

# YELLOW PEAS: LIKELIHOODS AND LIKELIHOOD RATIOS

ŚAUNAK SEN\*

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

sen@biostat.ucsf.edu

## INTRODUCTION

The solution to many statistical problems can be obtained by using the concept of a *likelihood*. In fact, most commonly-used statistical procedures such as t-tests, linear regression, logistic regression, and Cox regression are (at least approximately) based on likelihoods. The utility of likelihood methods is most apparent in complex problems, where it provides a logical basis for statistical inference. Such problems abound in genetic analysis, but the basic idea is easiest to see a simple example. Let us look at Mendel's original data that is believed to have started the current genetics revolution.

### Mendel's data

Mendel's seminal paper (<http://www.mendelweb.org/Mendel.html>) gives the following data from the seeds generated from the  $F_1$  hybrid between pea plants with yellow and green albumen. Mendel argued that the yellow albumen trait was dominant and was segregating in the ratio of 3:1.

Plants	Color of Albumen	
	Yellow	Green
1	25	11
2	32	7
3	14	5
4	70	27
5	24	13
6	20	6
7	32	13
8	44	9
9	50	14
10	44	18

Let us first look at the data from the first plant which had 25 yellow and 11 green albumen seeds. What was the strength of evidence in favor of a 3:1 segregation ratio? How does that compare with the evidence in favor of a 1:1 segregation ratio? What is our best guess about the segregation ratio? How confident are we about our guess based on the data at hand?

All these questions can be answered using the statistical concept of the likelihood.

## LIKELIHOOD AND LIKELIHOOD RATIO

Simply put, the likelihood of the data is the probability of the data given a specified statistical model for the data. For example, if we assume that the yellow and green color albumen phenotype is segregating in a 3:1 ratio, the likelihood of the data would be

$$\left(\frac{3}{4}\right)^{25} \times \left(\frac{1}{4}\right)^{11}$$

---

\* © 2006 Śaunak Sen; Last updated March 28, 2006.

More generally, if the probability of yellow seed was  $p$ , then the likelihood (function) would be

$$L(p) = p^{25}(1-p)^{11}.$$

Since we do not know what the actual value of  $p$  is, we want to infer it from data. A natural thing to do would be to compare the likelihood of the observed data for different values of  $p$  (Figure 1). We can do this graphically in Stata with the following commands:

```
set obs 101                                /* set number of points */
generate p = (_n-1)/100                    /* generate values 0.00,0.01,0.02,...,1.00 */
generate l = (p^25)*((1-p)^11)           /* calculate likelihood */
twoway line l p, clstyle(foreground)      /* plot it */
```

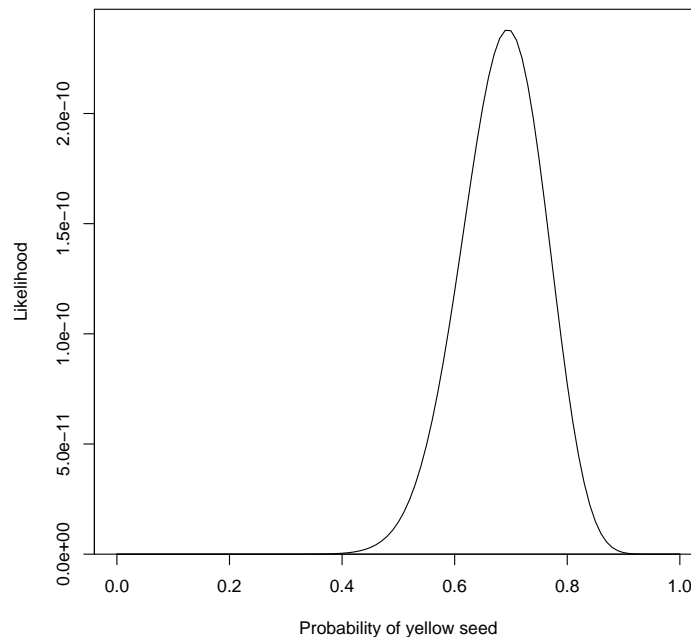


Figure 1: Plot of likelihood of probability of yellow seed ( $p$ ) for a plant with 25 yellow and 11 green seeds (Mendel's data). Using the likelihood plot, our best guess for  $p$  is about 0.7, and values between 0.40 to 0.90 are supported by the data.

Using this graph we can also find which value of  $p$  is best supported by the data. This is called the *maximum likelihood estimate*. For our data it will be  $p = \frac{25}{36} \simeq 0.7$ . You can also use it to get a confidence interval which would be a range of values supported by the graph (approximately 0.40 to 0.90). To compare the evidence between two possible values of  $p$ , say  $p=0.5$  and  $p=0.75$ , we use the *likelihood ratio* which is just the ratio of the likelihoods

$$\frac{L(0.75)}{L(0.5)} = \frac{0.75^{25} \times 0.25^{11}}{0.50^{25} \times 0.50^{11}} \simeq 13.6.$$

In other words the evidence is more than 13 times stronger in favor of the probability of yellow seed being 0.75 compared to being 0.50. In genetics, it is usual to plot not the likelihood ratio, but the base 10 logarithms of the likelihood ratio, which is called the *LOD score*. The graph of the LOD score is given in Figure 2.

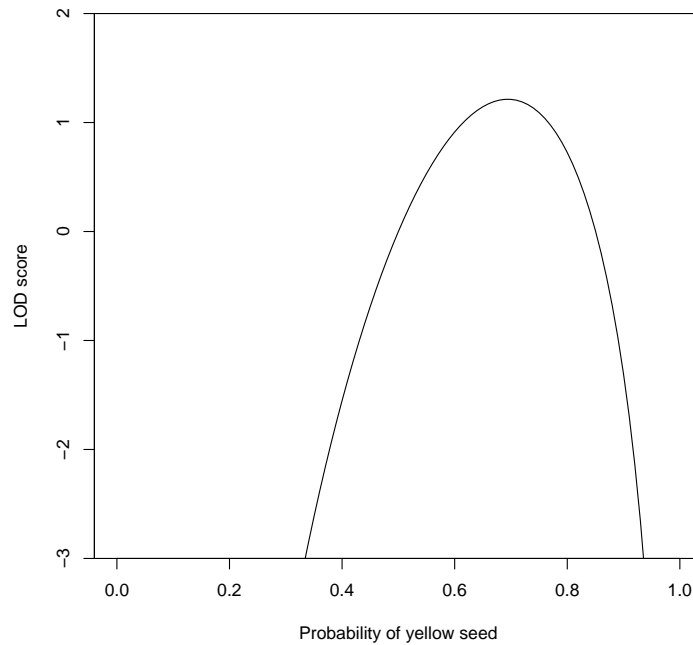


Figure 2: Plot of LOD score for Mendel's data. The LOD score is the base 10 logarithm of the likelihood as a function of  $p$ , compared to the hypothesis of  $p=0.5$ . Notice the approximate quadratic shape of the LOD score.

At this time, it is worth stating what the likelihood ratio does: (1) It compares evidence in data between competing hypotheses (in this case between  $p = 0.5$  and other values). (2) It is calculated on the log10 scale. This means that a LOD score of 1 indicates that the weight of evidence is 10 times as much. A LOD score of 2 indicates that the weight of evidence is 100 times, and a LOD score of -1 indicates that the weight of evidence is 1/10th. (3) We can use the LOD score to find the maximum likelihood estimate which is where the LOD score peaks, and this is the best guess we have for the parameter based on the data. (4) We can also get a range of values of the parameter which are supported by the data (the confidence interval).

There are two more properties that are not obvious, but worth stating. (5) As the sample size gets larger, the maximum likelihood estimate gets closer and closer to the true value of the parameter. For large samples, the maximum likelihood estimate has an approximately normal (Gaussian) distribution with mean equal to the true parameter, and variance inversely proportional to the sample size. The variance is approximately the inverse of the curvature of the (natural) log likelihood. This fact is used to construct approximate confidence intervals for the parameter of interest. (6) For large sample size, twice the natural logarithm of the likelihood ratio has a  $\chi^2$  distribution with degrees of freedom equal to the number of parameters. This fact is used to construct tests of significance.

#### Example: A linkage pedigree

In genetics, the LOD score is usually used for genetic mapping, where the parameter of interest is the putative position of the locus that is linked with our phenotype of interest (in our present problem, the parameter of interest is  $p$ , the probability of yellow seed). Let us consider this problem again in the context of a family pedigree of a rare allele with a dominant mode of action.

In this pedigree we do not have the genotype of the grandfather, but we know that he was affected. Also,

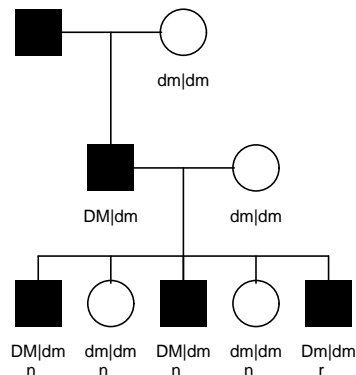


Figure 3: Pedigree for rare disease with dominant mode of action. Males are represented by square, and females by circles. An individual is shaded if they are known to be affected by the disease of interest. The disease allele is denoted by "D," and the healthy allele by "d." Marker alleles are denoted by "m" and "M." We want to see if the marker is linked to the disease. Recombinants are denoted by "r" and non-recombinants by "n".

since we know that the grandmother was homozygous, we know the phase of the father. For that reason, it is easy to figure out who was recombinant ("r") and who was not ("n"). If we denote the recombination fraction between the marker allele and the disease allele by  $\theta$ , then we get following LOD score plot (comparing the likelihoods to a value of  $\theta = 0.5$ ). Paralleling the discussion for Mendel's data, we can say that the LOD score does the following for us: (1) It compares the evidence between the hypothesis that the marker is unlinked to the disease to the possibility that it may be linked to various degrees. (2) Although there is evidence for linkage, it is pretty modest because the peak LOD score is 0.4. (3) Our best estimate for the recombination fraction is 0.2 (the maximum likelihood estimate). Although we can exclude the possibility that the marker is tightly linked (LOD scores less than 0), we are not in a position to completely exclude the possibility that the marker is unlinked ( $\theta=0.5$ ). For that we have to perform a  $\chi^2$  test.

*Exercise: Primula sinensis data*

The following data was one of the earliest examples of linkage observed in the *Primula sinensis* flower collected by de Winton and Bateson.

	Flat leaves	Crimped leaves	Total
Normal eye	328	77	405
Primrose Queen eye	122	33	155
Total	450	110	560

(1) Show that the data are consistent with a hypothesis that the traits are individually segregating in a 3:1 ratio as a dominant Mendelian trait. (2) Show using a  $\chi^2$  test that their joint segregation is not independent.

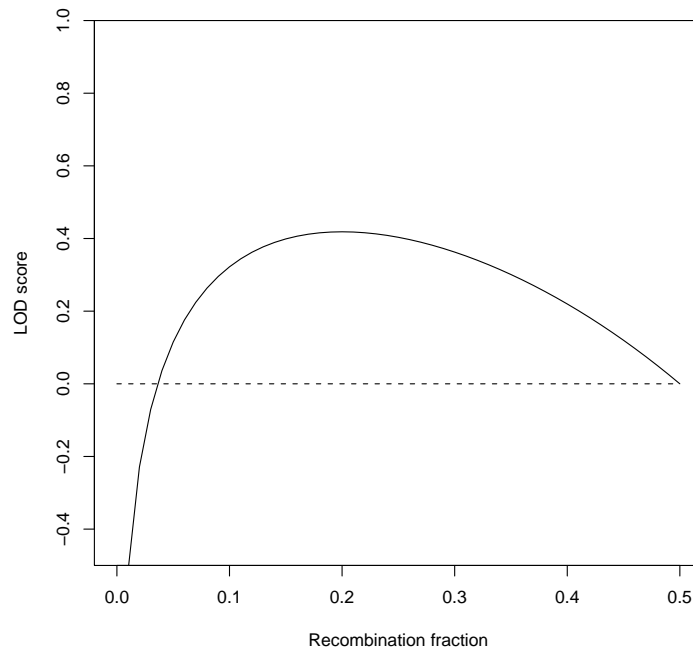


Figure 4: LOD score for pedigree. The LOD score is the base 10 logarithm of the likelihood as a function of the recombination fraction between the marker and the disease. The reference (denominator in the likelihood ratio) is the hypothesis that the marker is unlinked to the disease.

### Likelihoods in practice

*Analysis of variance:* When we want to test the hypothesis that the mean of a normally distributed outcome is the same for different levels of a categorical explanatory variable, we use analysis of variance (ANOVA). The hypothesis is tested using an F test, which can be approximated by a  $\chi^2$  test when the number of observations is large.

*LOD scores* in genetic studies are log likelihood ratios expressed in base 10 logarithms. In human genetics, for parametric linkage analysis, LOD scores compare the hypothesis that the locus being examined is unlinked ( $\theta=0/5$ ) to the hypothesis that it is linked.

*Logistic regression:* If we want to evaluate if a predictor variable is associated with the outcome, then we can use likelihood ratio tests. If the predictor is quantitative or dichotomous, we can just look at the t-value of the regression coefficient (beta). If the predictor is categorical, or if we want to test if a group of variables make a difference, examining the t-values of the regression coefficient is not sufficient. We can use a likelihood ratio test that compares two models – one with the predictor variables, to that without. In Stata, this can be done with the `lrtest` command.

*Complex problems:* The utility of likelihood methods is more apparent in complex settings where it is not easy to write down a test statistic, or to come up with an estimator of a parameter of interest. A good example is genetic mapping in large pedigrees. Another example is population-based association mapping when haplotypes are unknown.