E(z) = \alpha \log A + (1 - \alpha) \log B = E(n_c)E(z_c)$$

and

$$E(n_c) = E\left\{\sum_{i=1}^{E(n)} \left[\frac{E(z_k)}{E(z_c)}\right]_i\right\},$$

TABLE 9.—THE EFFICIENCY OF DIFFERENT TYPES OF DATA IN DOUBLE BACKCROSS SIB-PAIR EQUIVALENTS
$\theta = \frac{1}{2}$

Scoring Type	θ_1 .05	.20	.40	u score information $\theta_1 \rightarrow \frac{1}{2}$
A. Families of size s, phase unknown				
z_1, s = 2	1.0	1.0	1.0	1.0
z_2, s = 2	.1	.1	.1	.2
z_3, s = 2	.1	.1	.1	.2
z_1, s = 5	3.8	5.3	8.8	10.0
z_3, s = 5	.5	.5	.5	.8
z_1, s = 10	9.7	14.7	33.1	45.0
B. Single progeny, phase known				
double backcross	1.6	3.2	25.4	—
single backcross	.5	1.0	8.5	—
double intercross, coupling, both factors dominant	1.0	1.8	12.3	—

where i = 1, 2, ···, E(n) denotes successive observations from the distribution $p_k f_k(y;\theta)$. If we let c designate double backcross sibships of size 2, then the ratio $E(z_k)/E(z_c)$ may be called the *double backcross sib-pair equivalent* on the null hypothesis. It has the property that if $E(n_c)$ is the average number of double backcross sib-pairs required by a certain test when $\theta = \theta_0 = 1/2$, then $E(n_c)E(z_c)/E(z_k)$ is the average number of families of type k required for the same test, assuming in both cases that the excess over the boundaries at the termination of the test can be neglected. Furthermore, for small families $E(z_k)/E(z_c)$ is of the same order as the information weight k in Finney's (1940) system of u scores (table 9). It follows that if S is the number of units of u score information that can be obtained with "reasonable" effort, then S is an estimate of $\sum E(z_k)/E(z_c)$ and $E(n_c)$ also, and this correspondence may serve as a rough guide in the selection of a sequential test. If S is about 10, θ_1 should be chosen to be .05, since $E(n_c) = 9$ for $\theta_1 = .05$. Similarly, if S is about 70, θ_1 should be taken as .20, if S is as much as 350, θ_1 may be .30, and only if S is about 6000 should θ_1 be .40. For linkage of two common test factors (ABO, Rh, MN), S may be as much as 6000, and for two less common test factors (Le, Lu, P, Fy blood groups), S may be 350. In most other cases S is probably smaller than 100, and θ_1 should be chosen accordingly. If it turns out that S has been considerably underestimated, a second test with a larger value of θ_1 will not increase α beyond tolerable limits.

The restriction of the sampling procedure to stationary distributions has proscribed a valid sampling method that in some respects seems desirable. All types of data might be collected at the beginning of sampling and whenever linkage is suggested, but when there is no suggestion of linkage it would seem economical to investigate only highly informative families for which the double backcross sib-pair equivalent is large. This makes p_k dependent on $\sum z$, but $f_k(y;\theta)$ is not affected and the probability is still one that the procedure will eventually terminate. It is of course essential that data be reported without regard for whether they indicate linkage or not. Wald (1947) has shown that the postulated kind of dependence does

TABLE 10

z_1

s	a + d	b + c	θ_1				
			.05	.10	.20	.30	.40
2	0	2	.2577	.2148	.1335	.0645	.0170
	1	1	−.7212	−.4437	−.1938	−.0757	−.0177
	2	0	.2577	.2148	.1335	.0645	.0170
3	0	3	.5353	.4654	.3181	.1703	.0492
	1	2	−.7212	−.4437	−.1938	−.0757	−.0177
	2	1	−.7212	−.4437	−.1938	−.0757	−.0177
	3	0	.5353	.4654	.3181	.1703	.0492
4	0	4	.8140	.7201	.5171	.2979	.0940
	1	3	−.4636	−.2289	−.0603	−.0113	−.0007
	2	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	3	1	−.4636	−.2289	−.0603	−.0113	−.0007
	4	0	.8140	.7201	.5171	.2979	.0940
5	0	5	1.0927	.9753	.7200	.4358	.1486
	1	4	−.1860	.0217	.1242	.0945	.0315
	2	3	−1.4425	−.8874	−.3876	−.1514	−.0355
	3	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	4	1	−.1860	.0217	.1242	.0945	.0315
	5	0	1.0927	.9753	.7200	.4358	.1486
6	0	6	1.3715	1.2306	.9238	.5784	.2106
	1	5	.0927	.2764	.3233	.2222	.0763
	2	4	−1.1848	−.6726	−.2541	−.0870	−.0184
	3	3	−2.1637	−1.3311	−.5815	−.2272	−.0532
	4	2	−1.1848	−.6726	−.2541	−.0870	−.0184
	5	1	.0927	.2764	.3233	.2222	.0763
	6	0	1.3715	1.2306	.9238	.5784	.2106
7	0	7	1.6502	1.4859	1.1278	.7230	.2779
	1	6	.3715	.5316	.5262	.3601	.1309
	2	5	−.9072	−.4220	−.0696	.0188	.0138
	3	4	−2.1637	−1.3311	−.5815	−.2272	−.0532
	4	3	−2.1637	−1.3311	−.5815	−.2272	−.0532
	5	2	−.9072	−.4220	−.0696	.0188	.0138
	6	1	.3715	.5316	.5262	.3601	.1309
	7	0	1.6502	1.4859	1.1278	.7230	.2779

TABLE 14

e_1

s	s_1	s_2	θ_1				
			.05	.10	.20	.30	.40
2	0	2	.1367	.1042	.0555	.0238	.0058
	1	1	−.1038	−.0840	−.0492	−.0226	−.0058
	2	0	.1367	.1042	.0555	.0238	.0058
3	0	3	.1852	.1392	.0728	.0309	.0075
	1	2	−.0476	−.0380	−.0218	−.0098	−.0025
	2	1	−.0476	−.0380	−.0218	−.0098	−.0025
	3	0	.1852	.1392	.0728	.0309	.0075
4	0	4	.1991	.1447	.0719	.0295	.0071
	1	3	−.0186	−.0117	−.0037	−.0007	0
	2	2	−.0270	−.0245	−.0168	−.0084	−.0023
	3	1	−.0186	−.0117	−.0037	−.0007	0
	4	0	.1991	.1447	.0719	.0295	.0071
5	0	5	.1987	.1382	.0640	.0249	.0058
	1	4	−.0049	.0007	.0047	.0034	.0011
	2	3	−.0133	−.0120	−.0082	−.0041	−.0011
	3	2	−.0133	−.0120	−.0082	−.0041	−.0011
	4	1	−.0049	.0007	.0047	.0034	.0011
	5	0	.1987	.1382	.0640	.0249	.0058
6	0	6	.1921	.1273	.0542	.0197	.0043
	1	5	.0016	.0062	.0077	.0046	.0013
	2	4	−.0064	−.0054	−.0030	−.0012	−.0003
	3	3	−.0068	−.0065	−.0051	−.0028	−.0008
	4	2	−.0064	−.0054	−.0030	−.0012	−.0003
	5	1	.0016	.0062	.0077	.0046	.0013
	6	0	.1921	.1273	.0542	.0197	.0043
7	0	7	.1831	.1153	.0447	.0149	.0031
	1	6	.0046	.0083	.0081	.0044	.0012
	2	5	−.0030	−.0021	−.0005	.0002	.0001
	3	4	−.0034	−.0032	−.0025	−.0014	−.0004
	4	3	−.0034	−.0032	−.0025	−.0014	−.0004
	5	2	−.0030	−.0021	−.0005	.0002	.0001
	6	1	.0046	.0083	.0081	.0044	.0012
	7	0	.1831	.1153	.0447	.0149	.0031

TABLE 11

z_2

s	a	b	c	d	θ_1				
					.05	.10	.20	.30	.40
2	2	0	0	0	.0374	.0298	.0170	.0077	.0019
	1	1	0	0	−.1367	−.1042	−.0555	−.0238	−.0058
	1	0	1	0	−.0410	−.0320	−.0177	−.0078	−.0019
	1	0	0	1	.1038	.0840	.0492	.0226	.0058
	0	2	0	0	.2577	.2148	.1335	.0645	.0170
	0	1	1	0	.1038	.0840	.0492	.0226	.0058
	0	1	0	1	−.7212	−.4437	−.1938	−.0757	−.0177
	0	0	2	0	.0374	.0298	.0170	.0077	.0019
	0	0	1	1	−.1367	−.1042	−.0555	−.0238	−.0058
	0	0	0	2	.2577	.2148	.1335	.0645	.0170
3	3	0	0	0	.1038	.0840	.0492	.0226	.0058
	2	1	0	0	−.2596	−.1908	−.0969	−.0404	−.0098
	2	0	1	0	−.0410	−.0320	−.0177	−.0078	−.0019
	2	0	0	1	.2122	.1754	.1072	.0509	.0133
	1	2	0	0	.1038	.0840	.0492	.0226	.0058
	1	1	1	0	−.0410	−.0320	−.0177	−.0078	−.0019
	1	1	0	1	−.7212	−.4437	−.1938	−.0757	−.0177
	1	0	2	0	−.0410	−.0320	−.0177	−.0078	−.0019
	1	0	1	1	−.0410	−.0320	−.0177	−.0078	−.0019
	1	0	0	2	.3711	.3153	.2041	.1027	.0280
	0	3	0	0	.5353	.4654	.3181	.1703	.0492
	0	2	1	0	.3711	.3153	.2041	.1027	.0280
	0	2	0	1	−.7212	−.4437	−.1938	−.0757	−.0177
	0	1	2	0	.2122	.1754	.1072	.0509	.0133
	0	1	1	1	−.7212	−.4437	−.1938	−.0757	−.0177
	0	1	0	2	−.7212	−.4437	−.1938	−.0757	−.0177
	0	0	3	0	.1038	.0840	.0492	.0226	.0058
	0	0	2	1	−.2596	−.1908	−.0969	−.0404	−.0098
	0	0	1	2	.1038	.0840	.0492	.0226	.0058
	0	0	0	3	.5353	.4654	.3181	.1703	.0492
4	4	0	0	0	.1898	.1559	.0940	.0441	.0114
	3	1	0	0	−.3608	−.2532	−.1219	−.0494	−.0118
	3	0	1	0	−.0035	−.0022	−.0007	−.0001	0
	3	0	0	1	.3231	.2715	.1717	.0843	.0226
	2	2	0	0	−.0492	−.0442	−.0295	−.0144	−.0038
	2	1	1	0	−.1776	−.1362	−.0732	−.0316	−.0078
	2	1	0	1	−.6838	−.4139	−.1768	−.0681	−.0158
	2	0	2	0	−.0819	−.0641	−.0355	−.0156	−.0039
	2	0	1	1	.0628	.0519	.0315	.0148	.0038
	2	0	0	2	.4847	.4166	.2775	.1442	.0406

TABLE 11.—*Continued*

s	a	b	c	d	θ_1				
					.05	.10	.20	.30	.40
	1	3	0	0	.3804	.3311	.2245	.1178	.0332
	1	2	1	0	.2167	.1828	.1158	.0567	.0151
	1	2	0	1	−.8579	−.5479	−.2493	−.0995	−.0236
	1	1	2	0	.0628	.0519	.0315	.0148	.0038
	1	1	1	1	−.7622	−.4757	−.2115	−.0835	−.0197
	1	1	0	2	−.6174	−.3597	−.1446	−.0532	−.0120
	1	0	3	0	−.0035	−.0022	−.0007	−.0001	0
	1	0	2	1	−.1776	−.1362	−.0732	−.0316	−.0078
	1	0	1	2	.2167	.1828	.1158	.0567	.0151
	1	0	0	3	.6492	.5678	.3950	.2171	.0647
	0	4	0	0	.8140	.7201	.5171	.2979	.0940
	0	3	1	0	.6492	.5678	.3950	.2171	.0647
	0	3	0	1	−.4636	−.2289	−.0603	−.0113	−.0007
	0	2	2	0	.4847	.4166	.2775	.1442	.0406
	0	2	1	1	−.6174	−.3597	−.1446	−.0532	−.0120
	0	2	0	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	0	1	3	0	.3231	.2715	.1717	.0843	.0226
	0	1	2	1	−.6838	−.4139	−.1768	−.0681	−.0158
	0	1	1	2	−.8579	−.5479	−.2493	−.0995	−.0236
	0	1	0	3	−.4636	−.2289	−.0603	−.0113	−.0007
	0	0	4	0	.1898	.1559	.0940	.0441	.0114
	0	0	3	1	−.3608	−.2532	−.1219	−.0494	−.0118
	0	0	2	2	−.0492	−.0442	−.0295	−.0144	−.0038
	0	0	1	3	.3804	.3311	.2245	.1178	.0332
	0	0	0	4	.8140	.7201	.5171	.2979	.0940
5	5	0	0	0	.2879	.2396	.1486	.0716	.0189
	4	1	0	0	−.4307	−.2859	−.1294	−.0507	−.0118
	4	0	1	0	.0628	.0519	.0315	.0148	.0038
	4	0	0	1	.4354	.3703	.2407	.1219	.0335
	3	2	0	0	−.2006	−.1678	−.1004	−.0458	−.0116
	3	1	1	0	−.3006	−.2229	−.1146	−.0482	−.0117
	3	1	0	1	−.6174	−.3597	−.1446	−.0532	−.0120
	3	0	2	0	−.0819	−.0641	−.0355	−.0156	−.0039
	3	0	1	1	.1712	.1434	.0895	.0431	.0114
	3	0	0	2	.5985	.5185	.3527	.1886	.0546
	2	3	0	0	.2256	.1972	.1325	.0679	.0187
	2	2	1	0	.0628	.0519	.0315	.0148	.0038
	2	2	0	1	−.9809	−.6345	−.2907	−.1161	−.0275
	2	1	2	0	−.0819	−.0641	−.0355	−.0156	−.0039
	2	1	1	1	−.7622	−.4757	−.2115	−.0835	−.0197

Table 11.—*Concluded*

s	a	b	c	d	θ_l				
					.05	.10	.20	.30	.40
	2	1	0	2	−.5091	−.2683	−.0866	−.0248	−.0044
	2	0	3	0	−.0819	−.0641	−.0355	−.0156	−.0039
	2	0	2	1	−.0819	−.0641	−.0355	−.0156	−.0039
	2	0	1	2	.3301	.2832	.1864	.0949	.0261
	2	0	0	3	.7631	.6703	.4727	.2656	.0814
	1	4	0	0	.6591	.5855	.4211	.2401	.0741
	1	3	1	0	.4943	.4333	.3003	.1625	.0473
	1	3	0	1	−.6174	−.3597	−.1446	−.0532	−.0120
	1	2	2	0	.3301	.2832	.1864	.0949	.0261
	1	2	1	1	−.7622	−.4757	−.2115	−.0835	−.0197
	1	2	0	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	1	1	3	0	.1712	.1434	.0895	.0431	.0114
	1	1	2	1	−.7622	−.4757	−.2115	−.0835	−.0197
	1	1	1	2	−.7622	−.4757	−.2115	−.0835	−.0197
	1	1	0	3	−.3502	−.1284	.0103	.0269	.0103
	1	0	4	0	.0628	.0519	.0315	.0148	.0038
	1	0	3	1	−.3006	−.2229	−.1146	−.0482	−.0117
	1	0	2	2	.0628	.0519	.0315	.0148	.0038
	1	0	1	3	.4943	.4333	.3003	.1625	.0473
	1	0	0	4	.9279	.8228	.5958	.3489	.1130
	0	5	0	0	1.0927	.9753	.7200	.4358	.1486
	0	4	1	0	.9279	.8228	.5958	.3489	.1130
	0	4	0	1	−.1860	.0217	.1242	.0945	.0315
	0	3	2	0	.7631	.6703	.4727	.2656	.0814
	0	3	1	1	−.3502	−.1284	.0103	.0269	.0103
	0	3	0	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	0	2	3	0	.5985	.5185	.3527	.1886	.0546
	0	2	2	1	−.5091	−.2683	−.0866	−.0248	−.0044
	0	2	1	2	−1.4425	−.8874	−.3876	−.1514	−.0355
	0	2	0	3	−1.4425	−.8874	−.3876	−.1514	−.0355
	0	1	4	0	.4354	.3703	.2407	.1219	.0335
	0	1	3	1	−.6174	−.3597	−.1446	−.0532	−.0120
	0	1	2	2	−.9809	−.6345	−.2907	−.1161	−.0275
	0	1	1	3	−.6174	−.3597	−.1446	−.0532	−.0120
	0	1	0	4	−.1860	.0217	.1242	.0945	.0315
	0	0	5	0	.2879	.2396	.1486	.0716	.0189
	0	0	4	1	−.4307	−.2859	−.1294	−.0507	−.0018
	0	0	3	2	−.2006	−.1678	−.1004	−.0458	−.0116
	0	0	2	3	.2256	.1972	.1325	.0679	.0187
	0	0	1	4	.6591	.5855	.4211	.2401	.0741
	0	0	0	5	1.0927	.9753	.7200	.4358	.1486

Table 12

z_3

s	a	b+c	d	θ_1				
				.05	.10	.20	.30	.40
2	2	0	0	.0120	.0090	.0045	.0018	.0004
	1	1	0	−.0382	−.0281	−.0139	−.0056	−.0013
	1	0	1	.0979	.0747	.0392	.0164	.0039
	0	2	0	.0979	.0747	.0392	.0164	.0039
	0	1	1	−.6174	−.3597	−.1446	−.0532	−.0120
	0	0	2	.5154	.4297	.2671	.1289	.0341
3	3	0	0	.0373	.0277	.0139	.0056	.0013
	2	1	0	−.0740	−.0528	−.0249	−.0096	−.0022
	2	0	1	.1993	.1543	.0824	.0346	.0083
	1	2	0	.0542	.0386	.0175	.0063	.0014
	1	1	1	−.5782	−.3270	−.1244	−.0435	−.0094
	1	0	2	.6252	.5235	.3273	.1582	.0417
	0	3	0	.2076	.1680	.0984	.0451	.0115
	0	2	1	−.7622	−.4757	−.2115	−.0835	−.0197
	0	1	2	−.3502	−.1284	.0103	.0269	.0103
	0	0	3	1.0706	.9308	.6361	.3405	.0984
4	4	0	0	.0763	.0568	.0283	.0114	.0026
	3	1	0	−.1064	−.0732	−.0325	−.0121	−.0027
	3	0	1	.3034	.2378	.1293	.0547	.0131
	2	2	0	.0108	.0034	−.0026	−.0025	−.0008
	2	1	1	−.5261	−.2848	−.0995	−.0319	−.0065
	2	0	2	.7353	.6180	.3891	.1888	.0498
	1	3	0	.1632	.1298	.0727	.0317	.0078
	1	2	1	−.7877	−.4859	−.2092	−.0801	−.0185
	1	1	2	−.2462	−.0439	.0608	.0509	.0166
	1	0	3	1.1811	1.0270	.7031	.3775	.1093
	0	4	0	.3187	.2657	.1676	.0831	.0225
	0	3	1	−.6937	−.4465	−.2188	−.0938	−.0232
	0	2	2	−1.0746	−.5775	−.1905	−.0539	−.0092
	0	1	3	.1856	.3389	.3347	.2058	.0640
	0	0	4	1.6280	1.4403	1.0343	.5958	.1880
5	5	0	0	.1286	.0961	.0480	.0191	.0044
	4	1	0	−.1343	−.0884	−.0365	−.0128	−.0027
	4	0	1	.4092	.3245	.1795	.0766	.0183
	3	2	0	−.0322	−.0307	−.0208	−.0100	−.0026
	3	1	1	−.4620	−.2335	−.0698	−.0185	−.0031
	3	0	2	.8456	.7130	.4522	.2206	.0582
	2	3	0	.1189	.0920	.0478	.0193	.0044
	2	2	1	−.8075	−.4900	−.2029	−.0750	−.0169
	2	1	2	−.1404	.0435	.1142	.0764	.0233
	2	0	3	1.2917	1.1233	.7706	.4152	.1206

TABLE 12.—*Continued*

s	a	b+c	d	θ_1				
				.05	.10	.20	.30	.40
	1	4	0	.2741	.2268	.1397	.0671	.0177
	1	3	1	−.7331	−.4741	−.2286	−.0958	−.0233
	1	2	2	−1.0107	−.5235	−.1555	−.0361	−.0042
	1	1	3	.2959	.4340	.3987	.2396	.0737
	1	0	4	1.7386	1.5367	1.1031	.6366	.2015
	0	5	0	.4301	.3649	.2422	.1278	.0366
	0	4	1	−.5931	−.3728	−.1905	−.0879	−.0228
	0	3	2	−1.3547	−.7977	−.3166	−.1123	−.0244
	0	2	3	−.6897	−.2365	.0540	.0851	.0334
	0	1	4	.7420	.8444	.7200	.4434	.1445
	0	0	5	2.1855	1.9507	1.4400	.8717	.2972
6	6	0	0	.1932	.1451	.0728	.0289	.0066
	5	1	0	−.1561	−.0971	−.0364	−.0117	−.0023
	5	0	1	.5165	.4135	.2327	.1002	.0240
	4	2	0	−.0746	−.0633	−.0368	−.0160	−.0039
	4	1	1	−.3875	−.1738	−.0356	−.0031	.0008
	4	0	2	.9559	.8084	.5163	.2536	.0671
	3	3	0	.0747	.0545	.0240	.0078	.0015
	3	2	1	−.8201	−.4867	−.1923	−.0681	−.0148
	3	1	2	−.0331	.1330	.1701	.1034	.0305
	3	0	3	1.4023	1.2196	.8383	.4535	.1322
	2	4	0	.2296	.1882	.1123	.0518	.0132
	2	3	1	−.7718	−.4999	−.2360	−.0964	−.0231
	2	2	2	−.9364	−.4615	−.1163	−.0165	.0012
	2	1	3	.4062	.5295	.4636	.2744	.0838
	2	0	4	1.8492	1.6332	1.1720	.6778	.2154
	1	5	0	.3855	.3257	.2128	.1098	.0308
	1	4	1	−.6342	−.4049	−.2072	−.0942	−.0242
	1	3	2	−1.3672	−.7915	−.3002	−.1008	−.0208
	1	2	3	−.5826	−.1466	.1117	.1147	.0419
	1	1	4	.8525	.9408	.7880	.4827	.1571
	1	0	5	2.2961	2.0472	1.5093	.9143	.3129
	0	6	0	.5418	.4651	.3200	.1776	.0536
	0	5	1	−.4891	−.2888	−.1441	−.0693	−.0188
	0	4	2	−1.3191	−.8168	−.3702	−.1488	−.0353
	0	3	3	−1.4833	−.7448	−.1870	−.0182	.0070
	0	2	4	−.1435	.2505	.4115	.2985	.1043
	0	1	5	1.2994	1.3544	1.1224	.7109	.2465
	0	0	6	2.7430	2.4612	1.8476	1.1568	.4212

Table 12.—*Concluded*

s	a	b+c	d	θ_1				
				.05	.10	.20	.30	.40
7	7	0	0	.2684	.2032	.1028	.0408	.0093
	6	1	0	−.1703	−.0981	−.0319	−.0087	−.0015
	6	0	1	.6247	.5044	.2884	.1255	.0302
	5	2	0	−.1162	−.0939	−.0504	−.0206	−.0048
	5	1	1	−.3043	−.1068	.0029	.0141	.0051
	5	0	2	1.0663	.9041	.5814	.2876	.0764
	4	3	0	.0306	.0175	.0014	−.0025	−.0011
	4	2	1	−.8237	−.4752	−.1774	−.0594	−.0124
	4	1	2	.0750	.2243	.2281	.1318	.0380
	4	0	3	1.5129	1.3160	.9064	.4925	.1442
	3	4	0	.1852	.1497	.0855	.0374	.0091
	3	3	1	−.8095	−.5233	−.2406	−.0954	−.0223
	3	2	2	−.8534	−.3925	−.0732	.0048	.0070
	3	1	3	.5166	.6252	.5293	.3101	.0942
	3	0	4	1.9597	1.7297	1.2410	.7193	.2296
	2	5	0	.3409	.2866	.1838	.0925	.0253
	2	4	1	−.6751	−.4365	−.2225	−.0994	−.0252
	2	3	2	−1.3708	−.7769	−.2793	−.0875	−.0167
	2	2	3	−.4745	−.0551	.1712	.1455	.0508
	2	1	4	.9631	1.0372	.8563	.5224	.1700
	2	0	5	2.4067	2.1437	1.5786	.9571	.3287
	1	6	0	.4970	.4254	.2896	.1581	.0469
	1	5	1	−.5303	−.3219	−.1642	−.0790	−.0213
	1	4	2	−1.3565	−.8385	−.3696	−.1432	−.0331
	1	3	3	−1.4009	−.6749	−.1409	.0063	.0142
	1	2	4	−.0331	.3463	.4777	.3355	.1158
	1	1	5	1.4100	1.4509	1.1914	.7528	.2614
	1	0	6	2.8536	2.5577	1.9170	1.2003	.4384
	0	7	0	.6537	.5660	.4004	.2315	.0732
	0	6	1	−.3846	−.2024	−.0886	−.0413	−.0112
	0	5	2	−1.2229	−.7576	−.3720	−.1660	−.0422
	0	4	3	−1.9107	−1.0674	−.3649	−.1010	−.0153
	0	3	4	−1.0240	−.3345	.1158	.1639	.0677
	0	2	5	.4134	.7584	.8062	.5536	.1985
	0	1	6	1.8569	1.8649	1.5291	.9923	.3651
	0	0	7	3.3005	2.9718	2.2557	1.4460	.5559

TABLE 13

s	c_1					c_2					c_3				
	θ_1					θ_1					θ_1				
	.05	.10	.20	.30	.40	.05	.10	.20	.30	.40	.05	.10	.20	.30	.40
2	−.0374	−.0298	−.0170	−.0077	−.0019	−.0164	−.0130	−.0074	−.0033	−.0008	−.0197	−.0147	−.0075	−.0031	−.0007
3	−.0210	−.0167	−.0095	−.0042	−.0011	−.0121	−.0096	−.0054	−.0024	−.0006	−.0203	−.0150	−.0075	−.0030	−.0007
4	−.0105	−.0081	−.0044	−.0019	−.0005	−.0073	−.0057	−.0032	−.0014	−.0004	−.0175	−.0128	−.0062	−.0025	−.0006
5	−.0051	−.0038	−.0019	−.0008	−.0002	−.0041	−.0032	−.0018	−.0008	−.0002	−.0143	−.0103	−.0049	−.0019	−.0004
6	−.0025	−.0017	−.0008	−.0003	−.0001	−.0022	−.0017	−.0009	−.0004	−.0001	−.0113	−.0080	−.0037	−.0014	−.0003
7	−.0012	−.0008	−.0003	−.0001	0	−.0011	−.0009	−.0005	−.0002	0	−.0087	−.0061	−.0027	−.0010	−.0002
8	−.0006	−.0004	−.0001	0	0	−.0006	−.0004	−.0002	−.0001	0	−.0067	−.0046	−.0020	−.0007	−.0002
9	−.0003	−.0002	−.0001	0	0	−.0003	−.0002	−.0001	0	0	−.0050	−.0034	−.0014	−.0005	−.0001
10	−.0001	−.0001	0	0	0	−.0001	−.0001	−.0001	0	0	−.0038	−.0025	−.0010	−.0003	−.0001
11	−.0001	0	0	0	0	−.0001	−.0001	0	0	0	−.0028	−.0018	−.0007	−.0002	0
12	0	0	0	0	0	0	0	0	0	0	−.0021	−.0013	−.0005	−.0002	0
13											−.0016	−.0010	−.0003	−.0001	0
14											−.0011	−.0007	−.0002	−.0001	0
15											−.0008	−.0005	−.0002	0	0

not affect the validity of a sequential test, but his proof of the optimum character of the sequential test does not cover dependent observations. I suspect, but have not been able to prove, that the sequential probability ratio test is optimum for this class of dependence also.

The ease and exactness with which probability ratio scores may be combined is particularly important when the data are of mixed known and unknown phase, since the alternative u score theory provides only a rough approximation in small samples (Finney, 1943; Smith, 1953). This is a critical point, not only for human pedigrees, but especially in laboratory vertebrates where linkage studies are of secondary interest and the material on any particular pair of loci is usually heterogeneous and small.

16. INSTRUCTIONS FOR ANALYSIS

Although the simplicity of the sequential probability ratio test allows the investigator to modify his methods to fit particular situations, it may be useful to set down here instructions for the routine case of unrelated families, tested parents, known parental genotypes, and unknown phase.

Step 1. Define the method of selection. This comprehends both ascertainment of families and rejection of some kinds of ascertained families. Usually, families with untested parents or of doubtful mating type will be rejected; otherwise, cf. §§12–13. For each factor selection may be complete, truncate, or arbitrary (§7). With respect to the two factors in a linkage test, there are three important methods of selection:

(i) Complete selection of one or both factors.

(ii) Truncate selection of both factors.

(iii) Arbitrary selection of one factor (G), truncate selection of the other (T).

Step 2. Choose the alternative hypothesis (cf. §15). If the amount of data that can be obtained with "reasonable" effort is likely to be small, choose $\theta_1 = .05$ or $.10$; if a moderately large amount of data is hoped for, choose $\theta_1 = .20$ or $.30$; if an extraordinarily large amount is anticipated, take $\theta_1 = .40$. Usually, $\log B = -2$ and $\log A = 3$ are appropriate choices for the other parameters of the test.

Step 3. Classify the mating type of each family according to tables 4–8, and distribute the children among classes a, b, c, d, $\cdots$. In these tables, G_1, G_2 and T_1, T_2 denote factors without dominance or rare "dominants", while G, g and T, t are factors showing simple dominant-recessive relationships.

Step 4. Determine the score for each family from tables 10–18, or compute directly, using common logarithms. The following outline may be helpful in performing the above steps.

Classification of matings, methods of selection, and scores (z)

I. Double backcross, and single backcross with no dominance in the intercross factor.

(i) Complete selection of either factor	z_1
(ii) Truncate selection of both factors	$z_1 + c_1$
(iii) Arbitrary-truncate selection	$z_1 + e_1$

			TABLE 15					TABLE 16					TABLE 17				
			e_2					d_2					e_3				
s	s_1	s_2	θ_1					θ_1					θ_1				
			.05	.10	.20	.30	.40	.05	.10	.20	.30	.40	.05	.10	.20	.30	.40
2	0	2	.1367	.1042	.0555	.0238	.0058	.0534	.0416	.0229	.0100	.0025	.1708	.1200	.0562	.0220	.0051
	1	1	−.0374	−.0298	−.0170	−.0077	−.0019	−.0476	−.0380	−.0218	−.0098	−.0025	−.0447	−.0336	−.0172	−.0071	--.0017
	2	0	.0132	.0104	.0058	.0026	.0006	.0534	.0416	.0229	.0100	.0025	.0160	.0118	.0059	.0024	.0006
3	0	3	.1852	.1392	.0728	.0309	.0075	.0953	.0735	.0398	.0172	.0042	.2575	.1835	.0892	.0362	.0086
	1	2	.0171	.0134	.0075	.0033	.0008	−.0276	−.0220	−.0125	−.0056	−.0014	.0443	.0305	.0132	.0047	.0010
	2	1	−.0271	−.0215	−.0122	−.0055	−.0014	−.0276	−.0220	−.0125	−.0056	−.0014	−.0470	−.0348	−.0173	−.0069	−.0016
	3	0	.0171	.0134	.0075	.0033	.0008	.0953	.0735	.0398	.0172	.0042	.0294	.0214	.0104	.0041	.0010
4	0	4	.1991	.1447	.0719	.0295	.0071	.1271	.0968	.0515	.0221	.0054	.3022	.2150	.1068	.0446	.0108
	1	3	.0426	.0344	.0201	.0091	.0023	−.0016	−.0010	−.0003	−.0001	0	.1030	.0764	.0383	.0155	.0036
	2	2	−.0031	−.0028	−.0019	−.0009	−.0003	−.0332	−.0267	−.0155	−.0070	−.0018	.0051	.0016	−.0012	−.0011	−.0004
	3	1	−.0160	−.0126	−.0070	−.0031	−.0008	−.0016	−.0010	−.0003	−.0001	0	−.0416	−.0301	−.0143	−.0055	−.0012
	4	0	.0162	.0127	.0071	.0031	.0008	.1271	.0968	.0515	.0221	.0054	.0404	.0290	.0138	.0054	.0012
5	0	5	.1987	.1382	.0640	.0249	.0058	.1504	.1130	.0590	.0249	.0061	.3247	.2290	.1148	.0489	.0120
	1	4	.0530	.0419	.0236	.0105	.0026	.0216	.0175	.0103	.0047	.0012	.1389	.1043	.0545	.0232	.0057
	2	3	.0097	.0081	.0050	.0024	.0006	−.0226	−.0182	−.0105	−.0047	−.0012	.0448	.0331	.0160	.0062	.0014
	3	2	−.0052	−.0045	−.0029	−.0014	−.0004	−.0226	−.0182	−.0105	−.0047	−.0012	−.0096	−.0091	−.0063	−.0031	−.0008
	4	1	−.0087	−.0067	−.0036	−.0015	−.0004	.0216	.0175	.0103	.0047	.0012	−.0345	−.0242	−.0108	−.0039	−.0008
	5	0	.0134	.0104	.0058	.0025	.0006	.1504	.1130	.0590	.0249	.0061	.0493	.0348	.0161	.0061	.0014
6	0	6	.1921	.1273	.0542	.0197	.0043	.1667	.1235	.0630	.0263	.0063	.3347	.2330	.1168	.0503	.0125
	1	5	.0564	.0429	.0226	.0096	.0023	.0402	.0322	.0184	.0083	.0021	.1608	.1202	.0639	.0279	.0070
	2	4	.0154	.0128	.0078	.0036	.0009	−.0091	−.0071	−.0039	−.0017	−.0004	.0710	.0542	.0287	.0121	.0029
	3	3	.0012	.0011	.0007	.0003	.0001	−.0226	−.0183	−.0107	−.0049	−.0012	.0180	.0128	.0054	.0017	.0003
	4	2	−.0038	−.0033	−.0021	−.0010	−.0003	−.0091	−.0071	−.0039	−.0017	−.0004	−.0146	−.0126	−.0076	−.0034	−.0008
	5	1	−.0045	−.0033	−.0017	−.0007	−.0002	.0402	.0322	.0184	.0083	.0021	−.0275	−.0184	−.0075	−.0025	−.0005
	6	0	.0103	.0080	.0043	.0019	.0005	.1667	.1235	.0630	.0263	.0063	.0562	.0390	.0175	.0065	.0015

7	0	7	.1831	.1153	.0447	.0149	.0031	.1776	.1294	.0645	.0265	0063	.3371	.2314	.1153	.0499	.0125
	1	6	.0564	.0411	.0200	.0079	.0018	.0545	.0431	.0242	.0107	.0027	.1738	.1286	.0686	.0304	.0077
	2	5	.0177	.0142	.0082	.0037	.0009	.0035	.0031	.0022	.0011	.0003	.0881	.0675	.0368	.0162	.0040
	3	4	.0042	.0037	.0024	.0012	.0003	−.0162	−.0131	−.0076	−.0035	−.0009	.0371	.0285	.0148	.0061	.0014
	4	3	−.0006	−.0005	−.0003	−.0002	0	−.0162	−.0131	−.0076	−.0035	−.0009	.0049	.0028	.0002	−.0004	−.0002
	5	2	−.0023	−.0020	−.0013	−.0006	−.0002	.0035	.0031	.0022	.0011	.0003	−.0154	−.0128	−.0073	−.0031	−.0007
	6	1	−.0022	−.0015	−.0007	−.0002	0	.0545	.0431	.0242	.0107	.0027	−.0212	−.0133	−.0047	−.0013	−.0002
	7	0	.0075	.0058	.0031	.0013	.0003	.1776	.1294	.0645	.0265	.0063	.0613	.0419	.0182	.0066	.0015

TABLE 18.—LOD SCORES FOR INDIVIDUAL PROGENY WHEN THE PARENTAL PHASE IS KNOWN

$\frac{f(y;\theta_1)}{f(y;\frac{1}{2})}$	θ_1				
	.05	.10	.20	.30	.40
$2\theta_1$	−1.0000	−.6990	−.3979	−.2218	−.0969
$2(1-\theta_1)$	.2788	.2553	.2041	.1461	.0792
$2(2-\theta_1)/3$	.1139	.1027	.0792	.0544	.0280
$2(1+\theta_1)/3$	−.1549	−.1347	−.0969	−.0621	−.0300
$4(3-2\theta_1+\theta_1^2)/9$	.1106	.0965	.0694	.0440	.0207
$4(2+\theta_1-\theta_1^2)/9$	−.0410	−.0320	−.0177	−.0078	−.0019
$4(1-\theta_1+\theta_1^2)/3$	.1038	.0840	.0492	.0226	.0058
$4(2+\theta_1^2)/9$	−.0506	−.0490	−.0426	−.0320	−.0177
$2(1+2\theta_1-2\theta_1^2)/3$	−.1367	−.1042	−.0555	−.0238	−.0058
$2(1-2\theta_1+2\theta_1^2)$	.2577	.2148	.1335	.0645	.0170

II. Single backcross with dominance in the intercross factor.
  (i) Complete selection of either factor z_2
  (ii) Truncate selection of both factors $z_2 + c_2$
  (iii) Arbitrary selection of intercross factor, truncate selection of backcross factor $z_2 + e_2$
  (iv) Arbitrary selection of backcross factor, truncate selection of intercross factor $z_2 + d_2$

III. Double intercross with dominance in both factors
  (i) Complete selection of either factor z_3
  (ii) Truncate selection of both factors $z_3 + c_3$
  (iii) Arbitrary-truncate selection $z_3 + e_3$

IV. Double intercross with dominance in one factor
  (i) Complete selection of either factor z_4
  (ii) Arbitrary selection of factor with no dominance, truncate selection of dominant factor $z_4 + e_4$

V. Double intercross with no dominance in either factor
  (i) Complete selection of either factor z_5

Step 5. Accumulate the family scores (z). If $\sum z \leq \log B$, conclude that the frequency of recombination θ is significantly greater than θ_1 on the assumptions of §1. If $\sum z \geq \log A$, conclude that θ is significantly less than 1/2. Review the data and assumptions before deciding that true linkage is present. If $\log B < \sum z < \log A$, suspend judgment about linkage until further data lead to a decision. More data can also be used to estimate θ, after linkage has been detected, or to make a further test for linkage in the range $\theta_1 < \theta < 1/2$, if that seems advisable.

The following examples illustrate the scoring procedure.

Case 1. A mating of type GT × gt gives 2GT, 2Gt, and 1gt progeny. This is a double backcross (mating 1) with $s = 5$, $a + d = 3$. The score for complete selection is z_1 (table 10). For truncate selection of both factors, add the correction factor c_1 (table 13), and for truncate selection of the T factor but arbitrary selection of the G factor (which shows 4G:1g) add e_1 with $s_1 = 4$, $s_2 = 1$ (table 14). For $\theta_1 = .20$, we find $z_1 = -.3876$, $z_1 + c_1 = -.3895$, and $z_1 + e_1 = -.3829$.

Case 2. A mating of type GT × Gt gives 5GT, 2gT, 3Gt, and 1gt progeny. This is a single backcross (mating 9) with s = 11, a = 5, b = 2, c = 3, and d = 1. Families of this size are not given in table 11, but the score may quickly be obtained by factoring the expression for z_2 which is

$$z_2 = \log \frac{2^{10}}{3^8} [(2 - \theta_1)^5\theta_1^2 (1 + \theta_1)^3(1 - \theta_1) + (1 + \theta_1)^5 (1 - \theta_1)^2(2 - \theta_1)^3\theta_1]$$

$$= 3 \log [2(2 - \theta_1)/3] + 3 \log [2(1 + \theta_1)/3] + \log 2\theta_1 + \log 2(1 - \theta_1)$$

$$+ \log \frac{2^2}{3^2} [(2 - \theta_1)^2\theta_1 + (1 + \theta_1)^2(1 - \theta_1)].$$

The first four terms correspond to progeny of known parental phase (table 18), the last term to a single backcross family with s = 3, a = 2, b = 1, c = d = 0. For $\theta_1 = .20$, we find

$$z_2 = 3(.0792) + 3(-.0969) + (-.3979) + .2041 + (-.0969) = -.3438.$$

The corresponding scores for incomplete selection are $z_2 + c_2 = -.3438$ and $z_2 + e_2 = -.3439$. Here, as is usual in large families, the corrections for incomplete selection are negligible.

17. SUMMARY

The sequential probability ratio test for linkage detection in man is simple, exact and efficient. The basic assumptions of the linkage test are discussed, and criteria are developed for the choice of parameters in the sequential test. For the case of double backcross sib-pairs, the sequential tests considered here require less than 1/3 as many observations for a given risk of error as the Fisher-Finney u score method and about 1/5 as many observations as the Haldane-Smith nonsequential probability ratio test. Formulae for "lod" scores are given for a variety of mating types and methods of selection, and the research worker should have no difficulty extending the formulae to novel cases as they arise. The optimum property of the sequential probability ratio test holds for mixed data, the combination of which is easy and exact. Examples and tables of scores are given for the most important mating types.

The work for this paper was done under the direction of Dr. J. F. Crow, to whom the author is indebted for many stimulating discussions and constant encouragement. Drs. E. R. Immel, W. J. Schull, and C. A. B. Smith read the preliminary manuscript and offered helpful comments. Thanks are also due to the Numerical Analysis Laboratory of the University of Wisconsin, and especially to Mr. William Graebel, for assistance in computing the tables of linkage scores.

REFERENCES

Bailey, N. T. J. 1951. A classification of methods of ascertainment and analysis in estimating the frequencies of recessives in man. *Ann. Eugen.* 16: 223–225.

Bernstein, F. 1931. Zur Grundlegung der Chromosomentheorie der Vererbung beim Menschen mit besondere Berücksichtung der Blutgruppen. *Z. indukt. Abstamm. u. VererbLehre* 57: 113–138.

Bridges, C. B., and K. S. Brehme 1944. The mutants of Drosophila melanogaster. *Carnegie Inst. Wash. Publ.* 552.

Bross, I. 1952. Sequential medical plans. *Biometrics* 8: 188–205.

Carter, T. C. 1955. The estimation of total genetical map lengths from linkage test data. *J. Gener.* 53: 21–28.

Crew, F. A. E., and P. Ch. Koller 1932. The sex incidence of chiasma frequency and genetical crossing-over in the mouse. *J. Genet.* 26: 359–384.

Finney, D. J. 1940. The detection of linkage. *Ann. Eugen.* 10: 171–214.

Finney, D. J. 1941a. The detection of linkage. II: Further mating types; scoring of Boyd's data. *Ann. Eugen.* 11: 10–30.

Finney, D. J. 1941b. The detection of linkage. III: Incomplete parental testing. *Ann. Eugen.* 11: 115–135.

Finney, D. J. 1942. The detection of linkage. VI: The loss of information from incompleteness of parental testing. *Ann. Eugen.* 11: 233–242.

Finney, D. J. 1943. The detection of linkage. VII: Combination of data from matings of known and unknown phase. *Ann. Eugen.* 12: 31–43.

Fisher, R. A. 1935. The detection of linkage with "dominant" abnormalities. *Ann. Eugen.* 6: 187–201.

Haldane, J. B. S. 1934. Methods for the detection of autosomal linkage in man. *Ann. Eugen.* 6: 26–65.

Haldane, J. B. S. 1946. The cumulants of the distribution of Fisher's "u_{11}" and "u_{31}" scores used in the detection and estimation of linkage in man. *Ann. Eugen.* 13: 122–134.

Haldane, J. B. S., and C. A. B. Smith 1947. A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Ann. Eugen.* 14: 10–31.

Harris, H. 1948. On sex limitation in human genetics. *Eugen. Rev.* 40: 70–76.

Hogben, L. 1934. The detection of linkage in human families. *Proc. Roy. Soc. B* 114: 340–363.

Kosambi, D. D. 1944. The estimation of map distances from recombination values. *Ann. Eugen.* 12: 172–175.

Mohr, J. 1954. A study of linkage in man. *Op. Dom. Biol. Hered. Hum. Univ. Hafn.* 33: 1–119.

Neel, J. V. 1949. The detection of the genetic carriers of hereditary disease. *Amer. J. Hum. Genet.* 1: 19–36.

Penrose, L. S. 1953. The general purpose sib-pair linkage test. *Ann. Eugen.* 18: 120–124.

Rhoades, M. M. 1950. Meiosis in maize. *J. Hered.* 41: 59–70.

Slizynski, B. M. 1949. A preliminary pachytene chromosome map of the house mouse. *J. Genet.* 49: 242–245.

Smith, C. A. B. 1953. The detection of linkage in human genetics. *J. Roy. Stat. Soc. B* 15: 153–192.

Wald, A. 1947. *Sequential Analysis.* New York: Wiley.

Wald, A., and J. Wolfowitz. 1948. Optimum character of the sequential probability ratio test. *Ann. Math. Stat.* 19: 326–339.