THE FACTORIAL ANALYSIS OF SMALL FAMILIES WITH PARENTS OF UNDETERMINED GENOTYPE.

By LANCELOT HOGBEN.

(*Department of Social Biology, University of London.*)

In testing the statistical requirements of genetic hypothesis, two methods are employed. One is to predict the expected number of individuals belonging to a given genotype or phenotype and to determine the variance of the predicted value from the classical theory of sampling. The other approaches the problem from the standpoint of inverse probability, and seeks to determine from the observed values the most likely value which would be found in the population from which the sample is derived. The second method, which has been developed by Fisher(1), is the treatment appropriate to the problem of linkage. The first is customarily applied to test for single-gene substitutions. In testing the data derived from human families for single-gene substitutions, a special difficulty arises from the limitation of the sampling process, when neither parent belongs to the same phenotype as the offspring. Owing to the small size of human families, it will often happen that parents of a given phenotype capable of producing offspring of a second phenotype do not in fact do so. In a previous communication(3) this problem has been discussed, and a statistical method for dealing with it has been set forth. In a later publication Haldane(2) has suggested an alternative method based on the method of maximum likelihood. Haldane approaches the issue as a problem of inverse probability. This avoids the necessity of an *a priori* assumption which is inherent in the method proposed by the present writer. The object of this communication is to show that the same result may be obtained without introducing any such assumption as that to which Prof. Haldane takes exception.

The following preliminary statement of the symbols employed will elucidate what follows. The *observed* number of recessives in a group of n_s s-membered families each containing *at least one* recessive are denoted t_s. The number *predicted* by factorial hypothesis will be denoted r_s. The probability that an offspring of two heterozygous parents will be recessive is p, and the probability that such an offspring will be normal

is q. For single gene substitutions $p = \frac{1}{4}$ and $q = \frac{3}{4}$. If the method of sampling includes only fraternities with at least one recessive

$$\frac{\Sigma r_s}{\Sigma n_s c_s} = p,$$

where

$$c_s = \frac{s}{1 - q^s}.$$

By the use of a theorem due to Sheppard it can be shown that the standard deviation of p determined in this way is

$$\frac{\sqrt{\Sigma n_s k_s}}{\Sigma n_s c_s}, \qquad \ldots\ldots(1)$$

where

$$k_s = \frac{s\,(1-q)}{(1-q^s)^2} \cdot \{sq^s\,(q-1) + q\,(1-q^s)\}.$$

The method proposed in the communication cited above(2) was to compare p with the ratio

$$\frac{\Sigma t_s}{\Sigma n_s c_s}, \qquad \ldots\ldots(2)$$

and the difference between p and (2) with the standard deviation of p (1) determined in this way. An objection to this procedure is that the quantity defined by (2) contains both observed quantities (t_s) and a function (c_s) of the theoretical expectation. This objection can be removed by framing the problem in somewhat different terms leading to precisely the same numerical result. The method then has one clear advantage over any attempt to treat the problem from the standpoint of inverse probability. Tables of the requisite functions already given can be used to determine all the information required without the labour of undertaking an approximate solution of a fresh set of equations, whenever a new set of data is tested.

For a single s-membered fraternity it will be shown below that the standard deviation of the incomplete binomial distribution arising from the exclusion of the class of families with zero recessives is $k_s^{\frac{1}{2}}$. The expected number of recessives is $p.c_s$. For a pool of families of different sizes the net expectation is $\Sigma n_s p c_s$, and since there is zero correlation between successive samples made on different families the standard deviation is $\sqrt{\Sigma n_s k_s}$. Thus we may write

$$\Sigma r_s = \Sigma n_s p c_s \pm \sqrt{\Sigma n_s k_s}. \qquad \ldots\ldots(3)$$

All the quantities in (3) are defined by hypothesis. There now seems to be no formal objection to comparing the observed value Σt_s with the

expected value Σr_s and their difference with the standard deviation of the expected value as defined by (3).

The ratio of this difference to the standard deviation of the expected value is

$$[\Sigma t_s - \Sigma r_s] : \sqrt{\Sigma n_s k_s}.$$

This is numerically equivalent to

$$\left[\frac{\Sigma t_s}{\Sigma n_s c_s} - p\right] : \frac{\sqrt{\Sigma n_s k_s}}{\Sigma n_s c_s}.$$

If therefore the foregoing analysis is valid any deductions made by the previous treatment are also valid.

If a sibship contains s members and has parents $(Rr \times Rr)$ who may have affected offspring, the frequencies of families with 0, 1, 2, ..., r recessives form the binomial series of which the general term is ${}^sC_r p^r q^{s-r}$. Proceeding by the usual method for determining the mean and standard deviation for the complete binomial we may make a table of frequencies (f) and occurrences (r). The mean is thus given by

$$\frac{\Sigma r(f)}{\Sigma (f)},$$

and the mean square deviation referred to zero occurrences as origin is

$$\frac{\Sigma r^2 (f)}{\Sigma (f)}.$$

If all families with no recessives are excluded $\Sigma (f) = 1 - q^s$. Hence the mean (M) is

$$\frac{\sum_1^s r\, {}^sC_r p^r q^{s-r}}{1 - q^s}.$$

The standard deviation referred to the mean as origin is given by

$$\sigma_s^2 = \frac{\sum_1^s r^2 . {}^sC_r p^r q^{s-r}}{1 - q^s} - \left[\frac{\sum_1^s r\, {}^sC_r p^r q^{s-r}}{1 - q^s}\right]^2. \qquad \ldots\ldots(4)$$

The reduction may be effected thus:

$$\frac{\partial}{\partial p} . \sum_1^s {}^sC_r p^r q^{s-r} = \frac{1}{p} \sum_1^s r\, {}^sC_r p^r q^{s-r}, \qquad \ldots\ldots(5)$$

$$\frac{\partial^2}{\partial p^2} . \sum_1^s {}^sC_r p^r q^{s-r} = \frac{1}{p^2} \sum_1^s r^2\, {}^sC_r p^r q^{s-r} - \frac{1}{p^2} \sum_1^s r\, {}^sC_r p^r q^{s-r}. \qquad \ldots\ldots(6)$$

Also

$$\frac{\partial}{\partial p} \sum_1^s {}^sC_r p^r q^{s-r} = \frac{\partial}{\partial p}[(p+q)^s - q^s]$$

$$= s\,(p+q)^{s-1} = s, \qquad \ldots\ldots(7)$$

$$\frac{\partial^2}{\partial p^2}\sum_1^s {}^sC_r p^r q^{s-r} = \frac{\partial}{\partial p}[s(p+q)^{s-1}]$$

$$= s(s-1). \qquad \text{.....(8)}$$

Combining (5) and (7) we have

$$\sum_1^s r\,{}^sC_r p^r q^{s-r} = sp. \qquad \text{.....(9)}$$

Combining (6) and (8)

$$\sum_1^s r^2\,{}^sC_r p^r q^{s-r} = s(s-1)p^2 + sp. \qquad \text{.....(10)}$$

Substituting in (5) and (6) from (9) and (10)

$$M = \frac{p.s}{1-q^s} = p.c_s. \qquad \text{.....(11)}$$

Thus

$$\sigma_s^2 = \frac{s(s-1)p^2 + sp}{1-q^s} - \frac{s^2p^2}{[1-q^s]^2}$$

$$= \frac{s(1-q)}{[1-q^s]^2}\{sq^s(q-1) + q(1-q^s)\}$$

$$= k_s. \qquad \text{.....(12)}$$

The relation given in (1) is the proportion of recessives in a pool of fraternities each containing at least one recessive. If the mean number of recessives in a family of s is $p.c_s$ (11) the total number expected in a group of n_s families is

$$n_s p.c_s,$$

or if the pool contains families of all sizes from unity to c the maximum size of the family the expected number of recessives altogether is

$$\sum_1^c n_s p.c_s.$$

The standard deviation of $n_s p c_s$ is by (12)

$$\sqrt{k_s + k_s \ldots \text{ to } n_s \text{ terms}} = \sqrt{n_s k_s}.$$

Similarly since there is zero correlation between successive samples of families of different sizes the total expectation of recessives in a pool of families of various sizes is

$$\Sigma n_s p c_s \pm \sqrt{\Sigma n_s k_s}.$$

The restatement of the problem as given above makes the numerical computation less cumbersome, as the following example will illustrate. It also brings into clearer perspective the ratio of the discrepancy between observed and calculated values for families of different sizes. The table given below summarises matings of two normal parents with at least one offspring affected with total colour blindness. The data were taken from Dr Julia Bell's monograph and summarised for me by Mr E. A. Shrimpton, B.Sc. Thirty-seven sibships containing 196 individuals, of whom 91 are affected, are included. Of these, 10 affected are the offspring of first-cousin marriages and altogether 29 are the offspring of consanguineous unions. Four sibships have parents who are known to be first cousins and 13 sibships in all have consanguineous parents.

Size of fraternity (s)	Number of fraternities (n_s)	Number of affected: Observed (t_s)	Number of affected: Expected ($r_s=n_sp.c_s$)	Difference (Δ) t_s-r_s	$\sigma^2=n_sk_s$	Δ/σ
1	1	1	1·00	0·00	0·000	—
2	2	4	2·28	1·72	0·2449	3·9
3	6	12	7·78	4·22	1·5778	3·3
4	7	18	10·24	7·76	2·9403	4·5
5	5	8	8·19	−0·19	2·9589	0·1
6	3	7	5·47	1·53	2·3278	1·0
7	6	17	12·12	4·88	5·8214	2·0
8	4	13	8·89	4·11	4·6896	1·9
9	2	8	4·86	3·14	2·7604	1·9
10	1	3	2·65	0·35	1·5917	0·3
Total	37	91	63·5	—	24·913	—

For the entire series of sibships the expected number of affected is $63{\cdot}5 \pm \sqrt{24{\cdot}9}$. The difference between the expected and observed number is 27·5 or 5·5 times the standard deviation of the expected number. It is to be noted that the discrepancy between the observed and expected number is much greater for families of less than five members, as would be expected if it were due to biassed selection of the data.

I have to thank Prof. Haldane for kindly sending me the MS. of his alternative method.

SUMMARY.

A method by which the standard deviation of the expected *number* of recessives may be calculated for human or other data derived from small families is given. This involves the same functions as previously tabulated for a method of comparing the hypothetical expectation of recessives with an "adjusted" proportion based on the observed data. It leads to precisely the same numerical results and is free from an *a priori* assumption implicit in the earlier statement.

REFERENCES.

(1) FISHER, R. A. (1921). "On the mathematical foundations of theoretical statistics." *Phil. Trans. Roy. Soc.* A, **222**, 309.

(2) HALDANE, J. B. S. (1932). "A method for investigating recessive characters in Man." *Journ. Gen.* **25**.

(3) HOGBEN, L. T. (1931). "The genetic analysis of familial traits. I. Single gene substitutions." *Journ. Gen.* **25**.