Statistical Science
1997, Vol. 12, No. 1, 35–47

Three Early Papers on Efficient Parametric Estimation

A. W. F. Edwards

Abstract. Three papers from the early history of efficient parametric estimation are reprinted with commentary: (1) Fisher (1912), "On an absolute criterion for fitting frequency curves"; (2) Engledow and Yule (1914), "The determination of the best value of the coupling-ratio from a given set of data"; and (3) Fisher (1922), "The systematic location of genes by means of crossover observations."

Key words and phrases: Parametric estimation, maximum likelihood, minimum χ^2, genetic linkage.

0. INTRODUCTION

Many of the concepts which underlie efficient parametric estimation arose early in the history of mathematical statistics, especially in the work of Gauss and Laplace, but it was not until the time of R. A. Fisher that a recognizably modern theory emerged. His "On the mathematical foundations of theoretical statistics" appeared in 1922 (Fisher, 1922a), and it is a measure of its influence that among the words and phrases it introduced in their modern technical senses were *efficiency*, *parameter*, *consistency*, *statistic* and *method of maximum likelihood* (David, 1995), and probably *estimation* too. Fisher himself later wrote "This is the first large-scale attack on the problem of estimation" (Fisher, 1950). A recent introduction to this famous paper has been provided by Geisser (1992), and the paper itself has been reprinted in Fisher (1950) and Fisher (1971) as well.

We here reproduce three short papers from the decade 1912–1922 which provide some background to the development of efficient estimation. The first paper, "On an absolute criterion for fitting frequency curves" (Fisher, 1912), was not regarded by the author himself as worth reproducing in his *Contributions to Mathematical Statistics* (Fisher, 1950), nor did he refer to it in the first edition of *Statistical Methods for Research Workers* (Fisher, 1925). It did, however, receive mentions in 1915 and 1921 (Fisher, 1915, 1921), and also in the 1922 paper, as containing "My original statement of the Method of Maximum Likelihood," and it is of signal interest for that reason.

The second paper, "The determination of the best value of the coupling-ratio from a given set of data" (Engledow and Yule, 1914), is remarkable not only for having introduced the method of minimum χ^2 but for never once having been referred to, other than by Yates (1952) in his obituary of Yule. It should have had an honored place in the history of genetic linkage estimation (Edwards, 1996), but it was no more noticed in genetics than it was in statistics.

The third paper, "The systematic location of genes by means of crossover observations" (Fisher, 1922b), is of interest because it contains the first application of the method of maximum likelihood and was clearly intended by its author as a companion to the theoretical paper of that year. It applies the method to the same field as Engledow and Yule (1914), namely, genetic linkage, the practical area most closely associated with contemporary developments in estimation theory.

1. ON AN ABSOLUTE CRITERION FOR FITTING FREQUENCY CURVES (FISHER, 1912)

This, Fisher's first paper, was written and published while he was still an undergraduate (see Edwards, 1974, for a more extensive account than that given here). Very little is known about the original impetus for the paper other than what is provided by internal evidence. F. J. M. Stratton, whom Fisher thanks (though with his initials in the wrong order) was a young Fellow of Fisher's Cambridge College, Gonville and Caius, teaching the students mathe-

A. W. F. Edwards is Reader in Biometry, University of Cambridge, Department of Community Medicine, Forvie, Robinson Way, Cambridge CB2 2SR, United Kingdom (e-mail: awfe@medschl.cam.ak.uk).

matics and working as an Assistant in the University Observatory. In the Easter Term 1911 he had lectured at the Observatory on *Calculation of Orbits from Observations*, and during the next academic year on *Combination of Observations* in the Michaelmas Term (1911), the first term of Fisher's third and final undergraduate year. It is very likely that Fisher attended Stratton's lectures and subsequently discussed statistical questions with him during mathematics supervisions in College, and that he wrote the 1912 paper as a result.

The paper advocates the method of maximum likelihood, though not under that name. Fisher defined *likelihood* nine years later (Fisher, 1921) and coined the phrase *method of maximum likelihood* the year after (Fisher, 1922a); later on he sometimes preferred *maximal* to *maximum* (e.g., Fisher, 1937). In Section 1 Fisher points out that the multiplicity of different criteria for estimating parameters (to use the modern terms) is theoretically unsatisfactory. In Section 2 he mentions least squares and the method of moments, and in the first half of Section 3 he disposes of the former because of the arbitrariness in the choice of measure for the variate, and the latter because of the arbitrariness in the choice of the moments to equate. He then introduces his criterion of maximum likelihood, apparently basing it on an implied Bayesian uniform prior for the parameters, since he writes of the method as leading to "the most probable set of values" for the parameters.

In Section 4 Fisher applies the criterion to the parameters of the normal distribution and in Section 5 he investigates the likelihood surface ("inverse probability system") for the two parameters, adding a figure notable for the omission of the contours of the surface to which it refers. He remarks, in effect, that the maximum-likelihood estimate of the reciprocal of the variance (a word not coined by Fisher until 1918) and the mean jointly does not give the same result for the reciprocal of the variance as when the mean is integrated out—the difference being the well-known factor $n/(n-1)$—but he adds the all-important rider "that the integration with respect to m [the mean] is illegitimate and has no definite meaning with respect to inverse probability." This point of view is elaborated in the final Section 6, where Fisher insists that the likelihood "is a relative probability only, suitable to compare point with point, but incapable of being interpreted as a probability distribution over a region, or of giving any estimate of absolute probability."

The paper is thus paradoxical in its attitude to "inverse probability." Fisher appears to base his maximizing method on inverse probability assuming an implicit uniform prior, but then denies that the resulting "surface" for two parameters can be manipulated as if it were a probability distribution. He himself later drew attention to the apparent conflict (Fisher, 1922a). Forthcoming papers by J. Aldrich and me in *Statistical Science* will examine this question in more detail.

2. THE DETERMINATION OF THE BEST VALUE OF THE COUPLING-RATIO FROM A GIVEN SET OF DATA (ENGLEDOW AND YULE, 1914)

The estimation procedure known as the method of minimum χ^2, according to which the parameter values of a discrete distribution are chosen so as to minimize the value of goodness-of-fit χ^2, became well-known through its inclusion in the chapter, "The principles of statistical estimation," in the second and subsequent editions of Fisher's book, *Statistical Methods for Research Workers* (Fisher, 1928; see also Fisher and Balmukand, 1928). Fisher first discussed it in 1922 (Fisher, 1922a), when he pointed out that if the log-likelihood was expanded in a Taylor's series in x (the difference between the observed and expected values), then the first nonzero term in the expansion was equal to $-\frac{1}{2}\chi^2$. This well-known result was probably inspired by some rather cryptic mathematics by Haldane (1919a). Fisher mentioned that the method of minimum χ^2, of which he was critical, had been discussed by K. Smith (1916).

Kirstine Smith was a graduate student in Karl Pearson's Biometric Laboratory at University College London from 1915 (E. S. Pearson, 1990). Her paper does not give any earlier reference to the method, but it ends with the acknowledgment "The present paper was worked out in the Biometric Laboratory and I have to thank Professor Pearson for his aid throughout the work." Smith discussed the use of the method both with discrete distributions and grouped data from continuous distributions, a circumstance which prompted Fisher, in 1916, to send Pearson a note for *Biometrika* commenting on the unsatisfactory property that the operation of the method would be sensitive to the fineness with which the data were grouped (an echo of the similar point in the 1912 paper). Pearson declined to publish the note (E. S. Pearson, 1968).

In fact the method of minimum χ^2 had already been advocated by Engledow and Yule in 1914 in the paper here reprinted. They had invented it for the estimation of the recombination fraction in genetic linkage, an application in which the data are in the form of frequencies and thus discrete. Their parameter is $p = \frac{1}{2}(1-\theta)$, where θ is the modern recombination fraction. The coupling-ratio to which they refer is then $p : \frac{1}{2} - p$, or equivalently $1 - \theta : \theta$. In the original announcement of their paper in the *Cambridge University Reporter* (which I came across serendipitously) the order of the authors is Yule and Engledow (1914), as indeed it is also when referred to by Engledow (1914). Yule, who had been appointed University Lecturer in Statistics at Cambridge in 1912, had previously been in London and closely associated with Pearson. This fact, coupled with the original order of the authors and the fact that Engledow (later Sir Frank Engledow, Drapers Professor of Agriculture) was a plant physiologist, makes it probable that Yule was the originator. In the only comment on the paper of which I am aware, Yates (1952), in his Royal Society obituary of Yule, stated "Yule was also, in conjunction with Engledow, the first to put forward the use of minimum χ^2 for the estimation of linkage."

Engledow and Yule write simply that "the method suggested seems much better than any now in use" because it gives the best fit on the χ^2 criterion. Interestingly enough, the original announcement (Yule and Engledow, 1914) carries a summary in which the authors say "A method is given for determining the best value, i.e., the value that will make the probability of the observations a maximum on '*Professor Pearson's test*'." Had they not added the words in (my) italics, they would have been describing the method of maximum likelihood; as it is, they confusingly imply that the "P-value" in Pearson's test gives the probability of the observations.

The estimating equation is of the fourth degree, which the authors solve by Newton's method using data of Engledow's (1914). Two other sets of data are similarly treated, and the paper ends with a note on errors.

This paper has never appeared in the linkage literature (which is commonly taken to start with Haldane, 1919a, b, who did not refer to it; see Edwards, 1996) nor in the statistical literature. It was not included in *The Statistical Papers of George Udny Yule* (Yule, 1971), nor is it mentioned in *An Introduction to the Theory of Statistics* by Yule and Kendall (1937) or in the section on minimum χ^2 in Kendall (1955). The books on linkage estimation by Mather (1938) and Bailey (1961) are similarly silent, as is the literature on human linkage estimation (see Ott, 1991, and C. A. B. Smith, 1986). A summary was published in *The American Naturalist* (Engledow and Yule, 1915) but went unremarked.

Like Kirstine Smith, Engledow and Yule gave no reference for the method, but it is not unlikely that they will have seen an earlier paper by Harris (1912) which suggested employing Pearson's χ^2 test on genetical data and which went so far as to test two different hypotheses on the same data, indicating that the one with the lower value of χ^2 was to be preferred: "[χ^2's] applicability to the problem of testing the goodness of fit of Mendelian ratios seems obvious, but since, as far as I can ascertain, it has nowhere been applied to this problem, it seems worth while to call the attention of students of genetics to its usefulness." Harris was not quite correct, however, for Weldon (1902) had actually applied χ^2 to Mendel's three-factor data only two years after Pearson (1900) had invented the test. Yule might well have been familiar with this.

3. THE SYSTEMATIC LOCATION OF GENES BY MEANS OF CROSSOVER OBSERVATIONS (FISHER, 1922b)

This paper, the first application of the method of maximum likelihood under that name, can be viewed as the complement to Fisher (1922a). As Fisher himself wrote to his old biology teacher in 1929, "The fact is that nearly all my statistical work is based on biological material and much of it has been undertaken merely to clear up difficulties in experimental technique" (quoted by Box, 1978). Biological problems inspired his development of efficient estimation, and the estimation of genetic linkage, in particular, was a fertile proving-ground.

Section 1 of the paper discusses the nature of the linkage estimation problem and reminds readers that "It has been shown that the whole of the information supplied by the data (Fisher, 1922a) is made use of by the method of maximum likelihood." In Section 2 Fisher shows how to write down the likelihood function for the case of three loci on the sex chromosome of the fruit-fly *Drosophila willistoni*, the two parameters being the recombination fractions between adjacent loci. He assumes these are additive (on the grounds that the loci are close and crossing-over rare) and that the order of the loci is given. He then finds the maximum-likelihood equations by differentiating the log-likelihood (though without the explanation) and shows how they can be approximately linearized. Without solving these, Fisher turns immediately to a fuller example involving eight loci and thus seven parameters (Section 3), for which he achieves a complete approximate solution. J. H. Edwards (1989) has reworked this example and derived the exact maximum-likelihood estimates of the parameters. Fisher ends by noting that in another 1922 paper (Fisher, 1922c) he has shown how to find the correct number of degrees of freedom for goodness-of-fit χ^2, and he applies this test to the seven-parameter example.

It is remarkable that the first worked example of the use of the method of maximum likelihood should involve not one or two, but seven parameters. This paper is the only one that Fisher ever published in the *American Naturalist* and was clearly designed to introduce workers in genetic linkage, for whom the *American Naturalist* was the principal journal at the time, to efficient parametric estimation.

The year 1925 saw the publication of the first edition of Fisher's famous book *Statistical Methods for Research Workers.* In the Introductory chapter, Fisher discussed the need for efficient estimation procedures, and gave as an example of the method of maximum likelihood the estimation of the recombination fraction from F_2 data. For the second editon (Fisher, 1928), this section was expanded into a new chapter, "The principles of statistical estimation." The same material, treated more expansively, appeared in a joint paper in the same year (Fisher and Balmukand, 1928). Further details may be found in Edwards (1996).

REFERENCES

Bailey, N. T. J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage.* Clarendon, Oxford.

Box, J. F. (1978). *R. A. Fisher: The Life of a Scientist.* Wiley, New York.

David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *Amer. Statist.* **49** 121–133.

Edwards, A. W. F. (1974). The history of likelihood. *Internat. Statist. Rev.* **42** 9–15.

Edwards, A. W. F. (1996). The early history of the statistical estimation of linkage. *Ann. Human Genetics* **60** 237–249.

Edwards, J. H. (1989). The locus positioning problem. *Ann. Human Genetics* **53** 271–275.

Engledow, F. L. (1914). A case of repulsion in wheat. *Proc. Cambridge Philos. Soc.* **17** 433–435.

Engledow, F. L. and Yule, G. U. (1914). The determination of the best value of the coupling-ratio from a given set of data. *Proc. Cambridge Philos. Soc.* **17** 436–440.

Engledow, F. L. and Yule, G. U. (1915). The determination of the best value of the coupling-ratio from a given set of data. *American Naturalist* **49** 127–128.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41** 155–160.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **9** 507–521.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.

Fisher, R. A. (1922a). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

Fisher, R. A. (1922b). The systematic location of genes by means of crossover observations. *American Naturalist* **56** 406–411.

Fisher, R. A. (1922c). On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. Roy. Statist. Soc.* **85** 87–94.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

Fisher, R. A. (1928). *Statistical Methods for Research Workers*, 2nd ed. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1937). Professor Karl Pearson and the method of moments. *Annals of Eugenics* **7** 303–318.

Fisher, R. A. (1950). *Contributions to Mathematical Statistics*. Wiley, New York.

Fisher, R. A. (1971). *Collected Papers of R. A. Fisher* (J. H. Bennett, ed.) **1**. Univ. Adelaide, Australia.

Fisher, R. A. and Balmukand, B. (1928). The estimation of linkage from the offspring of selfed heterozygotes. *Journal of Genetics* **20** 79–92.

Geisser, S. (1992). Introduction to Fisher (1922). On the mathematical foundations of theoretical statistics. In *Breakthroughs in Statistics* **1**. *Foundations and Basic Theory* (S. Kotz and N. L. Johnson, eds.) 1–10. Springer, New York.

Haldane, J. B. S. (1919a). The probable errors of calculated linkage values, and the most accurate method of determining gametic from certain zygotic series. *Journal of Genetics* **8** 291–297.

Haldane, J. B. S. (1919b). The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8** 299–309.

Harris, J. A. (1912). A simple test of the goodness of fit of Mendelian ratios. *American Naturalist* **46** 741–745.

Kendall, M. G. (1955). *The Advanced Theory of Statistics* **2**, 3rd ed. Griffin, London.

Mather, K. (1938). *The Measurement of Linkage in Heredity.* Methuen, London.

Ott, J. (1991). *Analysis of Human Genetic Linkage.* Johns Hopkins Univ. Press.

Pearson, E. S. (1968). Some early correspondence between W. S. Gosset, R. A. Fisher and Karl Pearson, with notes and comments. *Biometrika* **55** 445–457. [Reprinted in *Studies in the History of Statistics and Probability* (E. S. Pearson and M. G. Kendall, eds.) 405–417. Griffin, London (1970).]

Pearson, E. S. (1990). *'Student.'* Clarendon, Oxford.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London, Edinburgh, and Dublin Philosophical Magazine, 5th Ser.* **50** 157–175.

Smith, C. A. B. (1986). The development of human linkage analysis. *Ann. Human Genetics* **50** 293–311.

Smith, K. (1916). On the 'best' values of the constants in frequency distributions. *Biometrika* **11** 262–276.

Weldon, W. F. R. (1902). Mendel's laws of alternative inheritance in peas. *Biometrika* **1** 228–254.

Yates, F. (1952). George Udny Yule. *Obituary Notices of Fellows of the Royal Society* **8** 309–323.

Yule, G. U. (1971). *Statistical Papers of George Udny Yule* (selected by A. Stuart and M. G. Kendall). Griffin, High Wycombe, U.K.

Yule, G. U. and Engledow, F. L. (1914). The determination of the best value of the coupling ratio from a given set of data (abstract). *Cambridge University Reporter* **44** 757.

Yule, G. U. and Kendall, M. G. (1937). *An Introduction to the Theory of Statistics*, 11th ed. Griffin, London.

On an Absolute Criterion for Fitting Frequency Curves[1]

R. A. Fisher
Gonville and Caius College, Cambridge

1. If we set ourselves the problem, in its essence one of frequent occurrence, of finding the arbitrary elements in a function of known form, which best suit a set of actual observations, we are met at the outset by an arbitrariness which appears to invalidate any results we may obtain. In the general problem of fitting a theoretical curve, either to an observed curve, or to an observed series of ordinates, it is, indeed, possible to specify a number of different standards of conformity between the observations and the theoretical curve, which definitely lead to different though mutually approximate results. This mutual approximation, though convenient in practice in that it allows a computer to make a legitimate choice of the method which is arithmetically simplest, is harmful from the theoretical standpoint as tending to obscure the practical discrepancies, and the theoretical indefiniteness which actually exist.

2. Two methods of curve fitting may first be noted, in which we shall use a sign of summation when the observations comprise a finite number of ordinates only, and an integral sign when the curve itself is observed, even though the integrals may in practice be estimated by a process of summation.

Consider f a function of known form, involving arbitrary elements, $\theta_1, \theta_2, \ldots, \theta_r$ and x the abscissa; let y be the observed ordinate corresponding to a given x. Then a natural method of getting suitable values for $\theta_1, \theta_2, \ldots, \theta_r$, that is of fitting the observations, is to make $\int_{-\infty}^{+\infty}(f-y)^2\,dx$ a minimum for variations of any θ; or if the ordinate is observed at finite and equal intervals of the abscissa, we should substitute $\sum(f-y)^2$ for the integral.

This method will obviously give a good result to the eye in cases where a good result is possible; the equations to which it gives rise are, however, often practically insoluble, a difficulty which renders the method less useful than the simplicity of its principle would suggest.

The method of moments is possibly of more value, though its arbitrary nature is more apparent. If we solve the first r equations of the type

$$\int_{-\infty}^{+\infty} f\,dx = \int_{-\infty}^{+\infty} y\,dx$$

or

$$\sum f = \sum y,$$

$$\int_{-\infty}^{+\infty} xf\,dx = \int_{-\infty}^{+\infty} xy\,dx$$

or

$$\sum xf = \sum x,$$

$$\int_{-\infty}^{+\infty} x^2 f\,dx = \int_{-\infty}^{+\infty} x^2 y\,dx, \quad \text{etc.}$$

or

$$\sum x^2 f = \sum x^2 y, \quad \text{etc.},$$

we may obtain values for the r unknowns, which will give a curve to the eye about as good as that of least squares, by a method which for some purposes is found to be more convenient.

3. The first of the above methods is obviously inapplicable to frequency curves, even if we wished to accept its standard of "goodness of fit." If we suppose that the observations comprise a complete and continuous curve, an arbitrariness arises in the scaling of the abscissa line, for if ξ, any function of x, were substituted for x, the criterion would be modified. While, if a finite number of observations are grouped about a series of ordinates, there is an additional arbitrariness in choosing the positions of the ordinates and the distances between them.

For a finite number, n, of observations the method of moments really gives the equations

$$\sum f = n, \quad \sum xf = \sum_1^n x,$$

$$\sum x^2 f = \sum_1^n x^2, \quad \text{etc.},$$

against which the above objections cannot be urged; still a choice has been made without theoretical

[1] Reprinted from *Messenger of Mathematics* **41** 155–160 (1912).

justification in selecting this set of r equations of the general form

$$\sum x^p f = \sum_1^n x^p .$$

But we may solve the real problem directly.

If f is an ordinate of the theoretical curve of unit area, then $p = f\,\delta x$ is the chance of an observation falling within the range δx; and if

$$\log P' = \sum_1^n \log p ,$$

then P' is proportional to the chance of a given set of observations occurring. The factors δx are independent of the theoretical curve, so the probability of any particular set of θ's is proportional to P, where

$$\log P = \sum_1^n \log f .$$

The most probable set of values for the θ's will make P a maximum.

If a continuous curve is observed—e.g., the period during which a barometer is above any level during the year is a continuous function from which may be derived the relative frequency with which it stands at any height—we should use the expression

$$\log P = \int_{-\infty}^{\infty} y \log f \, dx .$$

4. For example, let us take the normal curve of frequency of errors

$$f = \frac{h}{\sqrt{\pi}} - \exp[-h^2(x-m)^2],$$

where h and m are to be determined to fit a set of n observations. Our criterion gives, neglecting a constant term,

$$\begin{aligned}\log P &= n \log h - h^2 \sum (x-m)^2 \\ &= n \log\, h - h^2 n(m-\bar{x})^2 - h^2 \sum (x-\bar{x})^2,\end{aligned}$$

where $n\bar{x} = \sum x$.

Differentiating with respect to m, we get

$$-2h^2 n(m-\bar{x}) = 0,$$

and with respect to h

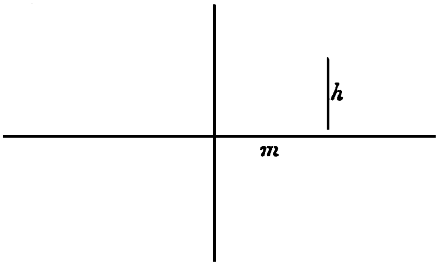
$$\frac{n}{h} = 2h\left\{n(m-\bar{x})^2 + \sum (x-\bar{x})^2\right\};$$

giving $m = \bar{x}\ 2h^2 = n/\Sigma v^2$, where v is written for $x - \bar{x}$; neglecting the solution $h = 0$, $m = \infty$, when P is a minimum. Since the value usually accepted is

$$2h^2 = \frac{n-1}{\sum v^2},$$

it will be necessary to examine one or two of the methods by which this answer is obtained.

5. Corresponding to any pair of values, m and h, we can find the value of P, and the inverse probability system may be represented by the surface traced out by a point at a height P above the point on a plane, of which m and h are the coordinates.



The actual maximum of P occurs, as we have shown, at the point

$$m = \bar{x},$$
$$2h^2 = \frac{n}{\sum v^2}.$$

(a) In an interesting investigation[*] Mr. T. L. Bennett takes the maximum value of

$$\int_{-\infty}^{+\infty} P\, dm,$$

for variations of h, i.e., of

$$h^n \exp[-h^2 \sum (x-\bar{x})^2] \cdot \int_{-\infty}^{+\infty} \exp[-h^2 n(m-\bar{x})^2]\, dm,$$

or of

$$\frac{\sqrt{\pi}}{h\sqrt{n}} h^n \exp\left[(-h^2)\sum v^2\right],$$

whence

$$(n-1)h^{n-2} = 2h^n \sum v^2,$$
$$2h^2 = \frac{n-1}{\sum v^2},$$

[*] *Errors of Observation,* Technical Lecture, No. 4, 1907–08, Survey Department, Egypt.

a determination which gives the section perpendicular to the axis of h, the area of which is a maximum, though it does not pass through the actual maximum point.

We shall see (in §6) that the integration with respect to m is illegitimate and has no definite meaning with respect to inverse probability.

(*b*) The usual text-book discussion* of the relation between h^2 and μ^2, where $n\mu^2 = \sum v^2$, assumes that the observed value of μ^2 is the same as the average value for a large number of sets of n observations each; thus the average value of $(x - m)^2$ being $1/(2h^2)$, the average value of $(\bar{x} - m)^2$—that is of

$$\frac{1}{n^2}(x_1 - m + x_2 - m \cdots x_n - m)^2$$

equals the average value of $(1/n^2)\Sigma_n^1(x - m)^2$, since the product terms go out—is

$$\frac{1}{n^2}\frac{n}{2h^2} = \frac{1}{2nh^2},$$

and the average value of $n\mu^2 = \sum(\bar{x} - x)^2$ is that of

$$\sum(m - x)^2 - n(\bar{x} - m)^2,$$

that is,

$$\frac{n}{2h^2} - \frac{1}{2h^2} = \frac{n-1}{2h^2};$$

and if the most probable value for h was such as to make the observed quantity μ^2 take up its average value we should have

$$h^2 = \frac{n-1}{2n\mu^2}.$$

The basis of the above method becomes less convincing when we consider that the frequencies with which different values of μ^2 occur, for a given value of h, cannot give a normal distribution, since μ^2 can only vary from 0 to $+\infty$; and that a frequency distribution might easily be constructed to have a zero at its mean, in which case the above basis would give us perhaps the only value for h, which could not possibly have given rise to the observed value of μ^2.

The distinction between the most probable value of h, and the value which makes μ^2 take up its average value, is illustrated by our treatment of the quantity $(\bar{x} - m)^2$, the average value of which is $1/(2nh^2)$, but the most probable value being zero, we say that the most probable value of m is $\bar{x}$, not

$$\bar{x} \pm \frac{1}{h\sqrt{(2n)}}.$$

If a frequency curve of unit area were drawn, showing the frequencies with which different values of μ^2 occur, for a given h, and if b were the ordinate corresponding to the observed μ^2, then we should expect the equation

$$\frac{\partial b}{\partial h} = 0$$

to give the most probable value of h. It is sufficient here, however, to point out the incorrectness of the assumption upon which some writers on the Theory of Errors have based their results.

6. We have now obtained an absolute criterion for finding the relative probabilities of different sets of values for the elements of a probability system of known form. It would now seem natural to obtain an expression for the probability that the true values of the elements should lie within any given range. Unfortunately we cannot do so. The quantity P must be considered as the relative probability of the set of values $\theta_1, \theta_2, \ldots, \theta_r$; but it would be illegitimate to multiply this quantity by the variations $d\theta_1, d\theta_2, \ldots, d\theta_r$ and integrate through a region, and to compare the integral over this region with the integral over all possible values of the θ's. P is a relative probability only, suitable to compare point with point, but incapable of being interpreted as a probability distribution over a region, or of giving any estimate of absolute probability.

This may be easily seen, since the same frequency curve might equally be specified by any r independent functions of the θ's, say $\phi_1, \phi_2, \ldots, \phi_r$, and the relative values of P would be unchanged by such a transformation; but the probability that the true values lie within a region must be the same whether it is expressed in terms of θ or ϕ, so that we should have for all values

$$\partial(\theta_1, \theta_2, \ldots, \theta_r)/\partial(\phi_1, \phi_2, \ldots, \phi_r) = 1$$

a condition which is manifestly not satisfied by the general transformation.

In conclusion I should like to acknowledge the great kindness of Mr. J. F. M. Stratton, to whose criticism and encouragement the present form of this note is due.

*Chauvenet, *Spherical Astronomy*, Note II., Appendix §17.

The Determination of the Best Value of the Coupling-Ratio from a Given set of Data[1]

F. L. Engledow and G. Udny Yule
St. John's College

Many workers in Mendelism who have come across cases in which coupling or repulsion occurred must have felt the necessity for some general method by which to determine from their data the best value to assign to the coupling-ratio, apart from any theory as to the ratios that are possible. Mr. G. N. Collins (*Am. Nat.* vol. XLVI., 1912) is, so far as we are aware, the only writer who has suggested any such method. He worked out the value of a coefficient of association for the whole series of possible ratios, 1 : 1 : 1 : 1, 2 : 1 : 1 : 2, etc., and then used the observed value of the same coefficient to decide which ratio gave the best agreement with the facts. While this method is very simple and convenient, it does not seem to lead to the most advantageous value for the ratio.

The test to be used for the closeness of agreement between the theoretical and observed frequencies seems clearly to be that developed by Professor Pearson (*Phil. Mag.*, vol. L., 1900). If $F_1 F_2 F_3 F_4$ etc. are a set of theoretical or expected frequencies, and $F'_1\, F'_2\, F'_3\, F'_4$ etc. are those observed, and if

$$\chi^2 = \sum \frac{(F' - F)^2}{F},$$

the probability P that in random sampling deviation-systems of equal or greater improbability will arise is a function of χ^2 which decreases continuously as χ^2 increases. The values of this function for any number of frequencies from 3 to 30 have been tabulated by Mr. Palin Elderson (*Biometrika*, vol I.). In order to measure the closeness of agreement between an observed set of the four frequencies for any pair of characters, and the expectation based on any assumed ratio, it is only necessary to work out the value of χ^2 and turn up in Mr. Elderton's table the column headed $n' = 4$, where the probability that an equally bad or worse set of deviations might arise in sampling will be found. If P is high, the agreement is good; if low, it is bad. That value of the ratio, then, which gives the most satisfactory agreement with the data is the value which makes the probability P a maximum or χ^2 a minimum. The value of P is not accurate if any frequencies are small, as a normal distribution of errors is assumed in the calculation of the tables, but even from the empirical point of view the method suggested seems much better than any now in use.

Suppose the two factors to be A and B, and let the gametes be produced by the heterozygote in the following proportions:

$$\begin{array}{cccc} AB & Ab & aB & ab \\ p & (0.5-p) & (0.5-p) & p \end{array}$$

then, assuming random mating, zygotic forms will be produced in the following proportions:

$$\begin{array}{cccc} AB & Ab & aB & ab \\ (p^2+0.5) & (0.25-p^2) & (0.25-p^2) & p^2. \end{array}$$

Let the observed *proportions* of zygotes be $f_1 f_2 f_3 f_4$, where $f_1 + f_2 + f_3 + f_4 = 1$. Then we have to make a minimum the quantity

$$\frac{(p^2+0.5-f_1)^2}{p^2+0.5} + \frac{(0.25-p^2-f_2)^2}{0.25-p^2} + \frac{(0.25-p^2-f_3)^2}{0.25-p^2} + \frac{(p^2-f_4)^2}{p^2}.$$

Differentiating with respect to p and equating to zero, we find

$$\begin{aligned} &(f_2^2+f_3^2-f_1^2-f_4^2)p^8+(0.5f_1^2+f_2^2+f_3^2-0.5f_4^2)p^6 \\ (1)\quad &+(0.25f_2^2+0.25f_3^2+0.1875f_4^2-0.0625f_1^2)p^4 \\ &+0.0625f_4^2\,p^2-0.015625\,f_4^2=0. \end{aligned}$$

This is an equation of the fourth degree for p^2. A first approximation to the root required may be obtained by Collins' method or from the formula

$$(2)\qquad p^2 = 0.25(f_1+f_4-f_2-f_3)$$

(which gives the value of p that makes the sum of the squares of differences least), or by comparison with various calculated series, and the solution is then readily obtained by Newton's method.

[1] Reprinted from *Proc. Cambridge Philos. Soc.* **17** 436–440.

To take the data of the preceding note as an illustration, the values of the proportions f are 0.5634, 0.2207, 0.2019, and 0.0141; writing x for p^2, this gives the equation

$$-0.228147x^4+0.248083x^3+0.002566x^2$$
$$+0.00001244x-0.0000031094=0\,.$$

The data suggested a gametic ratio 1 : 3 : 3 : 1, which gives $p = 0.125$, $p^2 = 0.015625$. Trial shewed that 0.0156 was not a very close approximation to a root; 0.02 proved nearer to a solution, and Newton's method gave by two approximations $p^2 = 0.019715\ldots$. Hence $p = 0.1404$ and this gives a ratio 1 : 2.56. The observed frequencies were then compared with the frequencies to be expected from this ratio, and from the ratio 1 : 3 : 3 : 1. The results obtained were:

Ratio 1 : 3	$\chi^2 = 2.0974$	$P = 0.554$
Ratio 1 : 2.561	$\chi^2 = 1.9918$	$P = 0.574$

It will be observed that while the calculated ratio does give the better agreement, the difference is slight. In both cases results equally or more divergent from expectation would occur nearly as often as not owing to mere fluctuations of sampling. The result is an illustration of the now recognized fact that a considerable alteration in the coupling-ratio may mean but a small alteration in the closeness of fit.

Two other cases have been tried and gave the following results. Collins (*loc. cit.*, p. 579) gives the following data for the characters coloured aleurone and horny endosperm in maize:

Coloured-horny	1774
Coloured-waxy	263
White-horny	279
White-waxy	420

We find $p = 0.3891$ or a ratio 3.509 : 1. For this value of the ratio χ^2 is 0.60435 or $P = 0.947$, the calculated frequencies 1782, 270, 270, 414 being in very close agreement with those observed. For the 3 : 1 ratio, χ^2 is 9.106 or $P = 0.028$, and the divergence is therefore one that would only be likely to occur once in some 36 trials owing to the fluctuations of random sampling.

Finally, we took the data given by Bateson, Saunders and Punnett in the Fourth Report of the Evolution Committee (p. 16) for coupling between dark axils and fertility in sweet peas. Here we find $p = 0.4745$, which is equivalent to a ratio 18.608 : 1, as compared with the ratio 15 : 1 suggested in the Report and a value "about 20 : 1" by Collins. The relative merits of the ratios are apparent from the following:

Ratio 18.608 : 1	$\chi^2 = 3.7539$	$P = 0.294$
Ratio 15 : 1	$\chi^2 = 5.9226$	$P = 0.116$
Ratio 20 : 1	$\chi^2 = 3.8975$	$P = 0.275$

The ratio 15 : 1 is clearly much the poorest of these three: a worse fit is only likely to occur, owing to fluctuations of sampling, some 12 times in 100. A worse fit than that given by 20 : 1 may occur some 27 times in 100, and a worse fit than that given by our calculated ratio some 29 times in 100. The figures again shew, however, how great differences may be made in the coupling-ratio assumed without creating an impossible discordance between assumptions and fact. The mere agreement of the data, within the possible limits of fluctuations of sampling, with the frequencies deduced from some assumed ratio—as in the case of the above data for peas and the ratio 15 : 1—is very slight evidence in favour of the truth of the assumption, especially where the coupling-ratio is high, at least with such moderate numbers of observations as are at present available. Some light might, however, be thrown on the theory of reduplication by carrying out an examination of all the available cases, determining p or the coupling-ratio for each by equation (1). Such an examination we hope to carry out.

As p is not expressed explicitly as a function of the proportionate frequencies f by equation (1), we do not see our way to give its probable error by this method of determination. The value given by (2), however, is in some cases close to the value given by (1), viz. if no one of the frequencies is very small (cf. the data below), and its standard error can be determined without difficulty on the usual, though hardly quite justifiable, assumption that deviations in the frequencies are small compared with their mean values. As the standard errors by the two methods of determination are likely to be of the same order of magnitude, it seems worth while stating the result as at least a rough guide to the possible magnitude of fluctuations. Differentiating both sides of equation (2), squaring and summing, we have, utilising known results for the sums of squares and product sums (cf., e.g., Yule, *Jl. Stat. Soc.*, 1912, p. 601),

$$\text{(3)} \qquad \varepsilon_p^2 = \frac{1}{4N}\frac{(f_1+f_4)(f_2+f_3)}{(f_1+f_4)-(f_2+f_3)},$$

where ε_p is the standard error of p (to be multiplied by 0.6745 to obtain the probable error) and N is the number of observations. If there is coupling ($p > 0.25$), the coupling-ratio $r = p/(0.5 - p)$. Differentiating, squaring and summing again, we have

$$\varepsilon_r^2 = \varepsilon_p^2 \frac{1}{4(0.5 - p)^4}. \tag{4}$$

	No. of observations	Value of p from		r from		Standard error of values from (2)	
Case		(1)	(2)	(1)	(2)	p	r
Wheat	213	0.1404	0.1968	2.56	1.54	0.0430	0.56
Maize	2736	0.3891	0.3885	3.51	3.48	0.0049	0.20
Peas	885	0.4745	0.4744	18.6	18.5	0.00385	2.94

If there is repulsion ($p < 0.25$), the repulsion-ratio is $(0.5 - p)/p$ and p^4 must be read for $(0.5 - p)^4$ in the denominator of the above expression. The table shews for comparison the values of p and r given by equation (1) and equation (2) respectively, and the standard errors of p and of r as obtained by the latter method.

In the first case the two equations give very divergent results, the unsuitability of equation (2) for general use being shewn by its failure to give a good approximation to the best value of the ratio. In this case, no doubt, we must also regard the standard error of r (0.56) as of very uncertain validity. The magnitude of the standard error of r in the last case—nearly 3 units—again emphasises the caution that must be used before attaching importance to the precise values of these high coupling-ratios.

The Systematic Location of Genes by Means of Crossover Observations[1]

R. A. Fisher
Rothamsted Experimental Station

1. INTRODUCTORY

In the construction of a chromosome map, the distances between neighboring genes are equated to the percentage of crossovers which have been observed between them. Owing to errors of random sampling, and some times to other disturbing causes, inconsistencies always arise between the distances so determined. For example, in the important data given by Lancefield and Metz for the sex chromosome of *Drosophila willistoni* [1, p. 241] we have the following values:

TABLE I

	Crossover percentage	Number of observations	Number of crossovers
Scute to Beaded	1.43	279	4
Beaded to Rough	2.42	455	11
Scute to Rough	7.09	6388	453

Within such a small range, double crossing over may be ignored; yet it would be wrong to use such inconsistencies as an argument against the linear arrangement of the genes. For although the true crossover values may be accurately additive, errors of random sampling will certainly disturb the observed percentages. The practical problem is to assign to the distances between the genes values which shall be as far as possible in accord with the whole of the observations available. In other words, we have to make use of as much as practicable, ideally the whole, of the information supplied by the data; giving due weight (i) to the greater accuracy of the values obtained from the larger number of observations, (ii) to the greater accuracy of values obtained from closer pairs. In general, too, we shall have to consider not three genes only, but a large number, lying sufficiently close together for double crossing over to be ignored, the percentage observed between each pair of which gives indirect information as to the position of all the others.

In its general character the problem resembles those problems involving errors of observation, where a smaller number of unknowns are determined from a larger number of inconsistent equations, and which are usually solved by the method of least squares. The practical solution depends on the construction of a number of "normal

[1] Reprinted from *American Naturalist* **56** 406–411 (1922).

equations" for the unknowns, in which the inconsistencies of the data are properly weighted and made to balance. To make the sum of the squares of the errors of the crossover percentages a minimum would, however, be wrong, and the method of least squares is not directly applicable. It has been shown that the whole of the information supplied by the data (2) is made use of by the method of maximum likelihood, and by a first approximation the required normal equations may be constructed.

2. MATHEMATICAL THEORY

In the above example, if we write p_1 and p_2 for the two adjacent crossover ratios, the probability of the actual series of observations will be proportional to

$$p_1^4(1-p_1)^{275}p_2^{11}(1-p_2)^{444}(p_1+p_2)^{453}(1-p_1-p_2)^{5935}$$

and the likelihood of any given pair of values for p_1 and p_2 will be proportional to the same quantity. In order to make this quantity a maximum for variations of p_1 and p_2, we have the equations

$$\frac{4}{p_1}-\frac{275}{1-p_1}+\frac{453}{p_1+p_2}-\frac{5935}{1-p_1-p_2}=0,$$
$$\frac{453}{p_1+p_2}-\frac{5935}{1-p_1-p_2}+\frac{11}{p_2}-\frac{444}{1-p_2}=0.$$

These equations are exact, but for practical purposes we need equations linear in p_1 and p_2, and a first approximation is sufficient; if p differs little from $x/(x+y)=x/n$, then

$$\begin{aligned}\frac{x}{p}-\frac{y}{1-p}&=0-\left(\frac{x}{p^2}+\frac{y}{(1-p)^2}\right)\left(p-\frac{x}{n}\right)+\cdots\\&=-\frac{n^3}{xy}\,p+\frac{n^2}{y}.\end{aligned}$$

So that we may rewrite equations (1) in the practical and approximate form

$$\begin{aligned}&\frac{279^3}{4\times 275}p_1+\frac{6388^3}{453\times 5935}(p_1+p_2)\\&\quad=\frac{279^2}{275}+\frac{6388^2}{5935},\\&\frac{6388^3}{453\times 5935}(p_1+p_2)+\frac{455^3}{11\times 444}p_2\\&\quad=\frac{6388^2}{5935}+\frac{455^2}{444}.\end{aligned}$$

For each percentage observation, therefore, we have merely to calculate the two quantities n^3/xy and n^2/y; then normal equations may be constructed in the form

$$\begin{aligned}&a_{11}p_1+a_{12}p_2+\cdots=b_1,\\&a_{12}p_1+a_{22}p_2+\cdots=b_2,\\&\cdots\cdots\cdots\cdots\cdots\cdots\end{aligned}$$

where a_{12} is the sum of the quantities n^3/xy for which both p_1 and p_2 are involved, a_{11} the corresponding sum for all in which p_1 is involved, and b_1 the sum of the quantities n^2/y for which p_1 is involved.

3. PRACTICAL EXAMPLE

In order to illustrate the practical application of this method to a complex case, we will consider the location of the 8 genes, from Reduced to Rimmed, in the middle of the sex chromosome of *Drosophila willistoni.* We have here 7 intervals to determine, and fifteen crossover percentages are given [1]. Table II shows the data, and the series of weighting quantities derived from them.

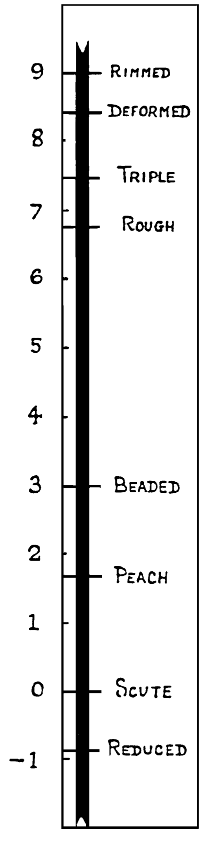


TABLE II

	Percentage	x	n	n^2/y	n^3/xy	Unknowns involved
Reduced–Scute	.95	27	2,848	2,875.26	303,287	p_1
Reduced–Rough	6.24	37	593	632.46	10,136	p_1, p_2, p_3, p_4
Scute–Peach	1.81	8	442	450.15	24,871	p_2
Scute–Beaded	1.43	4	279	283.06	19,742	p_2, p_3
Scute–Rough	7.09	453	6,388	6,875.58	96,956	p_2, p_3, p_4
Scute–Deformed	7.24	50	691	744.90	10,295	p_2, p_3, p_4, p_5, p_6
Scute–Rimmed	9.91	189	1,908	2,117.78	21,379	$p_2, p_3, p_4, p_5, p_6, p_7$
Peach–Beaded	1.70	3	176	179.05	10,504	p_3
Peach–Rough	5.05	33	654	688.75	13,650	p_3, p_4
Beaded–Rough	2.42	11	455	466.27	19,287	p_4
Rough–Triple	.49	4	809	813.02	164,433	p_5
Rough–Deformed	2.39	12	503	515.29	21,599	p_5, p_6
Rough–Rimmed	2.26	62	2,742	2,805.43	124,072	p_5, p_6, p_7
Triple–Rimmed	1.00	6	601	607.06	60,807	p_6, p_7
Deformed–Rimmed	4.17	2	48	50.09	1,202	p_7

From this table we write down the normal equations

$$313{,}423p_1 + 10{,}136(p_2 + p_3 + p_4) = 3{,}507.72,$$

$$10{,}136p_1 + 183{,}380p_2 + 158{,}509p_3 + 138{,}766p_4 + 31{,}674p_5 + 31{,}674p_6 + 21{,}379p_7 = 11{,}103.93,$$

$$10{,}136p_1 + 158{,}509p_2 + 182{,}663p_3 + 152{,}416p_4 + 31{,}674p_5 + 31{,}674p_6 + 21{,}379p_7 = 11{,}521.58,$$

$$10{,}136p_1 + 138{,}766p_2 + 152{,}416p_3 + 171{,}703p_4 + 31{,}674p_5 + 31{,}674p_6 + 21{,}379p_7 = 11{,}525.74,$$

$$31{,}674(p_2 + p_3 + p_4) + 341{,}778p_5 + 177{,}345p_6 + 145{,}451p_7 = 6{,}996.42,$$

$$31{,}674(p_2 + p_3 + p_4) + 177{,}345p_5 + 238{,}152p_6 + 206{,}258p_7 = 6{,}790.46,$$

$$21{,}379(p_2 + p_3 + p_4) + 145{,}451p_5 + 206{,}258p_6 + 217{,}460p_7 = 5{,}580.36.$$

Using a calculating machine, the work so far is rapid and mechanical; the solution of the normal equations may in this case be much simplified by observing the uniformity of some of the sets of coefficients, a type of uniformity which is probably characteristic of crossover data. Thus by considering $(p_2 + p_3 + p_4)$ as a single quantity, p_1 is immediately expressible in terms of it, and by solving the last three equations we may do the same for p_5, p_6 and p_7; substituting finally in equations (2), (3), (4) we solve them for p_2, p_3 and p_4, and obtain the values shown in Table III.

The seven values obtained give mutually consistent values for the crossover percentages between the fifteen pairs tested, and are therefore suitable

TABLE III

	Calculated	Observed	Difference d	Standard error σ	$\dfrac{d^2}{\sigma^2}$
Reduced–Scute	.90 p_1	.95	+.05	.18	.08
Reduced–Rough	7.66	6.24	−1.42	1.09	1.70
Scute–Peach	1.67 p_2	1.81	+.14	.61	.05
Scute–Beaded	2.98	1.43	−1.53	1.02	2.31
Scute–Rough	6.76	7.09	+.33	.31	1.13
Scute–Deformed	8.40	7.24	−1.16	1.06	1.20
Scute–Rimmed	8.97	9.91	+.94	.65	2.09
Peach–Beaded	1.31 p_3	1.70	−.39	.86	.21
Peach–Rough	5.09	5.05	−.04	.86	.00
Beaded–Rough	3.78 p_4	2.42	−1.36	.89	2.34
Rough–Triple	.69 p_5	.49	−.20	.29	.48
Rough–Deformed	1.64	2.39	+.75	.57	1.73
Rough–Rimmed	2.21	2.26	−.05	.28	.03
Triple–Rimmed	1.52	1.00	−.52	.50	1.08
Deformed–Rimmed	.57 p_7	4.17	+3.60	1.09	10.91
					$\chi^2 = 25.34$

for the construction of chromosome map. If the conditions of Maximum Likelihood had been exactly fulfilled they would agree better than any other consistent series of values with the percentages observed. As it is, it is only in the aberrant value of p_7 that the assumption that the observed values are approximately correct breaks down, and it is probable that such cases will only occur when the data are admittedly insufficient.

Table III is arranged to compare the differences between the calculated and the observed percentages with the standard errors due to sampling; except for p_7 all the differences are less than twice their standard errors; thus showing the general agreement between the data and the theory of linear arrangement of the genes. The fit, however, is not a close one, even if we omit p_7; in the present state of our knowledge this will not throw any doubt on the scheme of linear arrangement, but will suggest that the crossover ratios in this part of the chromosome were not constant in all the strains used to compile the data.

In estimating the Goodness of Fit of data of this kind, χ^2 may be calculated by summing the values of d^2/σ^2, as in Table III. Attention should, however, be called to the fact that it has been recently shown (3) that in entering Elderton's Table we must put n'equal to one more than the number of degrees of freedom, remaining after we have fitted our unknowns to the data. In the present case we have found 7 unknowns from 15 equations, leaving 8 degrees of freedom, so that n' should be 9, and not 16.

In conclusion it should be noted that to be available for the use of this process the crossover data should be stated in the form in which it is given by Lancefield and Metz, in which the crossovers tabled between any two genes do not include those experiments in which an intermediate gene was under observation. The practice of throwing together all the crossovers between two genes, in order to improve the ratios between the more distant points, causes the same crossover to appear repeatedly in different entries. The data are no longer the product of independent experiments, and must be re-summarized before reduction.

REFERENCES

1. R. C. Lancefield and C. W. Metz. The Sex-Linked Group of Mutant Characters in *Drosophila willistoni. American Naturalist*, LVI, pp. 211–241.
2. R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans. A*, CCXXII, pp. 309–368.
3. R. A. Fisher. On the Significance of χ' from Contingency Tables and on the Calculation of P. *Journal of Royal Statistical Society*, LXXXV, pp. 87–94.