17 March 1931.

J.B.S. Haldane, Esq., M.A.,
John Innes Horticultural Inst.,
Mostyn Road,
Merton Park, S.W. 19.

Dear Haldane,

Your problem is a very good example of the general use of maximum likelihood as explained in Section 57.1 of my book.

If you are satisfied in respect of viability etc., that the data ought to be combined, you proceed to write down the expectation and observed frequencies in each set of experiments:

(α) Backcrossed females

Expectation	$\frac{1}{2}n_1\{1-p,$	$p,$	$p,$	$1-p\}$
Observed	a_1	b_1	c_1	d_1

(β) Backcrossed males

Expectation	$\frac{1}{2}n_2\{1-q,$	$q,$	$q,$	$1-q\}$
Observed	a_2	b_2	c_2	d_2

(γ) F_2 Coupling

Expectation	$\frac{1}{4}n_3\{2+(1-p)(1-q),$	$1-(1-p)(1-q),$	$1-(1-p)(1-q),$	$(1-p)(1-q)\}$
Observed	a_3	b_3	c_3	d_3

(δ) F_2 Repulsion

Expectation	$\frac{1}{4}n_4\{2+pq,$	$1-pq,$	$1-pq,$	$pq\}$
Observed	a_4	b_4	c_4	d_4

Thence the logarithm of the likelihood is

$$(a_1+d_1)\log(1-p)+(b_1+c_1)\log p+(a_2+d_2)\log(1-q)+(b_2+c_2)\log q$$
$$+a_3\log\{2+(1-p)(1-q)\}+(b_3+c_3)\log\{1-(1-p)(1-q)\}+d_3\log(1-p)(1-q)$$
$$+a_4\log(2+pq)+(b_4+c_4)\log(1-pq)+d_4\log pq$$

and the two equations for p and q are found by differentiation, are

$$\frac{\partial L}{\partial p}=-\frac{a_1+d_1+d_3}{1-p}+\frac{b_1+c_1+d_4}{p}-\frac{a_3(1-q)}{2+(1-p)(1-q)}+\frac{(b_3+c_3)(1-q)}{1-(1-p)(1-q)}+\frac{a_4q}{2+pq}-\frac{(b_4+c_4)q}{1-pq}=0$$

and

$$\frac{\partial L}{\partial q}=-\frac{a_2+d_2+d_3}{1-q}+\frac{b_2+c_2+d_4}{q}-\frac{a_3(1-p)}{2+(1-p)(1-q)}+\frac{(b_3+c_3)(1-p)}{1-(1-p)(1-q)}+\frac{a_4p}{2+pq}-\frac{(b_4+c_4)p}{1-pq}=0$$

in which you will notice that all weighting is taken care of automatically by means of the observed frequencies.

Solution proceeds numerically, using probably the values from (α) and (β) as first approximations.

To systematise the process, and also to estimate the precision of the solution, it will be useful to calculate

$$-\frac{\partial^2 L}{\partial p^2},\quad -\frac{\partial^2 L}{\partial p\,\partial q},\quad -\frac{\partial^2 L}{\partial q^2}$$

for the trial values of p and q. There is no need to trouble about algebraic simplification.

If p_0, q_0 are the trial values, then the corrections δp_0, δq_0 are given by

$$\delta p_0\left(-\frac{\partial^2 L}{\partial p_0^2}\right)+\delta q_0\left(-\frac{\partial^2 L}{\partial p_0\,\partial q_0}\right)=\frac{\partial L}{\partial p_0}$$

$$\delta p_0\left(-\frac{\partial^2 L}{\partial p_0 \partial q_0}\right) + \delta q_0\left(-\frac{\partial^2 L}{\partial q_0^2}\right) = \frac{\partial L}{\partial q_0}$$

giving a new solution

$$p_1 = p_0 + \delta p_0, \quad q_1 = q_0 + \delta q_0$$

which will I think be efficient, though capable of further improvement in the same way. If a, b, c stand for the second differential coefficients of L, with changed sign, then $\frac{1}{a}$ gives the sampling variance of p <u>for given observed q</u>, while the total sampling variance of p is $\frac{c}{ac-b^2}$, and the sampling covariance of p and q is $b/(ac-b^2)$. Consequently, the sampling variance of $p-q$ is

$$\frac{a+c-2b}{ac-b^2}$$

and you will probably want this. A point which might be worth while following through in the algebra is the amount of information respecting $p-q$, i.e. $(ac-b^2)/(a+c-2b)$ to be expected from samples of n_1, n_2, n_3, and n_4, for this would show how much each type of experiment contributed to the total amount of information available.

Do you believe the Berlin tale? The fallacies of Edin's work on Stockholm are fairly easy to see, but I have not looked at the Berlin stuff.

Yours sincerely,

R. A. Fisher