ESTIMATION OF LINKAGE WITH CENSORED DATA

N. A. RAHMAN
Department of Mathematics, University of Leicester

Received 12.ii.64

1. In the usual examples of the measurement of linkage from two factor segregation, it occasionally happens that the observed frequency in one of the four distinct classes is very large as compared with the other class frequencies. For example, Imai (1931) gave the joint segregation of two factors A, a and B, b from F_2 families in *Pharbitis*, Morning Glory, as

Class	AB	Ab	aB	ab
Observed frequency	187	35	37	31
Expected proportion	$\frac{1}{4}(2+\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$

where $\theta \equiv (1-p_1)(1-p_2)$, and p_1 and p_2 are the true recombination fractions in male and female gametogenesis respectively. Thus approximately 64·5 per cent. of the sample observations were found in the AB class. Indeed this is not an exceptionally extreme case, for in the segregation of the two factors G, g and L, l in an F_2 of *Primula Sinensis* given by De Winton and Haldane (1935), out of a sample of 1372 observations, the frequency in the GL class was observed to be 977 *i.e.*, 71·2 per cent. of the total sample.

In such situations the standard maximum likelihood approach for estimating linkage is well-known, but it is possible to adapt the method for the estimation of linkage by completely ignoring the preponderant class in sampling. Mathematically this gives rise to a truncated multinomial distribution, and it may be shown that estimation in this case can lead to fully efficient results provided the sample size is appropriately chosen. If the labour involved in sample enumeration is at least approximately the same in the four classes then, with full efficiency in the estimation of θ, censored sampling in the *Pharbitis* example would lead to a minimum 48 per cent. saving in sampling cost.

This principle of truncation in a multinomial distribution is not new as, indeed, it was first applied by Daniels (1941) in an industrial example; but its use in the study of linkage does not seem to have been indicated specifically. The object of this note is to illustrate this technique with reference to the *Pharbitis* data cited above. The method, however, is of quite general applicability and can, of course, be adapted to any other linkage situation in which there is a similar preponderance of frequency in one class.

2. If, in general, the observed frequencies in the four classes are x_1, x_2, x_3 and x_4 respectively ($\Sigma x_i \equiv N_1$) then $\hat{\theta}$, the usual maximum likelihood estimate of θ, is obtained as a root of the quadratic equation

$$\frac{x_1}{(2+\hat{\theta})} - \frac{(x_2+x_3)}{(1-\hat{\theta})} + \frac{x_4}{\hat{\theta}} = 0,$$

and the large sample variance of $\hat{\theta}$ is

$$\text{Var}\ (\hat{\theta}) = \frac{2\theta(1-\theta)(2+\theta)}{(1+2\theta)N_1}.$$

With the numerical values of Imai, $\hat{\theta} = 0{\cdot}4835$ with a standard error of 0·04663.

Next, suppose the AB class is completely ignored in sampling, and that of a total of N_2 observations taken from the remaining three classes, the observed distribution is as follows:

Class	Ab	aB	ab
Observed frequency . . . . Expected proportion	y_2 $\left(\frac{1-\theta}{2-\theta}\right)$	y_3 $\left(\frac{1-\theta}{2-\theta}\right)$	y_4, $(\Sigma y_i \equiv N_2)$ $\left(\frac{\theta}{1-\theta}\right)$

It then follows that the equation for θ^*, the maximum likelihood estimate of θ in this case, is

$$-\frac{(y_2+y_3)}{(1-\theta^*)}+\frac{y_3}{\theta^*}+\frac{N_2}{(2-\theta^*)}=0$$

and the large sample variance of θ^* is

$$\text{Var}\ (\theta^*) = \frac{\theta(1-\theta)(2-\theta)^2}{2N_2}$$

Clearly θ^* will be fully efficient as compared with $\hat{\theta}$ if

$$\text{Var}\ (\theta^*) = \text{Var}\ (\hat{\theta}),$$

which gives

$$N_2 = N_1 \, . \, \frac{(1+2\theta)(2-\theta)^2}{4(2+\theta)} \equiv N_1 \, . \, g(\theta), \text{ say.}$$

It is now easily proved that for variation in θ within its permissible range $0<\theta<1$, $g(\theta)$ attains a maximum value 0·5174 for $\theta = 0{\cdot}1472$. Therefore, whatever be the true value of θ, the estimate θ^* must have a variance $\ngtr \text{Var}\ (\hat{\theta})$ if $N_2 = 0{\cdot}5174\ N_1$. This means that with censored sampling we need only 51·74 per cent. of the sample observations to have the assurance that Var $(\theta^*) \ngtr$ Var $(\hat{\theta})$. Alternatively, it may also be stated that each observation of the censored sample is at least equivalent to 1·93 observations of the uncensored sample so far as the estimation of θ is concerned. This information can be used to calculate the relative costs of censored and uncensored samples of different sizes.

To go back to our numerical illustration, suppose that θ is, in fact, equal to 0·4835. Then $g(\theta) = 0{\cdot}4554$, and with $N_1 = 290$, the upper limit for N_2 is 133. In fact, the total number of observations in the Ab, aB and ab classes was 103, and so with censored sampling the cost of enumerating 187 observations in the AB class is to be set against that of obtaining 30 additional observations in the Ab, aB and ab classes as a whole. On the other hand, without any specific information about θ apart from its permissible range of variation, the maximum value of N_2 is 151 compared with $N_1 = 290$ so that we would need an additional 48 observations in the Ab, aB and ab classes jointly for ignoring the 187 observations in the preponderant AB class. In either case the reduction in sample size is substantial.

3. From a general practical point of view, there seems little interest in truncation if the estimation of linkage is based on relatively small samples or when the cost of censored sampling is high. However, when the samples are really large or when estimates are required repeatedly, then the truncation of the distribution seems a useful possibility. Admittedly, the case of completely ignoring the preponderant class in sampling is rather extreme, but for large samples the principle of truncation can be combined with a double sampling procedure to attain any specified efficiency of estimation.

To illustrate this it is convenient to continue with the *Pharbitis* example. Suppose, then, that the estimation of θ is to be based on the enumeration of extensive data. Consider first an uncensored sample of ν_1 observations, and suppose the observed class frequencies are u_1, u_2, u_3 and u_4 respectively ($\Sigma u_i \equiv \nu_1$). We now assume that the preponderance of the AB class ensures that u_1 is sufficiently large but that the entire data are not fully recorded. At this stage, we introduce censored sampling, and suppose an independent sample of ν_2 observations is taken with observed frequencies v_2, v_3 and v_4 in the Ab, aB and ab classes respectively ($\Sigma v_i \equiv \nu_2$). We shall not specify the size ν_2 at the moment as it will be determined later to ensure any desired efficiency for the estimation of θ.

For any ν_2, the equation for obtaining θ^{**}, the maximum likelihood estimate θ based on the joint likelihood of $(\nu_1+\nu_2)$ observations, is

$$\frac{u_1}{(2+\theta^{**})}-\frac{(u_2+u_3+v_2+v_3)}{(1-\theta^{**})}+\frac{(u_4+v_4)}{\theta^{**}}+\frac{\nu_2}{(2-\theta^{**})}=0.$$

This is naturally a cubic in θ^{**} but its numerical solution is quite straightforward. Furthermore, the variance of θ^{**} is

$$\text{Var}\,(\theta^{**})=\frac{2\theta(1-\theta)(2-\theta)(4-\theta^2)}{\nu_1(1+2\theta)(2-\theta)^2+4\nu_2(2+\theta)}.$$

If we now postulate that the efficiency of θ^{**} to be at least equal to that of an estimate based on an uncensored sample of $k\nu_1$, $(k>1)$ observations, then

$$\nu_2=\frac{(1+2\theta)(2-\theta)^2}{4(2+\theta)}\cdot(k-1)\nu_1=g(\theta)(k-1)\nu_1.$$

Therefore the maximum value of ν_2 is $0{\cdot}5174(k-1)\nu_1$ whatever the true value of θ, and so the censored sample would at most be $(k-1)51{\cdot}74$ per cent. of the first. We could further improve upon this if we are prepared to use the first sample to provide an estimate of θ for determining ν_2 more precisely.

4. Finally, we may consider yet another scheme for censored sampling if with a single sample we are not prepared to sacrifice totally the information contained in the preponderant class for the estimation of a parameter. Again using the *Pharbitis* example, suppose it is decided to record only a fraction p, $0<p<1$, of the AB observations but otherwise the data are recorded completely. In this case the expected proportions in the four classes are proportional to $p(2+\theta)$, $(1-\theta)$, $(1-\theta)$ and θ respectively. Consequently, if in a total sample of N_3 observations with a censoring fraction p

the observed frequencies in the four classes are z_1, z_2, z_3 and z_4 respectively ($\Sigma z_i \equiv N_3$), then the equation for $\tilde{\theta}$, the maximum likelihood estimate of θ, is

$$\frac{z_1}{(2+\tilde{\theta})} - \frac{(z_2+z_3)}{(1-\tilde{\theta})} - \frac{z_4}{\tilde{\theta}} + \frac{N_3(1-p)}{[2(1+p)-(1-p)\tilde{\theta}]} = 0,$$

an equation which is only apparently a cubic since the coefficient of $\tilde{\theta}^3$ is zero. Furthermore, the large sample variance of $\tilde{\theta}$ may be shown to be

$$\text{Var}\,(\tilde{\theta}) = \frac{\theta(1-\theta)(2+\theta)[2(1+p)-(1-p)\theta]^2}{2N_3[2(1+p)+(1+7p)\theta]}.$$

Hence $\tilde{\theta}$ will be as efficient an estimate of θ as $\hat{\theta}$ obtained from a completely enumerated sample of N_1 observations if

$$N_3 = \frac{(1+2\theta)[2(1+p)-(1-p)\theta]^2 N_1}{4[2(1+p)+(1+7p)\theta]} \equiv h(\theta)\,.\,N_1, \text{ say.}$$

In the limiting case when $p = 1$, then evidently $N_3 = N_1$, whereas for $p = 0$ we have our earlier result $N_3 = N_1 g(\theta)$.

If an initial estimate of θ is available, then the appropriate censoring fraction can be determined to ensure any desired degree of precision for the estimate of θ. On the other hand, if no prior information is available about θ, then we can obtain the maximum value of N_3 as a function of $p \times N_1$ for variation over θ. Thus it can be proved that the maximum value of $h(\theta)$ is attained for $\theta \sim \dfrac{2(1+p)}{(13+19p)}$ so that approximately

$$N_3 \leqq \frac{4(1+p)(17+23p)(3+5p)^2}{(7+13p)(13+19p)^2}\,.\,N_1.$$

As a numerical illustration, suppose $p = 0{\cdot}1$. Then $N_3 \leqq 0{\cdot}5645 N_1$. This upper limit for N_3 is obviously greater than that obtained for $p = 0$, *viz.* $0{\cdot}5174 N_1$; and in a practical case the choice between alternative sampling procedures will be determined by whether it is cheaper to do a 10 per cent. count of the preponderant class or to take a somewhat smaller sample but ignoring this class. A decision of this kind would obviously depend upon the relative costs of sampling the preponderant and the other classes. For our present purpose the main point is to indicate that the method of maximum likelihood can be adapted to a flexible approach to sampling and efficient estimation of population parameters. This consideration is relevant since with estimates whose variance is proportional to the reciprocal of the sample size all assessments of efficiency can only be relative.

REFERENCES

DANIELS, H. E. 1941. A method of improving certain routine measurements. *J. Roy. Stat. Soc.*, *7* (Suppl.), 146.

IMAI, Y. 1931. Linkage studies in *Pharbitis*. Nil I. *Genetics*, *16*, 26.

WINTON, D. DE, AND HALDANE, J. B. S. 1935. The genetics of *Primula sinensis*. III. Linkage in the diploid. *J. Genetics*, *31*, 67.