Document for betabino (betabino.doc)

Ziheng Yang, version 1.1 January 2001
University College London

Getting started

Run the program betabino.exe from a Windows command prompt window (not from Windows explorer). The program will take input from the default data file betabino.dat and generate results in the default result file betabino.out. Compare betabino.out with betabino.rst to make sure that the program works. You may need to run the program a few times (see below). Print out this document, and read it together with the files betabino.dat and betabino.rst.

Model

This program fits the beta-binomial distribution to data of counts. Let the number of "successes" out of n_j independent trials be k_j in family j. Suppose that the probability of success in family j is p_j, so that k_j has a binomial distribution Bino(n_j, p_j); that is, the probability of observing k_j success out of n_j trials is given by

$$\text{Prob}(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad 0 \le k \le n. \tag{1}$$

Now suppose p_j varies among families according to a beta distribution Beta(α, β). This is a continuous distribution in the interval [0, 1], with the density

$$f(p; \alpha, \beta) = p^{\alpha-1}(1 - p)^{\beta-1} / B(\alpha, \beta), \quad 0 \le p \le 1, \tag{2}$$

where $B(\alpha, \beta)$ is the beta function. The counts among families then follow a beta-binomial distribution:

$$\text{Prob}(k; n, \alpha, \beta) = \binom{-\alpha}{k}\binom{-\beta}{n-k} \Big/ \binom{-\alpha-\beta}{n}, \quad 0 \le k \le n. \tag{3}$$

In this program, the mean m and variance v of the beta distribution are used as parameters to facilitate formulation of hypotheses, so the beta may also be specified as Beta(m, v). The program estimates parameters m and v (or α and β) of the beta distribution by maximum likelihood (ML) and calculates their SE's by using the curvature at the ML estimates.

The program has one extra layer of complexity. Suppose there are C classes, indexed by i. In the ith class, J_i families are tested in n_{ij} trials, with k_{ij} successes observed. In each family, the probability of success is given by the binomial probability Bino(n_{ij}, p_{ij}). The probability parameter p_{ij} varies according to a beta distribution Beta(m_i, v_i) using parameters m and v, or Beta(α_i, β_i) using parameters α and β. Different hypotheses can be formulated concerning whether those beta distributions across classes have the same or different means and variances. The problem is similar to a one-way analysis of variance. The following four models are automatically fitted by the program if the number of classes C > 1:

- *Model 0 (mv)* assumes that the classes have the same mean and variance. The model thus involve only two parameters: the mean and variance of the beta distribution.
- *Model 1 (mvvv)* assumes that the classes have the same mean but different variances. There are thus 1 + C parameters: one mean and C variances for the C classes.
- *Model 2 (mmmv)* assumes that the classes have different means and the same variance. There are thus C + 1 parameters: C means and one variance.
- *Model 3 (mmmvvv)* assumes that the classes have different means and variances. There are 2C parameters: C means and C variances.

The beta-binomial distribution is also known as the *negative hypergeometric distribution*, *inverse*

hypergeometric distribution, *hypergeometric waiting-time distribution*, and the *Markov-Polya distribution*. See Johnson, N.L., S. Kotz & A.W. Kemp (1993).

Likelihood ratio tests

You can compare the above models to test hypotheses using the likelihood ratio test. The most interesting two comparisons are perhaps those between models 0 and 2 and between models 1 and 3. Both test the hypothesis that the mean proportion is the same across classes. The first comparison assumes that the variance is the same across classes, while the second comparison assumes that the variances are different among classes. Suppose that the log likelihood under the simpler model (0 or 1 in the two tests, respectively) is l_0, and that under the more-general model (2 or 3 in the two tests, respectively) is l_1. The likelihood theory stipulates that one compares $2\Delta l = 2(l_1 - l_0)$ with a χ^2 distribution with the degree of freedom equal to the difference in the number of parameters in the two compared models ($C - 1$).

Suppose that there are $C = 6$ classes and the log likelihood values under models 0, 1, 2, and 3 are –329.43, –327.57, –310.72, and –305.22, respectively. The first test comparing models 0 against 2 will have the test statistic $2\Delta l = 2 \times ((-310.72) - (-329.43)) = 37.42$. This difference is significant, compared with the χ^2 distribution with d.f. = $C - 1 = 5$. The P value is 4.93×10^{-7}. So the mean proportions are significantly different among classes. The second test comparing models 1 against 3 has the test statistic $2\Delta l = 2 \times ((-305.22) - (-327.57)) = 44.70$. This difference is also significant, with $P = 1.67 \times 10^{-8}$.

Motivation

Suppose that you have several classes (mating types etc.), and you are interested in possible differences among the probabilities of success among the classes. Then you do an experiment using multiple families within each class. The motivation for using the beta-binomial distribution is that the success probabilities vary among families within a class. If there is little variation among the families within each class, you can combine the counts in the families in each class and use the binomial distribution to compare the different classes and see whether the success probabilities are the same in the different classes. However, if there is variation (in the success probability) among families within the class, you might want to account for the variation and instead test whether the mean success probabilities are different among the classes. One way of accounting for the variation among families is to use the beta binomial distribution, derived when you assume that the success probability varies among families (within a class) according to a beta distribution. The success counts from different families within each class have a beta-binomial distribution, which has the mean and variance of the success probability (across the families) as parameters. The likelihood ratio test then asks whether different classes have different mean success probabilities.

Running the program

Open a Windows command prompt box, and cd to the directory and then run the program

```
betabino
```

or

```
betabino <DataFile> <OutputFile>
```

The default data file name is betabino.dat and the default output file name is betabino.out. The output lists the log likelihood values under the four models, and the parameter estimates, and the SE's calculated using the curvature method.

The iteration routine has lower and upper bounds for parameters. The upper bound for the variance parameter in the beta depends on the mean parameter, and so the program may have problems setting the right bound. Run the program at least twice. This way the iteration will start from different places so that you can check the convergence. The initial values and iteration process are recorded in the rubbish file rub. If you do not see error messages and the results seem trustable and the ML estimates are not at the preset boundary, the results are trustable.

Format of data file

The default data file is betabino.dat and has Prof. Jim Mallet's butterfly data. The early part of the file is shown below. There are **6** cross types (classes), for which **14**, **13**, **15**, **13**, **20**, and **11** families, respectively, are examined for the egg hatch rate. For example, type (class) 1 has **14** families. In family 1, k_{11} = **21** out of n_{11} = **21** eggs were hatched. In family 2, k_{12} = **34** out of n_{12} = **66** eggs were hatched, and so on.

Note that numbers in the data file are separated by "white spaces" (spaces, tabs, line returns etc.), and not by commas.

Limitations

The program sets the maximum number of classes at 500, and the maximum number of families in each class at 500. If your data are larger, the values at the beginning of the source file betabino.c should be increased and the program recompiled.

An alternative parametrization

The program also has an option to use α and β as parameters exactly in the same way the mean (m) and variance (v) are used. To use this option, you need to uncomment the following line at the beginning of betabino.c and recompile the program.

```
#define useAlphaBeta
```

The compiled executable is renamed betabino2.
The four models are then M0' (*ab*), M1' (*abbb*), M2' (*aaab*), and M3' (*aaabbb*). M0' and M3' are equivalent to M0 and M3, where M1' and M2' are different from M1 and M2.

The parameters between the two parametrizations are related in the following way. The mean and variance of the Beta(α, β) distribution are

$m = \alpha / (\alpha + \beta)$,

$v = \alpha\beta / [(\alpha + \beta)^2 (\alpha + \beta + 1)]$.

From the mean (m) and variance (v) we can get the original parameters α and β as

$\alpha = m(m - m^2 - v) / v$

$\beta = (1 - m)(m - m^2 - v) / v$

References

1. Johnson, N.L., S. Kotz & A.W. Kemp (1993) *Univariate discrete distribtutions, 2nd edition*. New York: Wiley, Chapter 6 (section 2.2) equation 6.18
2. Professor Jim Mallet's butterfly paper.

Last modified 6 December 2001 by Ziheng Yang