Genetics Education

Innovations in Teaching and Learning Genetics

Edited by Patricia J. Pukkila

Do-It-Yourself Statistics: A Computer-Assisted Likelihood Approach to Analysis of Data From Genetic Crosses

Leonard G. Robbins

Dipartimento di Biologia Evolutiva, Università di Siena, 53100 Siena, Italy and Genetics Program and Department of Zoology, Michigan State University, East Lansing, Michigan 48824-1312

ABSTRACT

Graduate school programs in genetics have become so full that courses in statistics have often been eliminated. In addition, typical introductory statistics courses for the "statistics user" rather than the nascent statistician are laden with methods for analysis of measured variables while genetic data are most often discrete numbers. These courses are often seen by students and genetics professors alike as largely irrelevant cookbook courses. The powerful methods of likelihood analysis, although commonly employed in human genetics, are much less often used in other areas of genetics, even though current computational tools make this approach readily accessible. This article introduces the MLIKELY.PAS computer program and the logic of do-it-yourself maximum-likelihood statistics. The program itself, course materials, and expanded discussions of some examples that are only summarized here are available at http://www.unisi.it/ricerca/dip/bio_evol/sitomlikely/mlikely.html.

As most of us still impress on our introductory genetics students, genetics started with the counting of offspring produced by crosses. Although many of us now spend a large fraction of our time at a chemical bench, crosses, and the discrete data they generate, still remain a core tool in our work. Remarkably, the early synergism between genetics and statistics is now mostly absent from the pages of this journal. Virtually all of us are familiar with log of odds (LOD) score analysis of human genetic data, and most of us can do a χ^2 test against *a priori* expectations or a χ^2 contingency test. Nevertheless, in most articles in Genetics that contain cross data there is either no statistical analysis at all or transformation of the discrete data to frequencies—accompanied by confidence intervals and statistics, if any, that were originally devised for dealing with continuous variables. The questions we are actually interested in asking go well beyond the few methods of discrete-data analysis we've learned, so we either rely on an eyeball approach or fall back on the continuous-variable methods taught in the usual statistics courses. Given the loss of power and plethora of mathematical and inferential pitfalls that the latter entails, the eyeball may often be the better instrument.

This need not be the case, and in many areas of research this is not the case. Areas as diverse as animal behavior, clinical trials, and signal and image processing are replete with powerful examples of discrete analysis. (The CCAR database, for example, contains nearly 3000 entries for "maximum-likelihood" for the 3 years 1993–1995.) There are, of course, examples in genetics as well: LOD scores (see Crow 1993 and Morton 1995 for historical views of human gene mapping); sporadically appearing, but cumulatively numerous, applications of likelihood methods to problems in formal genetics (a far from exhaustive sample includes Kastenbaum 1958; Sandler and Kastenbaum 1958; Robbins 1971, 1977, 1999; Snow 1979; King and Mortimer 1991; Lyckegaard and Clark 1991; Hilliker *et al.* 1994; McPeeck and Speed 1995; Zhao *et al.* 1995a,b); and widespread application by the mathematical sophisticates of population genetics, quantitative genetics, and numerical taxonomy. Weir (1994, 1995) has also made a convincing case for the use of likelihood ratios in forensics, the newest area of applied population genetics. Yet the general picture in formal genetics may be seen with a simple count of a random issue of Genetics (May 1992; vol. 131, no. 4). Excluding population and quantitative genetics articles, there were 13 reports containing discrete

Address for correspondence: Dipartimento di Biologia Evolutiva, Università di Siena, Via P. A. Mattioli 4, 53100 Siena, Italy.
E-mail: robbins@unisi.it

data. Of the 13, only 2 (significantly $<\frac{1}{2}$; $\chi^2 = 8.07$, 1 d.f., $P \sim 0.004$) included any statistical analysis at all. Obviously, statistical tests may not have been needed in these articles; simple perusal of percentages can often be convincing. Nevertheless, methods for analysis of discrete data are available, their use is not difficult, and they can be revealing. They can also help us avoid designing complicated crosses that cannot, in the end, be analyzed.

Methods for discrete analysis have been a step-sister in statistics, but discrete multivariate analysis, with a thorough exegesis in Bishop *et al.*'s (1975a) classic book allows the same rigorous approach to discrete data that conventional analysis of variance provides for measurement data. Bishop *et al.* (1975a), however, is written for statisticians and can be intimidating. Perhaps that is why these methods have not found their way into most areas of genetics even though they are commonly used in wealthier fields such as clinical trials where professional statisticians are routinely members of the team, and analytical power must be kept high to keep the number of human subjects low. For analysis of crosses, however, the full-blown artistry of discrete multivariate analysis is not usually needed.

The value of a multivariate mode of thinking is well illustrated by the erroneous presentation of the *a priori* χ^2 test in a popular introductory genetics textbook (Griffiths *et al.* 1993; awkwardly corrected in Griffiths *et al.* 1999). In their example, data from a test cross with an observed recombinant fraction under 50% are used to test for linkage. Instead of the appropriate test of a 1:1 ratio of parental:nonparental, however, Griffiths *et al.* (1993) tests for a 1:1:1:1 ratio among all four products of the test cross. Unfortunately, this is a test for Mendelian independent assortment and not a test of whether the recombinant fraction is statistically different from 50%. This test compounds testing for linkage and testing for equal recovery of reciprocal products; two variables that really need to be separated. They assert that the sample data do not support linkage ($\chi^2 = 5.2$, 3 d.f., $P = 0.156$), but done correctly there is, in fact, a significant indication of linkage ($\chi^2 = 5.0$, 1 d.f., $P = 0.025$), while there are no significant differences (for example, there are no significant marker viability effects) in recovery of reciprocal products ($\chi^2 = 0.202$, 2 d.f., $P = 0.904$). In the new book, a contingency test of statistically independent recovery of the allelic combinations, in place of the *a priori* test for Mendelian independent assortment, yields a result ($\chi^2 = 5.02$, 1 d.f., $P = 0.025$) very close to that of the simpler test for deviation from a 1:1 ratio of parental:nonparental.

The foregoing example does not illustrate a need for analysis of maximum likelihood, nor even for the contingency test used in Griffiths *et al.* (1999); if the question had been correctly posed, a simple *a priori* χ^2 test would have been adequate. It does, however, illustrate how failing to separate different biological processes can lead one astray. Testing for linkage, a recombination fraction under 50%, is not the same as testing for all possible distortions of genotype frequencies.

Not every problem in formal genetics can be resolved just with clear thinking and χ^2 tests. The advent of powerful personal computers, however, makes the methods that are needed accessible to those who, like myself, are neither mathematicians, statisticians, nor professional programmers. Several of the most common questions geneticists must contend with can be asked: What are the best estimates of genetic parameters? Does a hypothesis adequately account for the observed effects? How can we test whether an experiment and control respond differently to a variable we're interested in, when both the experiment and control are also affected by some other variable? Is there significant variation in what we're scoring? Is there a correlation between two variables? How important is the correlation? Moreover, with use of the computer allowing us to strip away much of the mathematical complexity, the major task left for the geneticist is the clear definition of the question(s) to be asked.

With the hope of creating an enhanced awareness of these methods among the next generation of geneticists, a set of real-world examples and a program for numerical approximation of maximum likelihoods were used as the core of a graduate-level course offered first in 1996 at Michigan State University and again in 1998 at the University of Siena. The course presented a guide to this mode of analysis by means of examples, some already published and some new. In each case, I chose actual experiments rather than invented examples. For some of the examples, the genetics is nontrivial and the explanation of the crosses is lengthy, but this allows the student or reader to judge the value and difficulty of applying this method to real-world situations. The examples are as follows: (1) mapping a dominant of reduced penetrance; (2) testing for a correlation between two chromosome-behavior phenotypes; (3) testing whether a meiotic mutant affects chiasma interference; and (4) testing for the effects of a gene on viability in the presence of confounding variables. The first two examples are covered here, while the latter two are only briefly described with the full discussions included at the web site.

BASIC METHODOLOGY

Maximum-likelihood estimates and hypothesis testing: In the following sections, computer-assisted techniques for estimating parameters (such as map distances), testing for goodness-of-fit, and comparing hypotheses are described. All of them are based on the method of maximum likelihood (Fisher 1922; Edwards 1992).

Suppose that the probability of getting an offspring of a given class is p and that N of these offspring were observed in an experiment. The likelihood of getting those N offspring is defined as: $L = p^N$. Crosses yield multiple offspring classes, each with its own probability, but, because different offspring are independently produced, the likelihood for the entire experiment is the product of the likelihoods. For example,[1] a test cross involving two genes that are $\overline{ab}$ map units apart yields CO crossover offspring and NCO noncrossover offspring, and the likelihood is $L = \overline{ab}^{\mathrm{CO}}\ (1 - \overline{ab})^{\mathrm{NCO}}$. The value of $\overline{ab}$ that maximizes L is the estimate of $\overline{ab}$ that we use; in this case it is CO/(NCO + CO). If we were dealing with a more complex situation where there are many parameters, we would want to find the values of all of the parameters that simultaneously maximize L. Noting that as a number increases, its logarithm increases as well, we can, with the same effect and usually more easily, find the parameter values that maximize the logarithm of L. In the following, the maximum values of these functions are denoted $\hat{L}$ and $\ln \hat{L}$.

Most of the time we are not only interested in estimating the parameters, but in testing whether a hypothesis provides a sufficient explanation for the experimental variation or in testing whether there are significant differences between two (or more) hypotheses. For example, if we suspect a correlation between two variables, we would want to test three things. First, we need to test whether there is significant variation in these parameters in the first place. That is, does a hypothesis of no variation in one or the other parameter fail a goodness-of-fit test? Second, we will want to know whether a model that includes a correlation with slope other than zero is significantly better than the hypothesis of no variation (equivalent to a correlation with slope = 0). That is, we must compare two hypotheses. Third, we will want to know how much of the variation is explained by the correlation, *i.e.*, its sufficiency; another goodness-of-fit test.

In many cases, the obvious test for goodness-of-fit is a straightforward χ^2. That is, we use the maximum-likelihood estimates of the parameters to find the probabilities of each class, multiply these probabilities by the appropriate total(s) to get expected numbers, and calculate χ^2 as a measure of the difference between observations and expectations. The degrees of freedom are then the number of independent observations less the number of parameters estimated from the data.

When we wish to compare two hypotheses, H1 and H2, however, a different measure is often more appropriate or more convenient. This is the G (also known as G^2) statistic (Bishop *et al.* 1975b, Chapter 4):

$$G = 2 \times (\ln \hat{L}_{\mathrm{H1}} - \ln \hat{L}_{\mathrm{H2}}).$$

G is distributed approximately as χ^2 with degrees of freedom equal to the difference between the numbers of parameters of the two hypotheses. The approximation to χ^2 is asymptotic and becomes more exact as sample size increases.

Note that in many cases a test for sufficiency, usually stated as a test for goodness-of-fit, can also be described as a comparison of two hypotheses. For example, if H1 includes m parameters (unknowns to be solved for) and there are m independent observations (knowns), and the parameters can take any numerical value, the maximum-likelihood estimates of the parameters are identical to what would be obtained by solving m equations in m unknowns. Because H1 merely describes all of the variation, there is no test for its sufficiency (aside from the possibility of getting utterly absurd parameter values) and calculating χ^2 will yield a value of 0. A comparison of another hypothesis, H2, to H1 by a G test is then logically equivalent to testing H2 for goodness-of-fit. The values of G for H2 *vs.* H1 and the χ^2 for goodness-of-fit of H2 will generally be the same, or very nearly so. For such tests, the choice of whether it is done as a χ^2 or G test is largely a matter of convenience (if, for example, the values of $\ln \hat{L}_{\mathrm{H1}}$ and $\ln \hat{L}_{\mathrm{H2}}$ have already been found), esthetics, or habit.

How can we find the parameter values that maximize $\ln L$? For pedigree data, in years past we would have gone to Morton's (1955) tables, but we would now most likely use one of the readily available LOD score computer programs (Terwilliger 1994). In some other situations, we might also be able to turn to the literature for an analytical solution. If the crosses do not correspond to an already worked-out situation, but we are skilled in the calculus and linear algebra, we might try to find the partial derivatives of $\ln L$ with respect to each parameter, set them equal to zero, and solve the set of simultaneous equations. Failing that, and even a skilled mathematician sometimes will, we can turn to a computer to approximate the maximum by numerical methods. Indeed, if we are willing to travel this less elegant route, all we need to tell the computer is the probability for each offspring class, and the computer can do the rest.

The MLIKELY.PAS program: MLIKELY.PAS is a Pascal program that, in its current version, is compiled under TURBO PASCAL 4.0 (Borland Intl.). It is not user friendly—it lacks a graphical user interface, does not support a mouse, and requires that the user convert a few equations into Pascal syntax and paste them into the program, which then must be compiled and run. It is, however, geneticist friendly. It can work with virtually any set of crosses, whether simple or complex. The expressions the user needs to write are most often direct translations of a Punnett square or logic tree. And running the program requires only answering a series of questions and entering the data. The heuristic used is brutally simple; the user provides first guesses of the

[1] With apologies to the mathematically oriented, who might read NCO as N times C times O, common genetic abbreviations are used to name variables rather than following mathematical convention. Hence, NCO is to be read as "noncrossovers."

parameter values or accepts the program's defaults, and the computer increases and decreases those values, moving sequentially through the list of parameters using ever smaller intervals, until it finds the maximum of ln L to whatever precision is desired. A few tricks are used to speed operation:

1. The likelihood surface may be smooth in some areas and rough in others. Where rough, large increments may miss a peak. Where smooth, however, large increments are more efficient. Hence, if the iteration process continues for several rounds at a given increment, the interval changes to a larger value.
2. To even out the sensitivity of parameters that are very small and very large, the increments are made as fractions of the previous guess (as long as that prior guess was not exactly zero).
3. To ensure that a path through the likelihood space can never be retraced, the proportions by which parameters are increased and decreased are not the same but are relatively prime.
4. The user can specify limited ranges for the parameters (for example, it makes no sense to try crossover frequencies outside the range 0 to 0.5).
5. The size of the multipliers, and the number of cycles at an increment before reverting to a larger one, were optimized for a problem somewhat more complex than any reported here.

Although incorporated piecemeal in MLIKELY.PAS either intuitively or empirically, these procedures are not uncommon in optimization algorithms, and constraining parameter ranges is similar to the use of "hints" in speeding artificial intelligence schemes. MLIKELY.PAS provides output in a variety of formats: screen-readable, printable, and word-processor and spreadsheet importable, and saves the data in a reusable file so that they need be entered only once.

There are algorithms that can find a maximum more quickly. For example, the "optimizer" found in the QUATTRO PRO (Corel) spreadsheet package can estimate the derivatives first to speed the search for a maximum. MLIKELY.PAS is not, in any case, unreasonably slow. Iteration times are indicated in the examples that follow, in each case for runs on a 80486/33 computer with each parameter estimated to a precision of better than 1 part in 10^8. Even with a less-than-state-of-the-art PC, the running time is most often far less than the thinking time needed to define the problem in the first place.

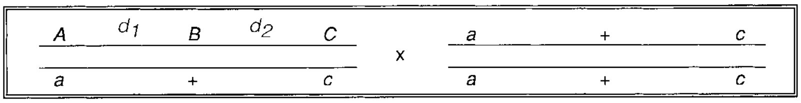
The simple heuristic used in MLIKELY.PAS can cause two problems that the user should be aware of. First, because the parameters are handled sequentially, if two or more of the starting guesses are impossible, *i.e.*, if they give negative expected frequencies, the program will not find the maximum, but will issue a warning message. Starting with more reasonable guesses is the cure for this problem. Second, a likelihood function may have more than one peak. As with other iterative peak-finding procedures, once in the neighborhood of a peak, even if it is not the highest in the entire landscape, the program may halt at that local maximum. It is even possible to have a model so badly structured that ln L is an oscillating function, such as a sine wave, but this is unlikely in any genetics application. A program designed to find likelihoods for only one class of problem can usually be rigged to avoid this. In contrast, although MLIKELY.PAS can be fooled by local maxima, and is not usable for every type of application to which maximum likelihood analysis applies, it can accommodate any model for which one can write the probabilities of getting each observed class.

That multiple peaks in an iterative process can be dangerous has certainly been seen in the study of human molecular evolution; the primacy of a mitochondrial Eve, while appealing, was supported by a maximum-parsimony tree that was not unique (Hedges *et al.* 1992; Templeton 1992). In more than 20 years of using MLIKELY.PAS and its ancestors, however, there has never been a false-peak problem except when I made a gross mistake in writing the probabilities in the first place, set absurd bounds for the parameters, or, more often, made a typographic error in putting them into the program.

The generally good behavior of the iteration algorithm used in MLIKELY.PAS could be a result of mere luck, but has probably occurred because formal genetics problems, as opposed to problems in taxonomy, are often well structured even when they involve many parameters. For example, in describing recombination in several regions, there will be several single-crossover frequencies to be estimated, but all of them behave in an algebraically similar fashion.

The behavior of MLIKELY.PAS during the iteration process as well as its output can provide useful indications of potential problems. For example, in the second example of this report, which considers testing for correlation using discrete data, an example of the effect of improperly bounding a parameter's search space is considered. It is nevertheless good practice to start with several widely different sets of parameter guesses to check that you always end up at the same peak.

MLIKELY.PAS also calculates χ^2 for a goodness-of-fit test of the hypothesis. The user must supply Pascal statements defining the sum(s) by which to multiply the probabilities to get expected numbers. For data from a single cross, this is simply the sum of all observations and that variable is already calculated by the program, but for a series of crosses the sum for each cross must be specified. Because the χ^2 calculation is included in MLIKELY, it is often more convenient to use this test of goodness-of-fit rather than an equivalent G test when only a single hypothesis is being tested; only a single set of equations need be written. In contrast, the likelihood-ratio approach of a contrast between the hypothesis of



ALGEBRAIC DESCRIPTION	PASCAL TRANSLATION
	VAR
	d1, d2, C, P, NCO, SCO1, SCO2, DCO :REAL;
	I :INTEGER;
	BEGIN
Note: d_1 & d_2 = map distances	d1 := Par[1];
C = coefficient of coincidence	d2 := Par[2];
P = penetrance	C := Par[3];
	P := Par[4];
$DCO = d_1 d_2 C$	DCO := d1*d2*C;
$SCO_1 = d_1 - DCO$	SCO1 := d1 - DCO;
$SCO_2 = d_2 - DCO$	SCO2 := d2 - DCO;
$NCO = 1 - SCO_1 - SCO_2 - DCO$	NCO := 1 - SCO1 - SCO2 - DCO;
A B C = ½ $NCO \cdot P$	ExpFr[1] := 0.5*NCO*P;
a + c = ½ NCO + ½ $DCO \cdot (1 - P)$	ExpFr[2] := 0.5*NCO + 0.5*DCO*(1-P);
A + c = ½ SCO_1 + ½ $SCO_2 \cdot (1 - P)$	ExpFr[3] := 0.5*SCO1 + 0.5*SCO2*(1-P);
a B C = ½ $SCO_1 \cdot P$	ExpFr[4] := 0.5*SCO1*P;
A B c = ½ $SCO_2 \cdot P$	ExpFr[5] := 0.5*SCO2*P;
a + C = ½ SCO_2 + ½ $SCO_1 \cdot (1 - P)$	ExpFr[6] := 0.5*SCO2 + 0.5*SCO1*(1-P);
A + C = ½ DCO + ½ $NCO \cdot (1 - P)$	ExpFr[7] := 0.5*DCO + 0.5*NCO*(1-P);
a B c = ½ $DCO \cdot P$	ExpFr[8] := 0.5*DCO*P;
	FOR I := 1 TO 8 DO ExpNo[I] := ExpFr[I]*SumObs;
	END;

Figure 1.—Mapping a dominant of reduced penetrance. (Top) Three genes are followed in a test cross. *A* and *C* are RFLP markers, while *B* (*Sp*^{d}; Asher *et al.* 1996) is a dominant mutation of reduced penetrance. (Bottom) Conventional genetic notation describing this cross and a Pascal translation. There are four parameters: two distances (expressed as recombination fractions rather than centimorgans for calculation purposes), one coefficient of coincidence, and one penetrance. Because of reduced penetrance, individuals of different genotypes can have the same phenotype. For example, the *a* + *c* phenotypic class includes both *a*/*a* +/+ *c*/*c* genotype individuals, = 1/2NCO, and individuals who are genotypically *a*/*a B*/+ *c*/*c* but are nonpenetrant for *B*, = 1/2DCO(1 − *P*). The Pascal version is inserted in MLIKELY.PAS, which is then compiled and run.

interest and a foil that explains all of the variation requires writing (or editing) two versions of the equations, compiling and running the program twice, and then calculating *G*.

Inclusion of these calculations in MLIKELY serves another purpose as well; seeing that the sum of the expected numbers equals the sum of the observed numbers. Moreover, examination of the listing of the χ^2 values of the individual cells gives a good check that the probabilities have been sensibly defined and accurately entered.

This article includes only enough information about the structure and running of the program to permit understanding how it serves the geneticist. MLIKELY (including all source code), sample sets of equations and data, as well as documentation files are available at the web site. Downloading carries two conditions: (1) neither the program, nor any substantial part of the program, may be used for commercial purposes nor incorporated into another program without my written permission; and (2) any improvements made, or versions modified for other Pascal compilers, will be shared with me so that they can be incorporated in future releases.

EXAMPLES

Parameter estimation—mapping a mutant of reduced penetrance: The top of Figure 1 illustrates a mouse genetics problem recently faced by J. Asher. [This question arose in work following from Asher *et al.* (1996). Unfortunately, Dr. Asher died before the work could be completed.] In this cross, he wished to map mutation *B*, a dominant of reduced penetrance, with respect to two RFLP (and, therefore, codominant) markers. Meiosis produces noncrossovers, single crossovers, and double crossovers, but because *B* is not fully penetrant, some *B*-bearing progeny will be B$^+$ in phenotype. For example, some of the *A B C* noncrossovers may be recovered as A + C phenotype progeny equivalent to one of the double crossover classes. As shown in the bottom panel of Figure 1, writing equations for the probabilities of DCO, SCO, and NCO, adding the effect of reduced penetrance to get the probabilities of each of the progeny types, and translation of the algebraic description of this situation into Pascal syntax are straightforward. The Pascal version includes a preamble declaring the names of the variables that will be used, and defining mnemonic designators for distances (expressed as crossover frequencies), the coefficient of coincidence and penetrance in terms of the array of parameters provided in the program. It also includes a statement that finds the expected numbers for each class by multiplying the probabilities by the sum of the observations. The Pascal translation of the genetics is inserted into the MLIKELY. PAS program, which is then compiled and run. The input needed consists of the eight observations, which can be entered in response to questions posed by the program at run time, or can be taken from a data file written (in ASCII text format) in advance.

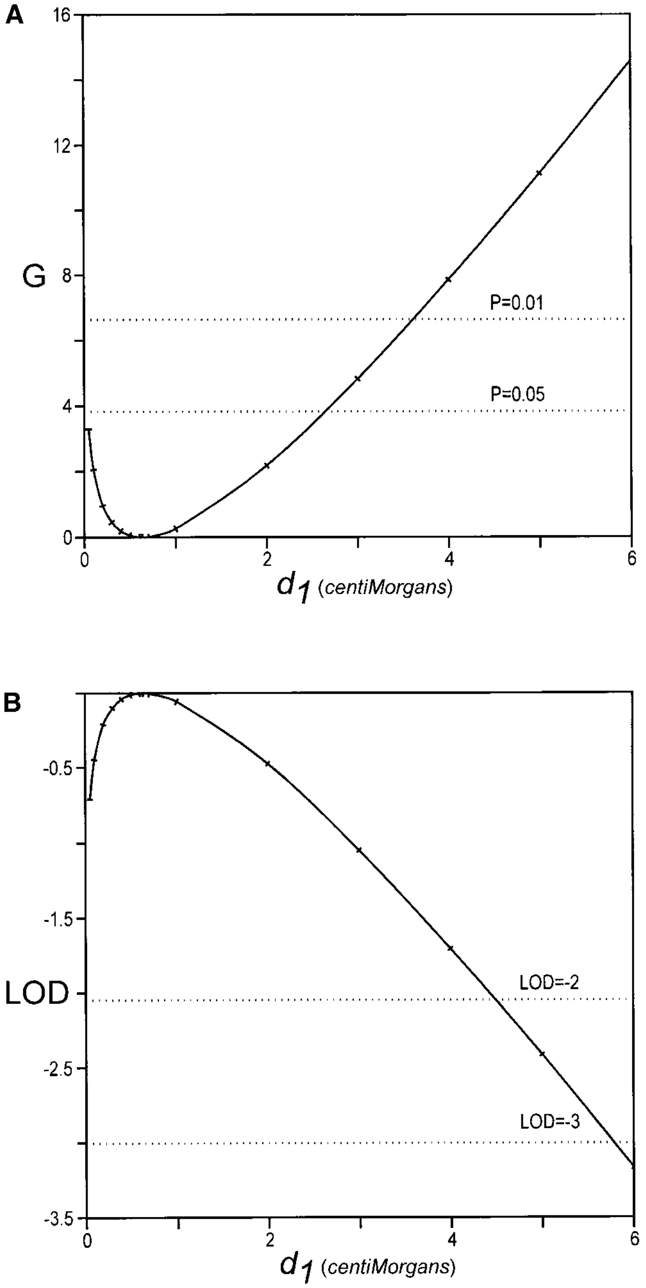


Figure 2.—Maximum-likelihood-derived confidence intervals for the distance between genes *A* and *B*. (A) *G*-test comparisons; (B) LOD score comparisons. The equations describing the cross shown in Figure 1 were changed so that distance d_1 was treated as a constant. Maximum likelihoods were obtained for a series of values of d_1 and $\ln \hat{L}$ for each of these fixed-d_1 hypotheses was compared to $\ln \hat{L}$ for the variable-d_1 hypothesis. The peak of the curve occurs at the estimate of d_1 obtained under the variable-d_1 hypothesis and the smallest and largest *A*-*B* map distances, in centimorgans, consistent with the data are those at which the curves cross the selected probability or LOD-score criterion (dotted lines). MLIKELY.PAS was used to find the $\ln \hat{L}$ values and the spreadsheet-compatible output file was imported into Quattro Pro, which was then used to calculate values of the *G* statistic and LOD scores. Graphs were prepared using Corel Draw; calculated points are shown by tick marks while the curves are Bezier interpolations.

Starting with some wild guesses ($d_1 = 0.1$, $d_2 = 0.1$, $C = 0.1$, and $P = 0.99$), in less than a second, the program finds the values of the two distances, the coefficient of coincidence, and the penetrance that maximize the ln likelihood of getting the observed results ($\hat{d}_1 = 0.00634$, $\hat{d}_2 = 0.0525$, $\hat{C} = 0.0$, and $\hat{P} = 0.6605$) and indicates, by the nonsignificant value of $\chi^2 = 3.547$ (3 d.f., $P = 0.315$), that this model provides a sufficient description of the data.

We can also take this a step further and examine the precision of these estimates. For example, we may be most interested in distance d_1, the short *A* to *B* interval. What is the largest, or smallest, estimate of this distance that is still consistent with the data? To do this[2], we (1) set a series of fixed values for d_1; (2) allow the program to find the values of the other parameters that maximize the likelihood; and then (3) compare the results with those for the maximum-likelihood estimate of d_1, *i.e.*, when $d_1 = 0.00634$.

We can use MLIKELY.PAS for the first two steps by changing just four lines of code, so that d_1 is treated as a constant,

```
d1 := Con[1];
d2 := Par[1];
C := Par[2];
P := Par[3];
```

and repeating the iteration several times for different values of Con[1]. The data can be reentered, or the original data file may be modified in any text editor to change the number of parameters from 4 to 3 and the number of constants from 0 to 1.

We then need a statistic that allows us to compare the results. Two related comparisons are shown in Figure 2, one using the *G* statistic (Figure 2A) and the other using LOD scores (Figure 2B). MLIKELY.PAS does not itself calculate either *G* or LOD scores, but both of those statistics are easy to calculate and graph using a spreadsheet, and MLIKELY.PAS does provide a spreadsheet-importable (comma and space delimited) output file.

The values of the maximum ln *L*'s were obtained with MLIKELY.PAS, and a spreadsheet program (QUATTRO PRO; Corel) was used to find $G = 2 \times [\ln \hat{L}_{(d_1 \text{ fixed})} - \ln \hat{L}_{(d_1 \text{ variable})}]$ for each fixed value of d_1. These results are shown in Figure 2A. There are four parameters when d_1 is allowed to vary, and three when it is fixed, giving 1 d.f., corresponding to $P = 0.05$ for $G = 3.841$ and $P = 0.01$ for $G = 6.635$. Thus, the 95% upper bound

[2] The procedure outlined here is decidedly inelegant and provides what is more properly termed a support interval rather than a conventional confidence interval, but it requires understanding only the basic concepts of statistical inference and does not require understanding variance and covariance nor knowledge of linear algebra. It is also practical.

TABLE 1

Progeny recovered from $\frac{Df(1)rJ1,\ Rex}{crossover} \times \frac{attached\text{-}XY}{O}$ crosses

Crossover	Regular female	Regular male	rDNA crossover (%)	Nondisjunctional (%)
y cv	193	179	8 (7.66)	14 (4.71)
y v f car	992	848	9 (1.78)	9 (0.66)
y cv v	339	234	10 (5.57)	5 (1.19)
y f car	983	672	13 (2.58)	16 (1.34)
y cv v f	69	83	6 (14.81)	5 (3.90)
y car	164	131	6 (6.83)	0 (0.00)
y v	91	90	1 (2.15)	2 (1.44)
y cv f car	148	142	3 (3.90)	3 (1.35)
y cv car	70	49	2 (5.41)	1 (1.16)
y f	22	17	1 (8.33)	1 (3.33)

A series of identical crosses were done of groups of females heterozygous for *Rex* and various marked *X* chromosomes that had different frequencies of an autosomal *Su(Rex)* in each group (L. G. Robbins, unpublished data). The crosses allowed detecting meiotic nondisjunction in the females and mitotic exchange between two rDNA arrays in their offspring.

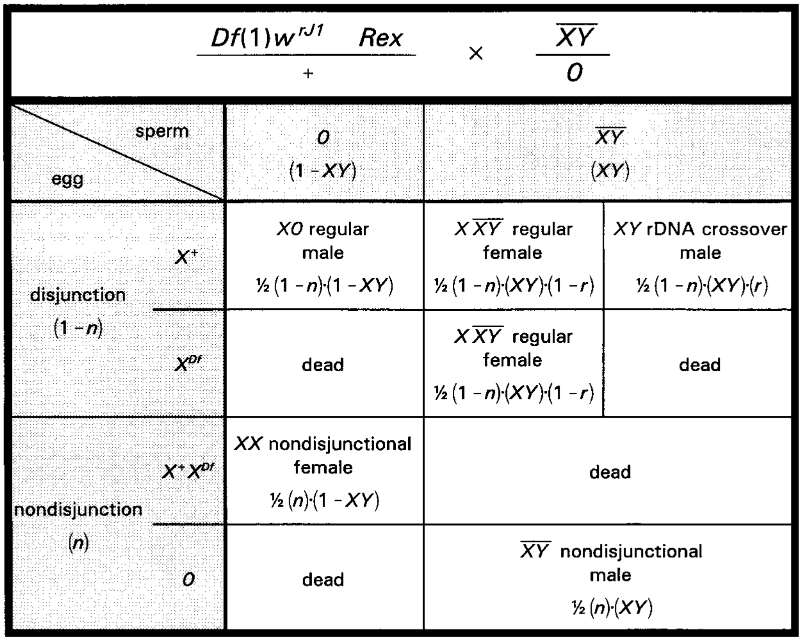
for d_1 is less than 3 map units, and the 99% upper bound is less than 4 map units. The probabilities provided by the *G* test correspond to those conventionally used in most hypothesis testing; they are the probabilities of getting a difference at least that large by chance alone.

A different comparison, shown in Figure 2B, is often used in human genetics. LOD (log of odds) scores are the $\log_{10}$ of the ratio of the likelihoods under two hypotheses, or, equivalently, the difference between the $\log_{10}$ *L*'s. The ln *L* output of MLIKELY.PAS can be converted to base 10 by multiplying by ln10 ≈ 2.30258, and the LOD scores are found by subtraction. Note that the conventions used in pedigree analysis, a LOD of +3 to demonstrate linkage and −2 to exclude linkage, are substantially more stringent than the usual critical values. This stringency is reasonable when dealing with the tests of multiple hypotheses implicit in using a progressive accumulation of families to decide whether there is linkage, but is overkill for most cross data. It is certainly overly stringent here, where we are already certain that the genes are linked.

Variation and correlation—the relationship between experimental variables: There are probably innumerable circumstances in which one observes two or more variable phenotypes and wants to know whether they are correlated. Where the phenotypes are metric, such as bristle lengths in Drosophila, conventional regression analysis can be appropriate, but regression analysis is also often used for counted variables, such as crossovers and disjunctional events, where a maximum-likelihood approach is more powerful and more revealing. To illustrate this, some unpublished data from my laboratory on the behavior of *Rex* are analyzed. The results of similar analyses may also be found in Palumbo *et al.* (1994), and some extensions to this approach are used in a recent article (Robbins 1999) that deals with sex-chromosome disjunction and meiotic drive produced by ribosomal-RNA gene deficiencies.

Rex is a repeated, heterochromatically located element of *Drosophila melanogaster*. Acting maternally, it promotes recombination between ribosomal-RNA gene arrays (rDNA) during early embryonic mitoses (Robbins 1981; Rasooly and Robbins 1991). We had repeatedly noted that crosses of *Rex* females also seem to produce more than the usual amount of sex-chromosome nondisjunction, amounting to ∼1% exceptions, and had wondered whether this is also an effect of *Rex*, or if it is an extraneous phenomenon unrelated to the presence of *Rex*. The frequency of nondisjunction, though elevated, is low enough that mapping it to *Rex* would be an uninviting task. If not *Rex*-related, this slight meiotic perturbation would also not be of much interest to us. Examination of data collected for other purposes, however, indicates that the frequencies of nondisjunction and rDNA recombination are correlated, suggesting that the two are functionally, even if not necessarily causally, related.

Those data came from crosses done along the way to mapping a suppressor of *Rex*, a *Su(Rex)*. At one point in this process, a series of chromosomes that carried different segments of the *X* chromosome were tested for suppression of *Rex* activity. As this particular *Su(Rex)* turned out to be autosomal, each genotype tested actually consisted of several flies bearing the same *X* segment, but a random sampling of *Su(Rex)* and non-*Su(Rex)* autosomes. The results of these crosses are shown in Table 1. Not only does the frequency of rDNA recombination appear to be (and is) heterogeneous because of the different frequencies of the *Su(Rex)* in the 10 samples, but the frequency of nondisjunction



$$\frac{Df(1)w^{rJ1}\quad Rex}{+} \times \frac{\overline{XY}}{0}$$

egg \ sperm		0 $(1 - XY)$	$\overline{XY}$ (XY)	
disjunction $(1 - n)$	X^+	XO regular male $\frac{1}{2}(1 - n)\cdot(1 - XY)$	$X\overline{XY}$ regular female $\frac{1}{2}(1 - n)\cdot(XY)\cdot(1 - r)$	XY rDNA crossover male $\frac{1}{2}(1 - n)\cdot(XY)\cdot(r)$
	X^{Df}	dead	$X\overline{XY}$ regular female $\frac{1}{2}(1 - n)\cdot(XY)\cdot(1 - r)$	dead
nondisjunction (n)	X^+X^{Df}	XX nondisjunctional female $\frac{1}{2}(n)\cdot(1 - XY)$	dead	
	0	dead	$\overline{XY}$ nondisjunctional male $\frac{1}{2}(n)\cdot(XY)$	

Figure 3.—Meiotic nondisjunction in *Rex*/+ females and mitotic exchange between two rDNA arrays in their offspring. Normal disjunction (1 − *n*) yields both *X*/*attached-XY* and *X*/*O* zygotes, but half of the latter die because they carry the lethal *rJ1* deficiency. A fraction (*r*) of the *X*/*attached-XY* zygotes are transformed to *X*/*Y* males or gynandromorphs by recombination between the two rDNA arrays of the *attached-XY*, but half of these also die because this exposes *Df(1)w*^{rJ1}. One-half of the products of nondisjunction also die because they are either *nullo-X* or metafemales.

varies as well. Are the two varying in a correlated fashion? We can find out by comparing the values of ln $\hat{L}$ under three hypotheses:

H1: The frequencies of both nondisjunction and *Rex*-induced exchange are different in each cross.
H2: The frequency of *Rex*-induced exchange differs among crosses, but the frequency of nondisjunction is the same in all 10 crosses.
H3: The frequencies of nondisjunction and *Rex*-induced exchange are related as nondisjunction = $m \times$ (rDNA exchange) + b. (Note that a linear correlation is considered here, but a correlation of any other form could be just as easily evaluated.)

There are three *G*-test comparisons to be made:

1. H1 "explains" all of the variation in the frequency of nondisjunction. H2 explains none of the variation in nondisjunction. Hence, the comparison of H1 *vs.* H2 tests whether there is statistically significant variation in the frequency of nondisjunction—it is equivalent to a goodness-of-fit test of H2.
2. H3 explains that part of the variation of nondisjunction that is linearly related to the frequency of *Rex*-induced exchange. H2 explains none of that variation. Hence, the comparison H3 *vs.* H2 is a measure of the variation explained by the correlation—it tests the significance of the correlation.
3. Last, H1 *vs.* H3 measures how much variation of nondisjunction is left unexplained after the relationship with *Rex*-induced exchange is accounted for. It tests the sufficiency of the correlation—it is equivalent to a goodness-of-fit test of H3.

The first step needed for making these comparisons is writing the probabilities of each of the progeny classes. Unfortunately, as illustrated in Figure 3, there are some complications caused by the actual cross used:

1. One of the *X* chromosomes of the *Rex* females also carried a deficiency that is recessive lethal. Thus, some offspring genotypes die because of the presence of the lethal.
2. The fathers carried an *attached-XY* ($\overline{XY}$) and therefore produce $\overline{XY}$ and *0* sperm, but the ratio of $\overline{XY}$:*0* sperm is not 1:1—$\overline{XY}$/*0* males produce an excess of *0* sperm.

3 Recoverable *Rex*-induced mitotic exchanges occur only in the $\overline{XY}$ embryos resulting from normal disjunction. The exchange product is an *X*/*Y* male (or gynandromorph), but if the *X* carries the lethal, it too dies. Thus, to completely describe each cross, we need parameters that describe (i) the frequency of nondisjunction(*n*); (ii) the frequency of *Rex*-induced exchange (*r*); and (iii) the proportion of sperm that carry the *attached-XY* (*XY*), and we must stay attentive to the classes that die.

The probabilities of the surviving genotypes among all zygotes are then

$$\text{Regular males} = \tfrac{1}{2}(1 - XY)(1 - n)$$

$$\text{Regular females} = (XY)(1 - n)(1 - r)$$

$$Rex\text{-induced mitotic exchanges} = \tfrac{1}{2}(XY)(1 - n)(r)$$

and

$$\text{Nondisjunctional males + females} = \tfrac{1}{2}n.$$

These are not, however, the probabilities of actually observing these offspring because we do not observe the lethals, which are $\frac{1}{2}(1 - XY)(1 - n) + \frac{1}{2}(XY)(r)(1 - n) + \frac{1}{2}n$. To get the probabilities among survivors, we must divide

Parameters: $XY = \dfrac{\overline{XY}\text{ sperm}}{\text{total sperm}}$ r = rDNA exchange rate n = nondisjunction rate

	Regular females	Regular males	rDNA recombinants	Nondisjunctional
Probability:	$XY(1-n)(1-r)$	$\frac{1}{2}(1-XY)(1-n)$	$\frac{1}{2}XY(1-n)r$	$\frac{1}{2}n$

```
VAR
  Cross, ObsGroup, ParGroup, I     :INTEGER;
  XY                               :REAL; {proportion XY sperm}
  n                                :REAL; {nondisjunction rate}
  r                                :REAL; {rDNA exchange rate}
  female                           :REAL; {probability of regular female}
  male                             :REAL; {probability of regular male}
  rDNAcrossover                    :REAL; {probability of rDNA crossover}
  nondisjunction                   :REAL; {probability of nondisjunctional offspring}
  survivor                         :REAL; {sum of probabilities of surviving offspring}
  total                            :INTEGER; {sum of observed in a cross}

BEGIN
 FOR Cross := 1 TO 10 DO BEGIN
  ObsGroup := 4*(Cross-1);

{H1: all parameters vary}      {H2: n same in all crosses}    {H3: n=m*r+b}
  ParGroup := 3*(Cross-1);       ParGroup := 2*(Cross-1);       ParGroup := 2*(Cross-1);
  XY := Par[1+ParGroup];         XY := Par[2+ParGroup];         XY := Par[3+ParGroup];
  r := Par[2+ParGroup];          r := Par[3+ParGroup];          r := Par[4+ParGroup];
  n := Par[3+ParGroup];          n := Par[1];                   n := Par[1]*r+Par[2];

  female := XY*(1-n)*(1-r);
  male := 0.5*(1-XY)*(1-n);
  rDNAcrossover := 0.5*XY*(1-n)*r;
  nondisjunction := 0.5*n;
  survivor := female+male+rDNAcrossover+nondisjunction;
  ExpFr[1+ObsGroup] := female/survivor;
  ExpFr[2+ObsGroup] := male/survivor;
  ExpFr[3+ObsGroup] := rDNAcrossover/survivor;
  ExpFr[4+ObsGroup] := nondisjunction/survivor;
  total := Obs[1+ObsGroup]+Obs[2+ObsGroup]+Obs[3+ObsGroup]+Obs[4+ObsGroup];
  FOR I := 1 TO 4 DO ExpNo[I+ObsGroup] :=ExpFr[I+ObsGroup]*total;
 END;
END;

Iteration times:
      49 seconds                   38 seconds                  2 minutes 15 seconds
```

Figure 4.—Correlation of two phenotypes associated with the *Rex* element of *Drosophila melanogaster*. (Top) The parameters used to describe this cross and the probabilities of the offspring types. Note that these are the probabilities among all zygotes, including those that are lethal, and do not sum to one. (Bottom) Pascal coding used to test for a correlation between the two phenotypes. Parameters are assigned in accord with three hypotheses: H1, that all parameters vary from cross to cross; H2, that the nondisjunction rate is the same in all crosses; and H3, that the nondisjunction rate is correlated with the rDNA exchange rate. Probabilities of each class among total zygotes are first calculated and then converted to expected fractions of each class among survivors by dividing by total surviving. Expected numbers are the expected fractions times the observed total for each cross. Iteration times for MLIKELY.PAS containing these equations are shown here, and the results are shown graphically in Figure 5.

the probability of each surviving genotype by the total probability of survival, $1 - \frac{1}{2}(1 - XY)(1 - n) - \frac{1}{2}(XY)(r)\ (1 - n) - \frac{1}{2}n$.

The equations needed to find the maximum-likelihood estimates of the parameters under the three models and the iteration times to find ln $\hat{L}$ are shown in their Pascal incarnation in Figure 4. Because only the parameter values change from cross to cross, a single set of equations is contained within a loop. Only four lines must be changed to accommodate each of the hypotheses.

Each cross yields four offspring classes, three of which are independent. Under H1, each cross is described by three separate parameters so there is a unique solution for each. They are

$$XY = (\text{regular females} + 2 \times \text{rDNA crossovers}) / (\text{regular females} + 2 \times \text{regular males} + 2 \times \text{rDNA crossovers}),$$

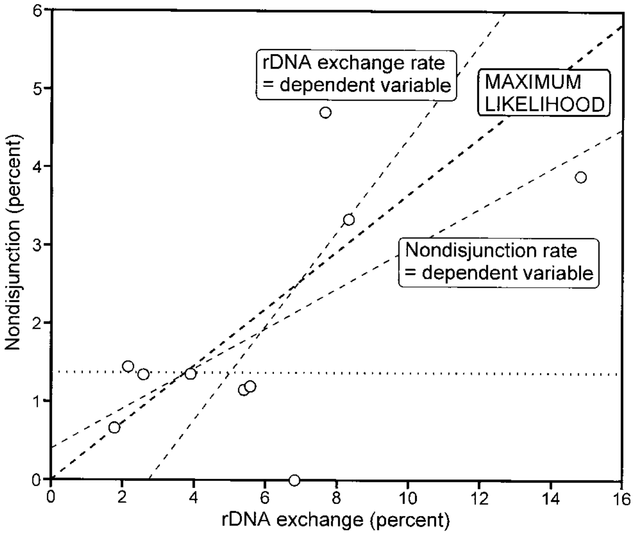
$$r = (2 \times \text{rDNA crossovers}) / (\text{regular females} + 2 \times \text{rDNA crossovers}),$$

and

$$n = (2 \times \text{nondisjunctional offspring}) / (\text{regular females} + 2 \times \text{regular males} + 2 \times \text{rDNA crossovers} + 2 \times \text{nondisjunctional offspring}).$$

If each of the three parameters is a probability with values between 0 and 1, MLIKELY.PAS must reach the same solutions, and the goodness-of-fit χ^2 at the end of the iteration process must be 0. Thus, even if the algebraic solutions for *XY*, *r*, or *n* were not reasonably obvious, MLIKELY.PAS would provide the solutions. In other words, whenever the number of parameters equals the number of independent observations, MLIKELY. PAS serves as a reasonably efficient equation solver.

The algebraic solutions could turn out to be <0 or >1 either because of sampling variation or because the three-probability model is truly nonsensical. Were that the case, as long as the parameters are constrained to the default 0–1 range, MLIKELY.PAS would yield parameter estimates that do not match the calculated values and we would get a positive χ^2 value. Either discrepancy, algebraic solutions that are <0 or >1, or a mismatch between the algebraic and numerical solutions, should certainly clue the investigator to question the adequacy



LIKELIHOOD ANALYSIS:	REGRESSION ANALYSIS:
VARIATION	
G=32.2, 9 d.f., P=0.00018	
CORRELATION	CORRELATION
Signficance:	Signficance:
G=20.0, 1 d.f., P=8×10^{-6}	t=2.44, 8 d.f., 0.025<P<0.01
Sufficiency:	Sufficiency:
G=12.2, 8 d.f., P=0.14	R^{2}=0.426

Figure 5.—Results of maximum-likelihood and regression analyses of the correlation of nondisjunction and rDNA exchange. *G*-test comparisons of the results of the MLIKELY.PAS runs described in Figure 4 indicate that there is highly significant variation in nondisjunction among the crosses (H2 *vs.* H1), provide a single, highly significant estimate of the correlation of the two phenotypes (H2 *vs.* H3), and indicate that the correlation accounts for all but a nonsignificant fraction of the variation in nondisjunction (H3 *vs.* H1). Regression analysis, arbitrarily treating either phenotype as the independent variable, provides two different estimates of the correlation, either of which is significant but not highly so, and leaves a substantial fraction of the variation of nondisjunction unexplained.

of the model. For the data in Table 1, the algebraic solutions for *XY*, *r*, and *n* are all in the 0–1 range, running MLIKELY.PAS for H1 yields the same values, and the χ^2 for H1 is 0. Note, however, that the proportion of *XY* bearing sperm is not actually involved in the hypotheses to be compared, and it would have been legitimate to assume that the value of the parameter *XY* was the same for all 10 crosses instead of separately evaluating it for each cross. An appendix that considers the pros and cons of different ways of formulating H1 is included at the web site.

Under both H2 and H3, there are fewer parameters than independent observations. Thus, there is more than one set of possible solutions, and the maximum-likelihood estimates are the minimum-variance, unbiased set. The estimates under the three hypotheses, and the *G*-test comparisons, are shown graphically in Figure 5. Under H1, we estimate the nondisjunction rates for each cross separately. Under H2, we obtain the maximum-likelihood estimate of a single nondisjunction rate for all of the crosses. Under H3, we obtain the maximum-likelihood estimates of the slope and intercept for correlated behavior of nondisjunction rate and rDNA exchange rate. In addition, Figure 5 shows the results that are obtained from conventional regression analysis that uses the frequencies of rDNA crossovers and nondisjunctional offspring rather than the actual progeny counts.

Likelihood and regression analyses give slightly different estimates of the slope and intercept. Indeed, with regression analysis there are two equally sensible lines of least-squares fit, with the best estimate of the underlying parameters somewhere in between. Regression analysis assumes that the values of one variable, the independent variable, are chosen by the experimenter and are not subject to sampling error. That is not in fact true in this kind of experiment, where both variables are actually determined by the data. Unless we have reason to believe that one parameter is known with greater precision than the other, either can be used as the independent variable. Maximum likelihood, in contrast, gives a single solution that takes account of the effects of sampling variation on both variables.

In neither analysis was the intercept constrained to pass through the origin, but the maximum-likelihood estimate of the intercept is 0 and the intercepts of both regression lines are not significantly different from 0. The statistics, however, are quite different. First, the maximum-likelihood method allows us to isolate a single variable and test whether it shows significant experimental variation in the first place; this cannot be parsed out with regression analysis. Second, the method of maximum likelihood provides a far more powerful test of whether the correlation is significant. In this instance, the regression analysis points to a significant correlation, but only at the 0.025 level; the *G* test indicates that it is actually very highly significant indeed. In other words, regression analysis, by using frequencies rather than the observed numbers, has thrown away a lot of information. Third, the likelihood analysis provides a direct test of whether the correlation adequately explains the experimental variation. Here, the unexplained variation is not only small, it is statistically insignificant. Regression analysis also provides a measure, if not a direct test, of sufficiency. As long as the intercept is calculated rather than forced through the origin, R^2 is the fraction of the variance that is explained by the correlation. Here, with less than half of the variance explained by the (albeit significant) correlation, regression analysis suggests that a substantial fraction of the experimental variation has not been accounted for, while the likelihood analysis tells us that only an insignificant fraction of the variation remains unexplained. In large measure this vagueness indicated by the regression analysis results from the lack of fit between the experimental design that has two variables subject to sampling errors and the assumption of regression analysis that one variable is error free.

A note is in order at this point about the need for care in defining the space within which MLIKELY.PAS searches for the maximum-likelihood solutions. In general, a slope and intercept can take on any positive or negative values, but allowing unconstrained iteration of the slope and intercept can lead to finding local and/or nonsensical bumps in the likelihood function. It is important to provide hints to the program in the form of constraints on the parameter ranges. Inspection of the data before running the program will generally suffice, and even if one fails to do that in advance, the absurdity of the result at a false maximum is quite evident. For these data, it is clear from inspection that the slope of the correlation must be positive. Given that, it is also evident that the intercept must be less than the maximum-likelihood estimate of the average rate of nondisjunction (H2). As long as either of these hints is provided to the program by setting the lower bound of the slope to zero or the upper bound of the intercept to the value previously found for the average, iteration proceeds quickly to the true maximum. If, however, a negative slope is allowed *and* an intercept greater than the average is allowed *and* the initial guess of the intercept is greater than that average, iteration to a local maximum is possible. The conjunction of these errors will be obvious, however. If a negative slope is allowed and the initial guess of the intercept is set greater than the average nondisjunction rate but less than the highest observed rate, the false solution under H3 (correlation) will be identical to the solution under H2 (invariant nondisjunction rate). If a negative slope is allowed and the initial guess of the intercept is set greater than the highest observed nondisjunction rate, the false solution for H3 will be even worse—if plotted, the line will not even remotely approach the data points. Even if inappropriate bounds are set, however, MLIKELY.PAS finds the correct solution as long as the initial guess of the intercept is less than the average nondisjunction rate.

Further examples of the range of problems amenable to this approach: The foregoing examples, estimating a parameter in the presence of nuisance variables and analysis of correlation, illustrate just two of the many problems in formal genetics that can be tackled using this approach. The web site, in addition to simpler, introductory examples, contains additional real-world examples that illustrate two hypothesis-testing problems that arise with regularity: (1) testing whether only a subset of parameters differ between a control and an experiment; and (2) taking account of sampling variation in control crosses done to evaluate confounding variables. In outline, those examples are as follows:

1. Sandler *et al.* (1968) suggested that a useful classification of recombination-defective meiotic mutants could be based on whether a mutant reduces map distances without affecting the coefficient of coincidence, or whether it affects both recombination and interference. In this example (abbreviated from Robbins 1977), mutant and control recombination in four marked regions are compared. A simple contingency test shows that the mutant suppresses recombination, but parsing crossover frequencies and coefficients of coincidence using maximum-likelihood methods is necessary to test whether the mutant affects interference *per se*.
2. In the first example described in this article, it was possible to eliminate the effects of a nuisance variable (penetrance) using a single set of data. Frequently, however, the effect of a confounding variable has to be evaluated in a separate cross and, when an effect is found in the control, it must be taken into account in assessing the experiment. Hearn *et al.* (1991) wanted to determine whether chromosomal rearrangements that variegate for the heterochromatic visible *lt* also variegate for nearby lethals by testing whether viability of the rearrangement is sensitive to modifiers of variegation. A simple contingency test would have sufficed were it not for the possibility that the modifier might have an effect on viability separate from its effect on variegation of the lethal locus. Recognizing this, they did control crosses that lacked the variegating rearrangement to expose the effects of the modifier alone.

Differences in the control crosses must be removed before deciding whether there is an effect in the experimental crosses. A simple, but flawed, approach would be adjusting the numbers in the experiment based on the ratios observed in the control. However, sampling errors are inherent in the control as well as the experiment, but "adjusting" the experimental data based on the controls assumes that the controls are error-free. The preferable approach, used by Hearn *et al.* and detailed in the example, is to construct a model for these viabilities and interactions and apply it simultaneously to all of the data.

DISCUSSION

Maximum-likelihood analysis of data from crosses: There are two themes running through the examples used to illustrate this approach. The first is the wide applicability of a simple numerical approximation approach to finding maximum-likelihood solutions. The second is the insight to be gained from partitioning of variation by even a primitive application of discrete multivariate analysis. In the teaching context, the first allows students to focus on the ideas without getting terribly involved in the mechanics, and the second forces a clear definition of the experimental design and the questions to be asked.

In many instances these ideas parallel each other, but that is certainly not always the case. For some problems, such as in the example of testing whether a meiotic mutant affects interference, only a test of a single hypothesis is needed, but finding the maximum-likelihood estimates of the multiple exchange and interference parameters is made easier by use of the computer. There are surely few geneticists who would be comfortable trying to solve a set of 14 simultaneous equations for the partial derivatives of L with respect to eight map distances and six coefficients of coincidence. Numerical analysis makes this kind of problem tractable.

There are also problems for which multiple hypotheses must be compared, but for which the maximum likelihood is readily found. For example, R. Morell recently posed the following. He was studying a human dominant of reduced penetrance for which genotypes could nevertheless be determined unambiguously by molecular means, even in many instances to the point of knowing whether the particular allele segregating was, for example, a frame-shift or a base substitution. Eyeball perusal of several pedigrees suggested that penetrance was not constant. There are several things worth examining in this situation. First, of course, is the question of whether these are significant differences in penetrance or merely stochastic variation. If there are significant differences, one might want to know, for example, whether penetrance is higher for clearly null alleles than for missense alleles or whether other loci affect expression of this trait. In other words, we need to ask, as we would in an analysis of variance were we following a measured variable rather than numbers of affected and unaffected individuals, whether there are significant differences in variation between and within groups. Testing a series of hypotheses was needed here, but, at the same time, there was no need to turn to numerical approximation to find the several $\ln \hat{L}$ values. Only one variable was involved, penetrance, and the analytic solutions were easily found (Morell *et al.* 1997).

There are also situations outside of formal genetics in which this approach may be of value. For example, we have recently used this kind of analysis in measurement of ribosomal RNA gene copy number (P. Crawley, unpublished results). Because copy numbers of a large number of genotypes were needed, dot-blot hybridizations, with a single-copy reprobe used to control for loading, were counted using a storage-phosphor screen device. The data were therefore in the form of discrete numbers (photons detected in each of many dots) and maximum-likelihood methods are appropriate for testing for differences among the genotypes. MLIKELY.PAS is not designed for the rather awkward bookkeeping involved in the complex data structure of multiple dots of multiple genotypes on multiple blots with probes that may differ in concentration and specific activity from run to run. Nevertheless, we used it to test the utility of this approach. It does work, and it certainly gives cleaner yes/no judgments of significance than does a chart of error bars.

Regardless of the particular questions being investigated, there are several reasons why this approach to teaching statistics is attractive, at least when a program like MLIKELY.PAS can be used to preempt the need for great mathematical competence:

1. There is no need to adapt methods designed for other purposes or for continuous data. There are always assumptions in doing that, which may not be obvious to the casual user of a statistics cookbook and may not hold. In writing the probabilities of each observed class, any assumptions are at least made evident. It forces us to understand our own experiment, and when an assumption is faulty it often becomes glaringly obvious by outcomes such as a value of 1 for the best estimate of a parameter that is a probability.
2. In this approach, there is no need to learn a myriad of different procedures nor to understand the fine points of when they should or should not be used. Here, laziness, rather than necessity, is the mother of invention.
3. If maximum-likelihood solutions exist, the parameter estimates are the minimum-variance unbiased estimates. There are some circumstances for which biased estimators exist that are nevertheless always closer to the population parameter (*e.g.*, pseudo-Bayesian estimators). Except for those cases, however, the maximum-likelihood estimates will provide the most powerful tests of significance possible. The maximum-likelihood approach may reveal significant differences in a given-sized sample when a method transplanted from continuous-variable statistics would not.
4. This approach gives a more comprehensive picture of what is going on than any single test of significance. Much as in conventional analysis of variance with continuous data, we can assess not only the significance of a suspected agent, but the strength of its effect and its sufficiency as an explanation of the observations. A correlation, for example, can be both statistically significant and, at the same time, unimportant. Is the phenomenon real? Is it strong enough that a biologist should care about it at all? Is it of primary or secondary importance? A correlation that explains only 1% of the experimental variation is probably not of much biological importance even if it is significant at a 0.0001 level.

Of course, this approach, particularly the use of numerical methods for solving a nearly unrestricted optimization problem, has its limitations as well. The MLIKELY.PAS algorithm in which parameters are varied in a fixed order sometimes requires that the user have an idea of reasonable guesses to enter as starting points; it may not recover from entirely unreasonable ones. The possibility of finding local maxima, and of missing the true maximum likelihood by an amount that would affect one's inferences, cannot be ruled out, even though it has been of little practical import in a fairly wide variety of applications. For some hypotheses, the interactions of the parameters can be so complex that iteration to a solution takes longer than is reasonable, even though no more than a few minutes are needed for each of the examples discussed here. Finally, likelihood methods are applicable in many situations other than formal genetics, but MLIKELY.PAS was written specifically with crosses in mind. MLIKELY only works for situations where the observations come in the form of a multinomial sample whose probabilities can all be explicitly stated in terms of the parameters to be estimated.

Some improvements to MLIKELY.PAS can also be envisioned, both for teaching purposes and for research uses. The interface might be improved to allow the user to judge when the precision reached is close enough to cease iteration, but a faster, current-generation computer makes the speed gain entirely trivial. The program could allow an option of varying the parameters in random sequence at each iterative step. At the cost of increased computation time, it would be more likely to recover from absurd initial guesses and less likely to halt at a local peak. A pseudo-Bayesian approach could be implemented for data sets that contain small numbers and many cells where the observed number is zero, by running two successive iterations, the first using the actual data, and the second using a set of numbers biased toward the initial maximum-likelihood expectations. Finally, MLIKELY.PAS' data-handling structures are not well suited to every application for which the maximum-likelihood approach would be an improvement over what one sees in the biological literature. Hopefully, colleagues working in other areas of genetics and molecular biology, or a programmer or two, will be intrigued enough by the power of these methods to adapt them to those situations as well.

I am grateful to Joe Felsenstein and Richard Lenski for their critical, helpful, and encouraging comments and for their stamina in wading through a manuscript that originally combined the contents of both this report and a paper on meiotic drive. I am especially grateful to Joe Felsenstein for the encouragement and time he gave a then-young graduate student more than 25 years ago when I wrote the first version of the MLIKELY program. I am also grateful to Rob Morell and Ellen Swanson for their criticisms and for the patience they have shown during the long gestation of this article. Finally, I thank the two anonymous reviewers and the corresponding editor who correctly suggested that the article would be more readable with some of the examples and the appendix removed to the website. Research in my laboratory has been supported by National Science Foundation grant MCB-9305846 and by start up funds from the Università di Siena.

LITERATURE CITED

Asher, J. H. Jr., R. W. Harrison, R. Morell, M. L. Carey and T. B. Friedman, 1996 Effects of *Pax3* modifier genes on craniofacial morphology, pigmentation, and viability: a murine model of Waardenburg syndrome variation. Genomics **34:** 285–298.

Bishop, Y. M. M., S. E. Fienberg and P. W. Holland, 1975a *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Bishop, Y. M. M., S. E. Fienberg and P. W. Holland, 1975b Formal goodness of fit: summary statistics and model selection, pp. 123–175 in *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Crow, J. F., 1993 Felix Bernstein and the first human marker locus. Genetics **133:** 4–7.

Edwards, A. W. F., 1992 *Likelihood* (expanded edition). Johns Hopkins University Press, Baltimore.

Fisher, R. A., 1922 On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. Lond. A **222:** 309–368.

Griffiths, A. J. F., J. H. Miller, D. T. Suzuki, R. C. Lewontin and W. M. Gelbart, 1993 Linkage I: basic eukaryotic chromosome mapping, pp. 132–135 in *An Introduction to Genetic Analysis*. W. H. Freeman, New York.

Griffiths, A. J. F., W. M. Gelbart, J. H. Miller and R. C. Lewontin, 1999 Recombination of genes, pp. 147–151 in *Modern Genetic Analysis*. W. H. Freeman, New York.

Hearn, M. G., A. Hedrick, T. A. Grigliatti and B. T. Wakimoto, 1991 The effect of modifiers of position-effect variegation on the variegation of heterochromatic genes of *Drosophila melanogaster*. Genetics **128:** 785–797.

Hedges, S. B., S. Kumar, K. Tamura and M. Stoneking, 1992 Technical comment on human origins and analysis of mitochondrial DNA sequences. Science **255:** 737–739.

Hilliker, A. J., G. Harauz, A. G. Reaume, M. Clark and A. Chovnick, 1994 Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. Genetics **137:** 1019–1026.

Kastenbaum, M. A., 1958 Estimation of relative frequencies of four sperm types in *Drosophila melanogaster*. Biometrics **14:** 223–228.

King, J. S., and R. K. Mortimer, 1991 A mathematical model of interference for use in constructing linkage maps from tetrad data. Genetics **129:** 597–601.

Lyckegaard, E. M., and A. G. Clark, 1991 Evolution of ribosomal RNA gene copy number on the sex chromosomes of *Drosophila melanogaster*. Mol. Biol. Evol. **8:** 458–474.

McPeeck, M. S., and T. P. Speed, 1995 Modelling interference in genetic recombination. Genetics **139:** 1031–1044.

Morell, R., T. B. Friedman, J. H. Asher and L. G. Robbins, 1997 The incidence of deafness is non-randomly distributed among families segregating for Waardenburg syndrome Type 1 (WS1). J. Med. Genet. **34:** 447–452.

Morton, N. E., 1955 Sequential tests for the detection of linkage. Am. J. Hum. Genet. **7:** 277–318.

Morton, N. E., 1995 LODs past and present. Genetics **140:** 7–12.

Palumbo, G., S. Bonaccorsi, L. G. Robbins and S. Pimpinelli, 1994 Genetic analysis of *Stellate* elements of *Drosophila melanogaster*. Genetics **138**: 1181–1197.

Rasooly, R. S., and L. G. Robbins, 1991 *Rex* and a suppressor of *Rex* are repeated neomorphic loci in the *Drosophila melanogaster* ribosomal DNA. Genetics **129:** 119–132.

Robbins, L. G., 1971 Nonexchange alignment: a meiotic process revealed by a synthetic meiotic mutant of *Drosophila melanogaster*. Mol. Gen. Genet. **110:** 144–165.

Robbins, L. G., 1977 The meiotic effect of a deficiency in *Drosophila melanogaster* with a model for the effects of enzyme deficiency on recombination. Genetics **87:** 655–684.

Robbins, L. G., 1981 Genetically induced mitotic exchange in the heterochromatin of *Drosophila melanogaster*. Genetics **99:** 443–459.

Robbins, L. G., 1999 Are unpaired chromosomes spermicidal? A maximum likelihood analysis of segregation and meiotic drive in ribosomal-DNA deficient *Drosophila melanogaster* males. Genetics **151:** 251–262.

Sandler, L., and M. A. Kastenbaum, 1958 A note on the frequency distribution of tetrads by rank in *Drosophila melanogaster*. Genetics **43:** 215–222.

Sandler, L., D. Lindsley, B. Nicoletti and G. Trippa, 1968 Mutants affecting meiosis in natural populations of *Drosophila melanogaster*. Genetics **60:** 525–558.

Snow, R., 1979 Maximum likelihood estimation of linkage and interference from tetrad data. Genetics **92:** 231–245.

Templeton, A. R., 1992 Technical comment on human origins and analysis of mitochondrial DNA sequences. Science **255:** 737.

Terwilliger, J. D., 1994 *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

Weir, B. S., 1994 The effects of inbreeding on forensic calculations. Annu. Rev. Genet. **28:** 597–621.

Weir, B. S., 1995 DNA statistics in the Simpson matter. Nat. Genet. **11:** 365–368.

Zhao, H. Y., M. S. McPeeck and T. P. Speed, 1995a Statistical-analysis of chromatid interference. Genetics **139:** 1057–1065.

Zhao, H. Y., T. P. Speed and M. S. McPeeck, 1995b Statistical-analysis of crossover interference using the chi-square model. Genetics **139:** 1045–1056.