# READER REACTION

# On a Likelihood-Based Goodness-of-Fit Test of the Beta-Binomial Model

**Steven T. Garren,[1] Richard L. Smith,[2] and Walter W. Piegorsch[3]**

[1]Department of Statistics, 104 Halsey Hall, University of Virginia,
Charlottesville, Virginia 22903, U.S.A.
[2]Department of Statistics, University of North Carolina,
Chapel Hill, North Carolina 27599-3260, U.S.A.
[3]Department of Statistics, University of South Carolina,
Columbia, South Carolina 29208, U.S.A.

## Summary

When faced with proportion data that exhibit extra-binomial variation, data analysts often consider the beta-binomial distribution as an alternative model to the more common binomial distribution. A typical example occurs in toxicological experiments with laboratory animals, where binary observations on fetuses within a litter are often correlated with each other. In such instances, it may be of interest to test for the goodness-of-fit of the beta-binomial model; this effort is complicated, however, when there is large variability among the litter sizes. We investigate a recent goodness-of-fit test proposed by Brooks, Morgan, Ridout, and Pack (1997, *Biometrics* **53**, 1097–1115) but find that it lacks the ability to distinguish between the beta-binomial model and some severely non-beta-binomial models. Other tests and models developed in their article are quite useful and interesting, but are not examined herein.

---

*Key words:* Beta-binomial; Goodness-of-fit; Likelihood; Overdispersion; Pearson statistic.

# 1 Introduction

A common form of discrete data in many biological experiments is the proportion $X/n$, where $X$ is the number out of $n$ of subjects responding to some stimulus. In some settings, data in the form of proportions can exhibit variability in excess of that assumed under the simple binomial model. A common probability distribution for describing such overdispersed data is the beta-binomial (Skellam, 1948), with mass function

$$P(X = x \mid n) = \binom{n}{x} \prod_{k=0}^{x-1}(\pi + k\theta) \prod_{k=0}^{n-x-1}(1 - \pi + k\theta) \Big/ \prod_{k=0}^{n-1}(1 + k\theta),$$

for $x = 0, \ldots, n$. Under this model, the expected value of $X$ is $\mathcal{E}(X|n) = n\pi$ and the variance includes a nonnegative dispersion parameter $\theta$, so that $Var(X|n) = n\pi(1-\pi)(1+\theta)^{-1}(1+n\theta)$. A special case of the beta-binomial is the simple binomial, which occurs for $\theta = 0$.

One of the most successful uses of the beta-binomial model is in laboratory studies of developmental toxicity, where $X$ is the number of fetuses within a litter of size $n$ exhibiting toxic effects after their parents' exposure to some chemical agent (Haseman and Piegorsch, 1994). An important, but oft-overlooked issue in such studies is whether the beta-binomial adequately represents the extra-binomial variation; i.e., how well does the model fit the data?

To address this issue, Brooks, Morgan, Ridout, and Pack (1997) – hereafter BMRP – considered a series of models for overdispersed developmental toxicity data that included the beta-binomial, but that also allowed for various finite mixtures of binomials and beta-binomials. These alternative constructions offer new insight for modeling overdispersed proportions. BMRP also proposed a series of useful procedures for testing model fit for this type of data. In this short note, we center attention on one of these tests: a novel omnibus goodness-of-fit test for the beta-binomial model that avoids specifying an alternative model by working with the maximized likelihood itself, rather than with a likelihood ratio.

Specifically, BMRP first maximized the likelihood under a beta-binomial model, and then simulated from the fitted beta-binomial distribution conditional on the observed values of $n$ to determine a null distribution of the maximized likelihood. By implication, they rejected the beta-binomial model when the observed test statistic fell in either tail of the

null distribution. This produced a two-sided test, where a small $p$-value indicated departure from beta-binomial variability in the original data.

BMRP applied their omnibus goodness-of-fit test to a succession of six developmental toxicity data sets whose response proportions were thought to be overdispersed. They found that the beta-binomial model provided a reasonable fit in all cases. They were able to show, however, that for five of these six data sets, some other model – made up of various finite mixtures of binomials and beta-binomials – provided a better fit than the single beta-binomial. This may lead one to ponder whether their omnibus test statistic lacks the ability to detect deviations from the beta-binomial in certain instances.

As relates to goodness-of-fit testing for the beta-binomial, we feel that unless one can identify *a priori* a specific distribution or class of distributions to define the alternative space, the omnibus alternative is quite the favorable choice, since it gives an absolute criterion against which to assess the model fit. Selecting a valid test statistic within this context, however, is not trivial. For instance, likelihood-based tests can become problematic, as we illustrate with a short example in Section 2. We show that the BMRP maximized likelihood statistic can lack the ability to detect departure from the beta-binomial, compared to a Pearson-type chi-squared statistic, even when the data are grossly non-beta-binomial and the sample size is large. Our point is that while likelihood ratio tests are widely accepted for a broad variety of testing situations (including cases of testing model adequacy), they nonetheless require a class of structured or parametric alternatives against which to operate. When no such class seems suitable, one might be tempted to use the maximized likelihood under the null hypothesis as a test statistic in its own right, as BMRP did for the beta-binomial problem. The present discussion is intended to warn against relying on such a procedure.

## 2 Example

We consider the performance of both the BMRP test and an alternative Pearson chi-squared test, on the following artificial data. First, fix the constant $\lambda \in (0, 1)$. The artificial data consist of $J$ litters each of size $n = 3$, where exactly $\lambda J$ litters have the response $X = 0$,

and the remaining $(1-\lambda)J$ litters have $X = 2$. We ignore any slight discrepancy arising from the fact that $\lambda J$ may have to be rounded to the nearest integer, as our calculations are essentially large-sample approximations ($J \to \infty$) and these will not be affected by rounding error.

This example is unlikely to occur in practice but we are using it to illustrate a general point: that even for a data set such as this which is obviously not beta-binomial, the BMRP test may fail to detect that fact (unless $J$ is extremely large). We believe many other examples could have been constructed to illustrate this point.

Under a beta-binomial model with $n = 3$, let $p_x(\pi, \theta)$ denote the probability of response $x$, for $x = 0, 1, 2, 3$, as a function of the parameters $\pi$ and $\theta$. The maximum likelihood estimators (MLEs) in our example are therefore the values of $\pi$ and $\theta$ that maximize $\lambda \log p_0(\pi, \theta) + (1-\lambda) \log p_2(\pi, \theta)$. Since these depend solely on $\lambda$, we write them as $\pi_\lambda$ and $\theta_\lambda$. Let $f_1 = \lambda \log p_0(\pi_\lambda, \theta_\lambda) + (1-\lambda) \log p_2(\pi_\lambda, \theta_\lambda)$ and $f_0 = \sum_x p_x(\pi_\lambda, \theta_\lambda) \log p_x(\pi_\lambda, \theta_\lambda)$. Then $Jf_1$ is (modulo rounding error) the actual value of the maximized log likelihood test statistic for our artificial data configuration, and $Jf_0$ is its expected value under the hypothesis $H_0$ that the beta-binomial model with parameters $(\pi_\lambda, \theta_\lambda)$ is correct. Also, let $\sigma^2 = \sum_x p_x(\pi_\lambda, \theta_\lambda) \log^2 p_x(\pi_\lambda, \theta_\lambda) - f_0^2$; then $J\sigma^2$ is the variance of the maximized log likelihood statistic for a sample of $J$ litters when $H_0$ is correct. For $\lambda \leq 0.25$, it can be shown that $\theta_\lambda = 0$, so the estimated beta-binomial model is in fact a binomial distribution. In that case $f_0$, $f_1$, and $\sigma^2$ are defined the same way, using the binomial distribution.

Figure 1 plots the values of $f_0$ and $f_1$ across the entire range of $\lambda$. It can be seen that, throughout the range $0.25 \leq \lambda < 1$, the two curves are very close to each other, and at two points ($\lambda = 0.25$ and $\lambda \approx 0.53$) they intersect. Throughout this range, therefore, we might expect a test based on the maximized likelihood statistic to have difficulty discriminating our artificial configuration from a beta-binomial distribution.

For large $J$, the two-sided BMRP test will reject $H_0$, at significance level 0.05, if $J(f_1 - f_0)^2/\sigma^2 > (1.96)^2$, approximately. The sample size $J$ required, to demonstrate that our artificial data set is not beta-binomial, is therefore $(1.96)^2\sigma^2/(f_1 - f_0)^2$. Note that to determine a $p$-value, the actual BMRP test compares the maximized likelihood function of the original data to simulated maximized likelihood functions, which are determined by

re-estimating $\pi$ and $\theta$ from data simulated using the original MLE. Our discussion of their approach appeals to large-samples, but for large $J$, the maximized log likelihood statistic under $H_0$ will have an asymptotic normal distribution, justifying our approximation.

Now for the same artificial data, consider an alternative Pearson chi-squared test. For $J$ litters, the observed frequencies of $x = 0, 1, 2, 3$ are $\lambda J,\ 0,\ (1-\lambda)J,\ 0$. Their expected values under $H_0$ are $Jp_x,\ x = 0, 1, 2, 3$. The Pearson test therefore rejects $H_0$ if $Jf_2 > c$, where $f_2 = (\lambda - p_0)^2/p_0 + p_1 + (1 - \lambda - p_2)^2/p_2 + p_3$ and $c$ is the test critical value. In the calculations to follow, we take $c = 7.8147$, the upper 5%-point of $\chi^2_3$. Therefore, the sample size $J$ at significance level 0.05 required to detect that the data set is not beta-binomial is $7.8147/f_2$. Here again, we have ignored any possible adjustment to $c$ that accounts for estimating the unknown beta-binomial parameters. For large $J$, however, the MLEs are consistent under $H_0$ so the $\chi^2_3$ approximation for the distribution of the test statistic is justified.

In Figure 2, we show the critical values of $J$ at significance level 0.05 required to detect that the artificial data set is not beta-binomial, for both tests. For all except very small $\lambda$, the sample size required under the BMRP test is larger than that required under the Pearson test. For most of the range, the difference between the two sample sizes is several orders of magnitude, diverging to infinity near $\lambda = 0.25$ or 0.53.

## 3 Discussion

The short example in Section 2 demonstrates that omnibus likelihood-based tests of fit for the beta-binomial distribution can be problematic, so caution is advised regarding the use of such tests. Alternative models can exist whose expected log likelihood does not differ greatly from that of the null model. This suggests that there may be cases where the type of goodness-of-fit test proposed by BMRP is unable to detect non-beta-binomial data.

As we noted above, although we believe that the maximized-likelihood test proposed by BMRP suffers from some shortcomings, a number of perfectly sound procedures for comparing models are included in their article. Moreover, BMRP suggested some useful alternative mixture models, including mixtures of binomial distributions and mixtures of

binomial and beta-binomial distributions. These alternative models add greatly to the scope of available models for this type of data.

As regards testing fit of the beta-binomial model, the effort to develop a good omnibus test is difficult, especially when large variation exists in $n$. If a well-defined alternative model can be postulated, then a likelihood ratio test is appropriate (as was done, in fact, in the latter portions of the BMRP article), but a likelihood-based approach for omnibus alternatives seems unreliable. We have considered generalizations of the Pearson chi-squared test to handle the omnibus setting, and will present those results elsewhere.

We should note in closing that for readers wishing to study the BMRP data sets in more detail, the tables in BMRP contain some minor typographical errors. In Table 1 the entry at position (8, 14) should be moved to (9, 14). In Table 2 the entry of "1" should appear at position (11, 16). In Table 4 the entry at position (9, 10) should be moved to (10, 10).

## Acknowledgements

We thank Dr. Steve Brooks for clarifying the few typographical errors in the tables of Brooks et al. (1997), and also for other useful comments on our work in this area. We also thank the editor for providing valuable comments for improving the presentation of this article. This research was partially supported by NIMH grant MH53259-01A2 (STG), NSF grants DMS-9205112 and DMS-9705166 (RLS), and NCI grant CA76031 (WWP).

## References

Brooks, S. P., Morgan, B. J. T., Ridout, M. S., and Pack, S. E. (1997), Finite mixture models for proportions, *Biometrics* **53**, 1097–1115.

Haseman, J. K. and Piegorsch, W. W. (1994), "Statistical analysis of developmental toxicity data" in *Developmental Toxicology*, 2nd edition, editors C. Kimmel and J. Buelke-Sam, New York: Raven Press, pp. 349–361.

Skellam, J. G. (1948), A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials, *Journal of the Royal Statistical Society, Series B* **10**, 257–261.

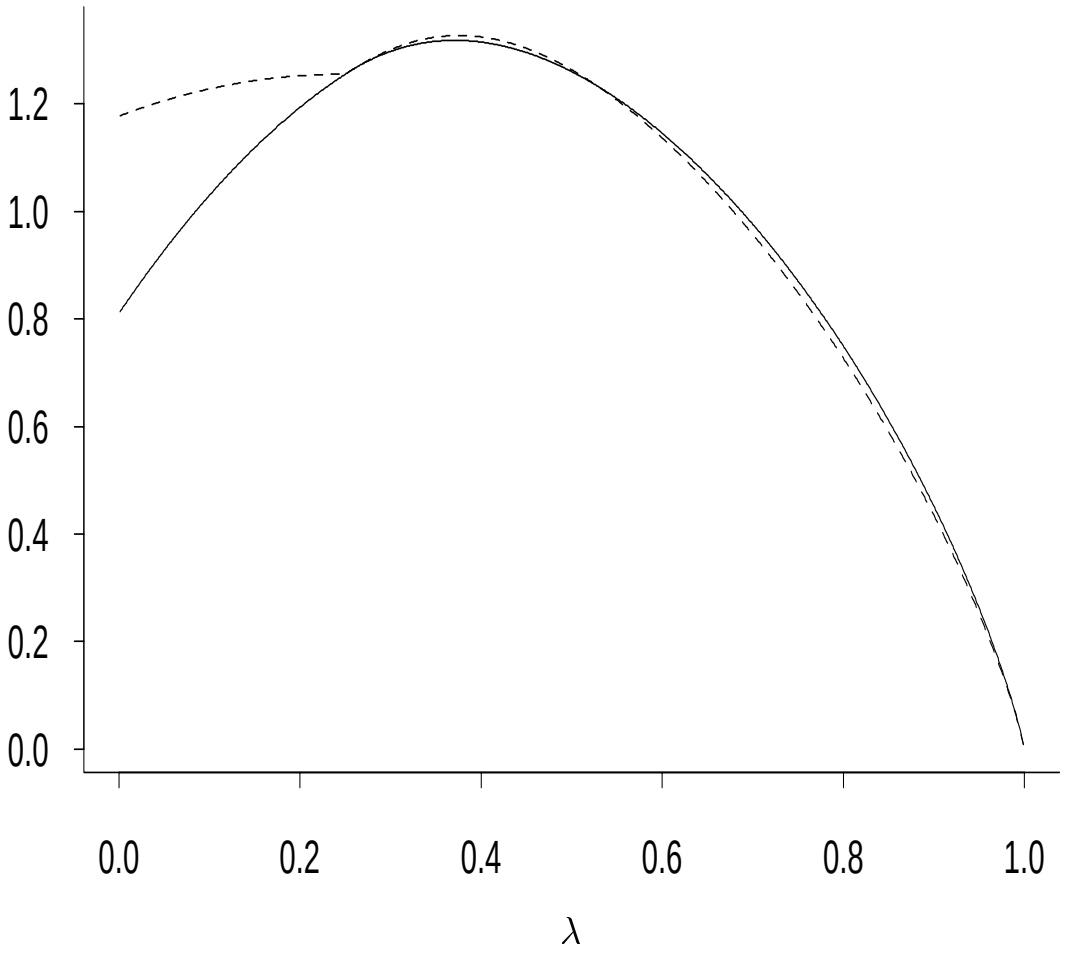


**Figure 1.** The mean values of the negative log likelihood test statistic computed for the beta-binomial model when the beta-binomial model is true ($f_0$, dashed curve) and when the alternative model described in Section 2 is true ($f_1$, solid curve), divided by $J$. The two curves intersect at $\lambda = 0.25$ and $\lambda \approx 0.53$, and are very close throughout the range $0.25 < \lambda < 1$.

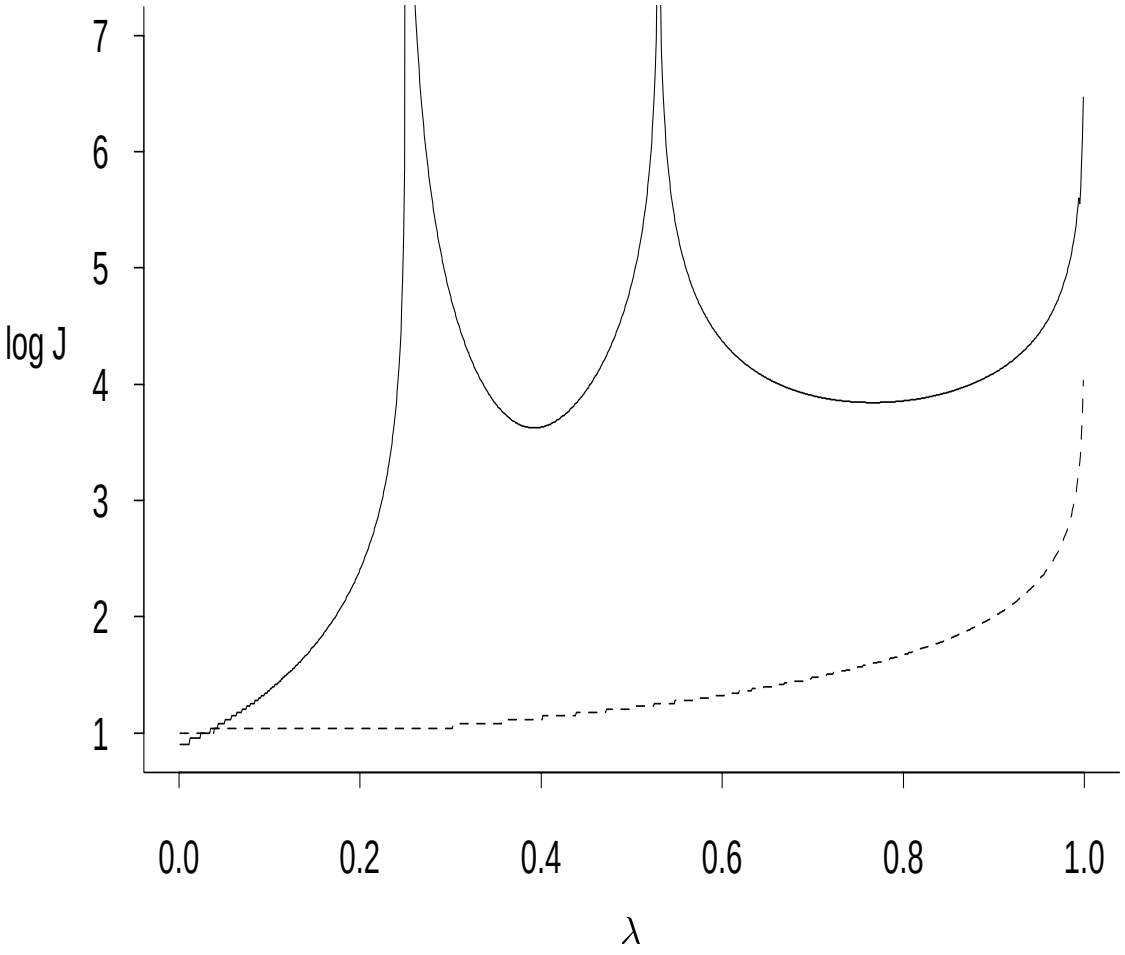


**Figure 2.** Sample size needed to reject the beta-binomial model when the alternative model described in Section 2 is true, assuming a 0.05 significance level. Solid curve: likelihood test proposed by BMRP. Dashed curve: Pearson chi-squared test. The logarithm of the sample size $J$ is in base 10.