Asymptotic distributions of linear combinations of logs of multinomial parameter estimates

Roger B. Newson

23 July, 2008

1 Formulas

Suppose that Y has a multinomial distribution with parameters $(p_1, \ldots, p_m)$. (That is to say, suppose that $\Pr(Y = i) = p_i$ for an integer i such that $1 \leq i \leq m$.) Given a sample $Y_1, \ldots, Y_n$ of multinomial random variables with vector parameter $(p_1, \ldots, p_m)$, the maximum–likelihood (and method of moments) estimator of p_i is

$$\hat{p}_i = \#\{j : Y_j = i\}/n = n_i/n \tag{1}$$

where $n_i = \#\{j : Y_j = i\}$ is the number of sample indices j such that $Y_j = i$. The dispersion matrix of the $\hat{p}_i$ is defined by

$$\mathrm{Cov}[\hat{p}_i, p_j] \quad = \quad \begin{cases} n^{-1}p_i(1-p_i), & \text{if } i = j, \\ -n^{-1}p_i p_j, & \text{if } i \neq j. \end{cases} \tag{2}$$

We aim to estimate the $\ln p_i$ with a view to estimating linear combinations of these logs. For each i, we have

$$\frac{\partial}{\partial p_i} \ln p_i = \frac{1}{p_i}, \tag{3}$$

implying that the covariance of $\ln \hat{p}_i$ and $\ln \hat{p}_j$ is given by

$$\mathrm{Cov}[\ln \hat{p}_i, \ln \hat{p}_j] \quad = \quad \frac{1}{p_i p_j}\mathrm{Cov}[\hat{p}_i, \hat{p}_j] + o_p(n^{-1}) \quad = \quad \begin{cases} n^{-1}(1-p_i)/p_i + o_p(n^{-1}), & \text{if } i = j, \\ -n^{-1} + o_p(n^{-1}), & \text{if } i \neq j, \end{cases} \tag{4}$$

where $o_p(n^{-1})$ is a term with the feature that $o_p(n^{-1})/n^{-1}$ is consistent for zero. It follows that, if $(\gamma_1, \ldots, \gamma_m)$ is a vector of coefficients, and we estimate the linear combination $\lambda = \sum_{i=1}^m \gamma_i \ln p_j$ using the estimator $\hat{\lambda} = \sum_{i=1}^m \gamma_i \ln \hat{p}_j$, then the variance of $\hat{\lambda}$ is expressed as

$$\begin{aligned} \mathrm{Var}\left[\sum_{i=1}^m \gamma_i \ln \hat{p}_i\right] &= \sum_{i=1}^m \sum_{j=1}^m \gamma_i \gamma_j \mathrm{Cov}[\ln \hat{p}_i, \ln \hat{p}_j] \\ &= n^{-1}\sum_{i=1}^m \gamma_i^2(1-p_i)/p_i \quad - \quad n^{-1}\left(\sum_{i=1}^m \sum_{i \leq j \leq m,\, j \neq i} \gamma_i \gamma_j\right) \quad + \quad o_p(n^{-1}) \\ &= n^{-1}\sum_{i=1}^m \gamma_i^2/p_i \quad - \quad n^{-1}\left(\sum_{i=1}^m \sum_{j=1}^m \gamma_i \gamma_j\right) \quad + \quad o_p(n^{-1}) \\ &= n^{-1}\sum_{i=1}^m \gamma_i^2/p_i \quad - \quad n^{-1}\left(\sum_{i=1}^m \gamma_i\right)^2 \quad + \quad o_p(n^{-1})\,. \end{aligned} \tag{5}$$

In the special case where $\sum_{i=1}^m \gamma_i = 0$, this simplifies to

$$\mathrm{Var}\left[\sum_{i=1}^m \gamma_i \ln \hat{p}_i\right] \quad = \quad n^{-1}\sum_{i=1}^m \gamma_i^2/p_i \quad + \quad o_p(n^{-1})\,. \tag{6}$$

It follows that a consistent standard error formula for $\hat{\lambda}$ is given in the general case by

$$\widehat{\mathrm{SE}}[\hat{\lambda}] \quad = \quad \sqrt{n^{-1}\sum_{i=1}^m \gamma_i^2/\hat{p}_i \quad - \quad n^{-1}\left(\sum_{i=1}^m \gamma_i\right)^2} \quad , \tag{7}$$

and, in the special case where $\sum_{i=1}^{m} \gamma_i = 0$, this simplifies to

$$\widehat{\mathrm{SE}}[\hat{\lambda}] \quad = \quad \sqrt{n^{-1} \sum_{i=1}^{m} \gamma_i^2 / \hat{p}_i} \quad = \quad \sqrt{\sum_{i=1}^{m} \gamma_i^2 / n_i} \quad . \tag{8}$$

Confidence intervals for λ, calculated using these standard errors, are typically exponentiated to derive confidence intervals for $\exp(\lambda)$, which are usually easier for non–mathematicians to understand.

In the case where there are 2 vectors of coefficients $(\alpha_1, \ldots, \alpha_m)$ and $(\beta_1, \ldots, \beta_m)$, and the linear combinations of logs are $\nu = \sum_{i=1}^{m} \alpha_i \ln p_i$ and $\xi = \sum_{i=1}^{m} \beta_i \ln p_i$, then, by an argument similar to (5), their covariance is of the form

$$\mathrm{Cov}[\nu, \xi] \quad = \quad n^{-1} \sum_{i=1}^{m} \alpha_i \beta_i / p_i \quad - \quad n^{-1} \left(\sum_{i=1}^{m} \alpha_i \right) \left(\sum_{i=1}^{m} \beta_i \right) \quad + \quad o_p(n^{-1}) \, , \tag{9}$$

and the second term is zero if *either* $\sum_{i=1}^{m} \alpha_i$ *or* $\sum_{i=1}^{m} \beta_i$ is zero. This covariance can be used to derive standard errors for transformations, or for linear combinations of linear combinations.

2 Examples

2.1 The odds ratio

Suppose that there is 1 random sample of n units, in which 2 binary variables are measured on the jth unit, with possible values 0 and 1 and denoted X_{j1} and X_{j2}, respectively, and common probabilities for all j

$$\begin{aligned} P_{00} &= \Pr\{X_{j1} = 0 \wedge X_{j2} = 0\}, \qquad P_{01} = \Pr\{X_{j1} = 0 \wedge X_{j2} = 1\} \, , \\ P_{10} &= \Pr\{X_{j1} = 1 \wedge X_{j2} = 0\}, \qquad P_{11} = \Pr\{X_{j1} = 1 \wedge X_{j2} = 1\} \, . \end{aligned} \tag{10}$$

If we define $Y_j = 2X_{1j} + X_{2j} + 1$, then the Y_j are multinomial, with possible integer values 1 to 4 and probabilities

$$p_1 = P_{00}, \quad p_2 = P_{01}, \quad p_3 = P_{10}, \quad p_4 = P_{11} \, . \tag{11}$$

If the coefficients are $\gamma_0 = \gamma_4 = 1$ and $\gamma_2 = \gamma_3 = -1$, then the linear combination $\lambda = \sum_{i=1}^{4} \gamma_i \ln p_i$ is the familiar expression for the log of the common odds ratio, measuring the association between X_{j1} and X_{j2} for all j. As the coefficients sum to zero, the variance of the sample log odds ratio $\hat{\lambda} = \sum_{i=1}^{4} \gamma_i \ln \hat{p}_i$ is of the form (6), and is given by

$$\mathrm{Var}[\hat{\lambda}] \quad = \quad n^{-1}(1/P_{00} + 1/P_{01} + 1/P_{10} + 1/P_{11}) \quad + \quad o_p(n^{-1}) \, , \tag{12}$$

and the sample standard error is of the form (8), and is given by the familiar formula

$$\widehat{\mathrm{SE}}[\hat{\lambda}] \quad = \quad \sqrt{1/N_{00} + 1/N_{01} + 1/N_{10} + 1/N_{11}} \, , \tag{13}$$

where $N_{gh} = nP_{gh}$ for g and h in the set $\{0, 1\}$.

2.2 Hardy–Weinberg disequilibrium

In the genetics of diploid organisms such as humans and fruit flies, a typical 2–allele polymorphism has a commoner allele A, a rarer allele a, and possible genotypes AA, Aa and aa, with population prevalences P_{AA}, P_{Aa} and P_{aa}, respectively. If n individuals are sampled randomly from a population and genotyped, then we can estimate sample prevalences $\hat{P}_{AA} = N_{AA}/n$, $\hat{P}_{Aa} = N_{Aa}/n$ and $\hat{P}_{aa} = N_{aa}/n$, where N_{AA}, N_{Aa} and N_{aa} are the respective sample frequencies of the 3 genotypes.

Lindley (1988) proposed a reparameterization of the 3–dimensional vector parameter (P_{AA}, P_{Aa}, P_{aa}) to a 2–dimensional vector parameter (α, β), defined by

$$\begin{aligned} \alpha &= \tfrac{1}{2} \ln \left(\tfrac{4 P_{AA} P_{aa}}{P_{Aa}^2} \right) = \ln 2 + \tfrac{1}{2} \ln \left(\tfrac{P_{AA} P_{aa}}{P_{Aa}^2} \right) \, , \\ \beta &= \tfrac{1}{2} \ln \left(\tfrac{P_{AA}}{P_{aa}} \right) \, . \end{aligned} \tag{14}$$

(This is possible because of the constraint $P_{AA} + P_{Aa} + P_{aa} = 1$.) The parameter α is zero if the paternal and maternal alleles of a randomly–sampled member of the population are statistically independent, as they will be if their mothers and fathers selected each other at random (at least with respect to genotype).

If $\alpha = 0$, then the polymorphism is said to be in Hardy–Weinberg equilibrium. A positive value for α indicates a systematic tendency for the maternal and paternal alleles to be the same ("inbreeding"), whereas a negative value for α indicates a systematic tendency for the maternal and paternal alleles to be different ("outbreeding"). The parameter β is the log of the square root of the ratio between the population prevalences of the two homozygous genotypes AA and aa, and will be zero if the two homozygous genotypes are equally common, and equal to the log of the ratio of the allelic frequencies of A and a, if the population is indeed in Hardy–Weinberg equilibrium.

The parameters $\alpha - \ln 2$ and β are clearly linear combinations of logs of multinomial proportions. If we denote $p_1 = P_{AA}$, $p_2 = P_{Aa}$ and $p_3 = P_{aa}$, then the index of each multinomial proportion will be one greater than the number of copies of the rarer allele. The vector of coefficients for the linear combinations of the $\ln p_i$ will be $(0.5, -1, 0.5)$ in the case of $\alpha - \ln 2$, and $(0.5, 0, -0.5)$ in the case of β. Both of these vectors of coefficients sum to zero. The maximum–likelihood (and method of moments) estimators of the parameters $\alpha - \ln 2$ and β will be the corresponding linear combinations of the $\ln \hat{p}_i$. By the transformation–invariance property of maximum–likelihood estimators and the location–invariance property of moments, the estimate of α is derived by adding $\ln 2$ to the estimate of $\alpha - \ln 2$. Therefore, the estimates of α and β are

$$\begin{aligned} \hat{\alpha} &= \tfrac{1}{2} \ln \hat{P}_{AA} + \tfrac{1}{2} \ln \hat{P}_{aa} - \ln \hat{P}_{Aa} + \ln 2 \,, \\ \hat{\beta} &= \tfrac{1}{2} \ln \hat{P}_{AA} - \tfrac{1}{2} \ln \hat{P}_{aa} \,. \end{aligned} \tag{15}$$

The variances and covariances of these parameters are of the form (6) and (9), respectively, and variances and covariances involving α are the same as the corresponding variances and covariances involving $\alpha - \ln 2$. The variance–covariance matrix of $\hat{\alpha}$ and $\hat{\beta}$ is therefore given by

$$\begin{aligned} \mathrm{Var}[\hat{\alpha}] &= n^{-1} \left(\tfrac{1}{4P_{AA}} + \tfrac{1}{4P_{aa}} + \tfrac{1}{P_{Aa}} \right) + o_p(n^{-1}) \,, \\ \mathrm{Var}[\hat{\beta}] &= n^{-1} \left(\tfrac{1}{4P_{AA}} + \tfrac{1}{4P_{aa}} \right) + o_p(n^{-1}) \,, \\ \mathrm{Cov}[\hat{\alpha}, \hat{\beta}] &= n^{-1} \left(\tfrac{1}{4P_{AA}} - \tfrac{1}{4P_{aa}} \right) + o_p(n^{-1}) \,. \end{aligned} \tag{16}$$

The sample standard errors of $\hat{\alpha}$ and $\hat{\beta}$ are of form (8), as follows:

$$\begin{aligned} \widehat{\mathrm{SE}}[\hat{\alpha}] &= \sqrt{\tfrac{1}{4N_{AA}} + \tfrac{1}{4N_{aa}} + \tfrac{1}{N_{Aa}}} \,, \\ \widehat{\mathrm{SE}}[\hat{\beta}] &= \sqrt{\tfrac{1}{4N_{AA}} + \tfrac{1}{4N_{aa}}} \,. \end{aligned} \tag{17}$$

A variety of end–point transformations may be carried out on the parameters and their confidence intervals, to make the parameters more easy for non–mathematicians to understand. In fact, the parameter α has repeatedly been reinvented, with a variety of transformations. The parameter $\theta = \exp(-\alpha)$ was proposed in Olson (1993) and Olson and Foley (1996) as a measure of Hardy–Weinberg disequilibrium, apparently independently of Lindley (1988). The present author proposed the equivalent parameter $H = \exp(\alpha - \ln 2)$, for the same purpose, at some point in the late 1990s after 1996, without knowledge of any of the aforementioned references, naming it the "geometric mean homozygote–heterozygote ratio". It seemed strange to the present author that most geneticists seemed to be using a chi–squared test for Hardy–Weinberg equilibrium, and thereby discarding information about the direction of the disequilibrium. And it was a surprise to find so few prior references in the literature to this parameter, which can easily be computed, together with its standard error, using 1940s technology. And it was even more of a surprise to find that Lindley (1988) had derived a standard error for $\hat{\alpha}$ by a totally different methodology, based on likelihood functions, which leads to a *much* more complicated expression for the standard error than the one presented here. *However*, Lindley seems to have had an ulterior motive of developing a Bayesian methodology, rather than arriving at a confidence interval formula for frequentists to use.

3 References

Lindley DV. Statistical inference concerning Hardy–Weinberg equilibrium. *Bayesian Statistics* 1988; **3**: 307-326.

Olson JM. Testing the Hardy–Weinberg law across strata. *Annals of Human Genetics* 1993; **57**: 291–295.

Olson JM, Foley M. Testing for homogeneity of Hardy–Weinberg disequilibrium. *Biometrics* 1996; **52**: 971–979.