Bootstrap Goodness-of-Fit Test for the Beta-Binomial Model

STEVEN T. GARREN[1], **RICHARD L. SMITH**[2] **& WALTER W. PIEGORSCH**[3],

[1] *Department of Mathematics and Statistics, James Madison University, Harrisonburg, Virginia, USA,* [2] *Department of Statistics, University of North Carolina, Chapel Hill, USA and* [3] *Department of Statistics, University of South Carolina, Columbia, USA*

ABSTRACT *A common question in the analysis of binary data is how to deal with overdispersion. One widely advocated sampling distribution for overdispersed binary data is the beta-binomial model. For example, this distribution often is used to model litter effects in toxicological experiments. Testing the null hypothesis of a beta-binomial distribution against all other distributions is difficult, however, when the litter sizes vary greatly. Herein, we propose a test statistic based on combining Pearson statistics from individual litter sizes, and estimate the p-value using bootstrap techniques. A Monte Carlo study confirms the accuracy and power of the test against a beta-binomial distribution contaminated with a few outliers. The method is applied to data from environmental toxicity studies.*

Correspondence: S. T. Garren, Department of Mathematics and Statistics, Burruss Hall, MSC 7803, James Madison University, Harrisonburg, Virginia, 22807, USA. E-mail: garrenst@jmu.edu

1 Introduction: Extra-Binomial Variability

In many experiments encountered in the biological and biomedical sciences, data are generated in the form of proportions, Y/n, where Y is a non-negative count and is bounded above by the positive integer n. When n is assumed fixed and known, Y might be modeled as binomial(n, p); i.e., view Y as the sum of n independent Bernoulli random variables, W_m $(m = 1, \ldots, n)$, with $p = \mathcal{E}W_m$. If some correlation existed among the W_m, then Y would no longer be distributed as binomial. This situation is not uncommon; e.g., in laboratory tests for developmental toxicity the W_m can represent the binary responses of fetuses within a litter of size n from a female rodent exposed to some toxic stimulus (Haseman and Piegorsch, 1994). Since the pregnant rodent is the experimental unit in this situation, the litter-mates represent correlated binary observations, and the sum of those observations may not fit the binomial sampling model. Correlated Bernoulli responses within a litter create what is known as a *litter effect* and often such an effect is modeled hierarchically. The introduction of heterogeneity in p induces correlation between the $\{W_m\}$ and therefore may be used to model litter effects. If p is beta distributed and $(Y|n, p) \sim$ binomial(n, p), then the marginal density of Y given n is beta-binomial (Williams, 1975; Haseman and Kupper, 1979). Additional details of this model are provided in Section 2.

In cases with strong evidence of extra-binomial variability, the beta-binomial model is preferable to the binomial model. Testing for departure from the binomial distribution has been discussed by Cochran (1954) and Tarone (1979) amongst others. Risko and Margolin (1996) provide a recent review and commentary on these methods. In contrast to the binomial distribution, however, formal testing of whether the beta-binomial distribution fits overdispersed proportion data has not been discussed as thoroughly in the literature.

In Section 3 a goodness-of-fit test for the beta-binomial model is constructed by bootstrapping chi-squared tests. Simulation results in Section 4 show that the method provides reasonably accurate estimates of the size of the test, and that the test is powerful against a beta-binomial model contaminated with outliers. Section 5 applies the test to toxicological data sets, and we end in Section 6 with a short discussion.

2 Description of Beta-Binomial Model

One characterization of the beta-binomial model employs the following hierarchy: within the context of a developmental toxicity experiment, assume that a given study consists of J litters of animals, and that n_i is the number of pups in the ith litter, for $i = 1, \ldots, J$. The litter sizes n_i are treated as fixed constants. Let Y_i denote the number of responses in the ith litter. Conditional on p_i, the Y_i are independent binomial random variables

$$(Y_i \mid n_i, p_i) \sim \text{binomial}(n_i, p_i), \quad i = 1, \ldots, J.$$

The random variables $\{p_i, i = 1, \ldots, J\}$ are independent and have the common beta density

$$f(p \mid \alpha, \beta) = [B(\alpha, \beta)]^{-1} p^{\alpha-1}(1-p)^{\beta-1} \quad (0 < p < 1),$$

where α and β are unknown positive constants, and $B(\cdot, \cdot)$ is the beta function. The unconditional distribution of Y_i is expressed by the beta-binomial probability

$$P(Y_i = y \mid n_i, \alpha, \beta) = \binom{n_i}{y} \frac{B(\alpha + y, \beta + n_i - y)}{B(\alpha, \beta)}, \tag{1}$$

for $y = 0, \ldots, n_i$; $i = 1, \ldots, J$. If one defines the strictly positive parameters μ and θ by

$$\mu = (\alpha + \beta)^{-1}\alpha \quad \text{and} \quad \theta = (\alpha + \beta)^{-1} \tag{2}$$

as suggested by Williams (1975), then the mean and variance of Y_i can be expressed by

$$\mathcal{E}(Y_i \mid n_i, \mu, \theta) = n_i\mu \quad \text{and} \quad \text{Var}(Y_i \mid n_i, \mu, \theta) = n_i\mu(1-\mu)(1+\theta)^{-1}(1+n_i\theta)$$

for $i = 1, \ldots, J$. The parameter μ may be referred to as the *mean parameter* of the marginal proportions, where $0 < \mu < 1$. The parameter θ is called the *dispersion parameter*, and if $\theta > 0$, then the data are said to be *overdispersed*. In some settings the variance of Y_i may appear smaller than that for the binomial distribution, suggesting *underdispersion* (Prentice, 1986; Engel and te Brake, 1993), but this is not common in developmental toxicology and hence we will not study it here.

The maximum likelihood estimator (MLE) of (μ, θ) can be shown to be consistent (Lehmann, 1983, pp. 409–413) and is determined numerically. As $\theta \downarrow 0$, the variance of $(Y_i|n_i, \mu, \theta)$ monotonically decreases to $n_i\mu(1-\mu)$, and $(Y_i|n_i, \mu, \theta)$ converges to a binomial random variable.

3 Approaches to Goodness-of-Fit Testing

Suppose the data consist of independent pairs $\{(Y_i, n_i), i = 1, \ldots, J\}$ as described in Section 2, and suppose the goal is to test the null hypothesis that the data follow a beta-binomial distribution (1) against the alternative hypothesis that the distribution is not of this form. Throughout, we will use the (μ, θ) parametrization (2) to represent a specific member of the beta-binomial family.

3.1 *Previous Test Statistics*

For discrete distributions, Pearson's χ^2 statistic is often used for testing goodness-of-fit. The difficulty with this in the case of toxicity experiments is that the data usually represent information from litters of different sizes, and in this case it is not easy to apply the χ^2 test. Mantel and Paul (1987) resolved this problem by assuming that the litter sizes $\{n_i, i = 1, \ldots, J\}$ are themselves random variables from some known distribution, and based the Pearson statistic on the unconditional probability distribution of the Y_i using the MLE of (μ, θ). This approach, however, loses information about the individual litter sizes when determining the observed numbers of the Y_i, and thus could conceal large variations in the proportions of responses among litters.

A different approach is based on likelihood ratio tests. Pack (1986) proposed using the likelihood ratio for testing the specific question of whether two groups of beta-binomial data have the same or two different values of (μ, θ). Lockhart et al. (1992) used the likelihood ratio for testing the beta-binomial model against the alternative that the Y_i have independent binomial distributions with parameters (n_i, p_i). There are some technical difficulties with this latter approach, however; since the number of unknown parameters p_i goes to infinity as $J \to \infty$, the usual asymptotic theory developed for likelihood ratio tests cannot necessarily be applied in this situation.

Liang and McCullagh (1993) proposed a test for determining whether the mean-variance relationship across different litter sizes is consistent with the beta-binomial model. Their method accounted for extra-binomial variability of a form that included the beta-binomial, but employed only a quasi-likelihood fitting algorithm to estimate the model parameters. Thus the formal beta-binomial assumption in (1) was not used. The procedure was presented

more as an approach for comparing different extra-binomial variance structures when the mean-variance relationship is the only distributional property under study, rather than as a formal test for goodness-of-fit to the beta-binomial.

More recently, Brooks et al. (1997) considered overdispersion models for developmental toxicity data that included the beta-binomial, but that also allowed for various finite mixtures of binomials and beta-binomials. In fact, the main emphasis of their paper was to employ finite mixture models to determine which model had the best fit, by examining the maximized likelihood function. For assessing the quality of the beta-binomial assumption, they avoided specifying an alternative model by working with the maximized likelihood itself, rather than with a likelihood ratio. However, Garren et al. (2000) showed that the omnibus goodness-of-fit test proposed by Brooks et al. is not necessarily sensitive to non-beta-binomial data even as the number of litters gets large. In the setting considered by Garren et al. (2000), in which the Brooks test was compared with a Pearson test in a situation with common litter sizes, the Pearson test was overwhelmingly better. The Pearson test is not directly applicable to unequal litter size problems, but the results of Garren et al. (2000) suggest that it would be profitable to look for something which generalizes the Pearson test to this setting.

3.2 *A Bootstrap Test Statistic*

The difficulties of constructing a simple, consistent goodness-of-fit test prompt us to return to the Pearson statistic as in Mantel and Paul (1987), but with some modifications to employ information in the different litter sizes, and to use bootstrapping to determine the null reference distribution. This approach is in line both with the general approach to bootstrapping goodness-of-fit statistics advocated by Romano (1988), and with the concept of an omnibus test advocated by Zhang (1999).

Suppose there are J_1 litters of size n_1, J_2 litters of size n_2, and so on up to J_K litters of size n_K, where $J_k > 0$ $(k = 1, \ldots, K)$ and $\sum_k J_k = J$. Our beta-binomial goodness-of-fit test statistic τ is constructed as follows:

1. Calculate individual Pearson goodness-of-fit test statistics for each litter size. Thus, for each litter size $n = n_k$ $(k = 1, \ldots, K)$, let $O_{y,n}$ denote the observed number of litters of

size n which contain y responses. Similarly, let $E_{y,n}$ denote the expected value under (1) of $O_{y,n}$, assuming J_k litters of this size, where (μ, θ) is estimated by the MLE $(\hat{\mu}, \hat{\theta})$. Define the individual Pearson statistic to be

$$Q_n = \sum_{y=0}^{n} (O_{y,n} - E_{y,n})^2 / E_{y,n}.$$

For a particular realization, let q_n denote the observed value of Q_n.

2. Estimate the distribution function of the Q_n by simulation. Thus, for each litter size $n = n_k$ $(k = 1, \ldots, K)$, generate a large number $J_n M$ of beta-binomial pseudo-random variates with parameters $(\hat{\mu}, \hat{\theta})$, where M does not depend on n. Repeat the calculation in step 1 to generate a parametric bootstrap sample $Q^*_{n,1}, \ldots, Q^*_{n,M}$ such that each $Q^*_{n,m}$ is based on J_n litters. Note that μ and θ are re-estimated for each bootstrap sample. The distribution function of the Q_n is estimated to be

$$\rho_n = M^{-1} \sum_{m=1}^{M} I(Q^*_{n,m} < q_n),$$

where $I(\cdot)$ is the indicator function. The null distribution of ρ_n is approximately uniform $(0, 1)$ though not exactly uniform, even in the limit as $M \to \infty$, since Q_n is discrete. Asymptotic uniformity would be achieved if ρ_n were instead defined as a randomized statistic (c.f., Hogg and Craig, 1995, p. 291), but we chose not to utilize randomized statistics herein.

3. Combine the estimated distribution functions ρ_n by computing the estimated p-value

$$\tau = 1 - \left(\max_{k=1,\ldots,K} \rho_{n_k} \right)^K. \tag{3}$$

Intuitively, the null hypothesis should be rejected if any ρ_n is too large; i.e., if τ is too small. The power transformation in (3) ensures that, if each ρ_n is approximately uniform $(0, 1)$, then τ also is approximately uniform $(0, 1)$, using a standard approach to order statistics (c.f., Hogg and Craig, 1995, pp. 193–200). Hence, the statistic τ estimates the p-value, and an approximate level τ_0 test is obtained by rejecting the null hypothesis whenever $\tau < \tau_0$.

Since the distribution of τ is discrete, our proposed bootstrap method theoretically can be replaced by computation of the exact distribution, but the amount of computing time would be enormous. We prefer, therefore, to use the bootstrap. We explore the operating characteristics of this test in a modest Monte Carlo study in the next section.

4 Monte Carlo Study

We performed a Monte Carlo study to determine how well our test statistic proposed in Section 3 performs with a sample size of $J = 50$. The accuracy of the test's size was examined by simulating a beta-binomial distribution and testing for departure from it, while the power of the test was examined by simulating a beta-binomial distribution mixed with a binomial distribution and again testing for departure from the beta-binomial. Both sets of simulations used litter sizes from a toxicological study considered below, in which the 50 litters ranged in size from 6 to 18 (see Table 1a). To study size and power characteristics, the underlying parameters of the beta-binomial distribution were chosen to vary over all combinations of $\mu \in \{0.05, 0.1, 0.15\}$ and $\theta \in \{0, 0.05, 0.1\}$. In the alternative distribution for the power study the binomial distribution was assigned the response probability $p \in \{0.7, 0.8, 0.9\}$, and the beta-binomial distribution was chosen with mixing probability $\{0.85, 0.9, 0.95\}$. Notice that when the alternative model has a large mixing probability, the model can be viewed as a beta-binomial model contaminated with a few outliers. The number of bootstrap samples was set to $M = 1000$, and the number of independent replications was 2000. The nominal levels, τ_0, were taken as 0.1, 0.05, 0.025, and 0.01. The estimated rejection probabilities (size or power) are the proportions of the 2000 replicates where $\tau < \tau_0$. These appear in Tables 2a (size) and 3a (power). With 2000 replicates, our estimates of the rejection probability, γ, have approximate standard errors of $\sqrt{\gamma(1-\gamma)/2000}$. For example, these estimates of standard error are 0.0067, 0.0049, 0.0035, and 0.0022 when γ has values 0.1, 0.05, 0.025, and 0.01, respectively.

Table 2a illustrates that the estimated size is somewhat close to its nominal level, τ_0, although the tendency is to be slightly above it. Table 3a tends to suggest that large mixing probabilities produce large power when the overall mean, μ, of the beta-binomial model differs greatly from the response probability, p, in the binomial model. Large mixing probabilities (≈ 0.95) frequently produce at least one extreme value of y. This increases the corresponding statistics $O_{y,n}$ and ρ_n and decreases τ, since the $E_{y,n}$ typically are not greatly influenced by a small number of extreme values of y.

To explore further the operating characteristics of this bootstrap approach, we increased the number of litters to $J = 100$ in the Monte Carlo evaluations. Tables 2b and 3b were

produced, respectively, in the same way as Tables 2a and 3a, except Tables 2b and 3b used $J = 100$ litters instead of 50. These additional litters were generated by doubling the 50 litter size frequencies, J_k, in Table 1a. Results from Tables 2a and 2b are similar, indicating that perhaps some of the error when estimating size comes from using a limited number of bootstrap samples and replications, rather than from using only 50 or 100 litters. Table 3b shows greater power than Table 3a for the binomial probability $p = 0.9$. The other values of p did not result in much difference in power between Tables 3a and 3b. The Fortran program which produced Tables 2 and 3 is quite versatile and is available at

http://www.stat.unc.edu/postscript/rs/betabin.

5 Examples

As an illustration we applied our goodness-of-fit test statistic to data from three experiments involving pregnant mice, studied originally by Lockhart et al. (1992). Those experiments involved matings between a male and a female mouse to examine damage in the resulting embryos based on dominant lethal mutations. To assess such damage, approximately two weeks after mating, the pregnant females were sacrificed and their uterine contents were examined. For each litter the number of viable implants and the number of non-viable implants were determined, where *viable* was defined before the experiment begins (Lockhart et al., 1991). None of these parent mice were exposed to any toxic chemicals before or during the experiment.

Lockhart et al. (1992) noted that the majority of proportions from these studies exhibited significant departure from the simple binomial model, and considered use of the beta-binomial in (1) to model the overdispersion. A concern of interest was whether this was an adequate assumption, i.e., was there adequate goodness-of-fit for the beta-binomial model? To answer this question, we can apply the bootstrap method from Section 3.2. Our goodness-of-fit results are summarized in Table 4, which are based on the data from Tables 1a-1c.

In Table 4, the number of bootstrap samples used is $M = 100\ 000$ when determining the observed significance level, τ. The table indicates that data sets in Tables 1b and

1c are significant at level 0.05. A more detailed examination of the data set in Table 1c indicated the presence of three unlikely (y, n) pairs: $(7, 7)$, $(9, 9)$, and $(5, 8)$. These appear to be outliers since they occur with very small probabilities under the associated MLE $(\hat{\mu}, \hat{\theta}) = (0.068, 0.064)$. When these three data pairs were removed, the p-value increased from $\tau = 0.000$ to $\tau = 0.054$. This shows that although the outliers partly explain the failure of the beta-binomial distribution, they are not the sole reason for it. Even when the outliers are removed, the test is borderline significant.

Six toxicological data sets analyzed and published by Brooks et al. (1997) also are analyzed by our techniques. Our results are summarized in Table 5, again using $M = 100\ 000$ bootstrap samples. The first two data sets are published in Brooks et al. (1997). Data sets #3-5 were first published by Haseman and Soares (1976), and the sixth one was first published by Aeschbacher et al. (1977). [The tables in Brooks et al. (1997) contain some minor topographical errors: In their Table 1 the entry at position (8, 14) should be moved to (9, 14); in Table 2 the entry of "one" should appear at position (11, 16); in Table 5 the entry at position (9, 10) should be moved to (10, 10).] Our test statistic produced p-values of 0.144, 0.231, 0.000, 0.000, 0.009, and 0.375 for data sets #1-4, respectively. We, therefore, conclude that only data sets #3-5 from Brooks et al. depart significantly from the beta-binomial model.

Although satisfactory results could have been obtained in Tables 4 and 5 with as few as $M = 1000$ bootstrap samples, we used $M = 100\ 000$ because we wanted to determine with high accuracy how well the bootstrap method really works. Using $M = 100\ 000$ and 205 litters, the first data set published in Brooks et al. (1997), requires about 7 1/2 hours on a Sun Ultra 2 computer, and using $M = 1000$ requires less than five minutes. The Fortran program which produced Tables 4 and 5 is applicable to any data set involving proportions and is available at the same web site referenced in Section 4.

6 Discussion

The beta-binomial model is a common choice when analyzing proportion data with some form of litter effect. This model is quite rich, has some intuitive appeal, and is relatively

simple to use since its probability distribution is tractable. Although the model is popular for the above reasons, it has not undergone much goodness-of-fit analysis in the literature. Our method is an attempt at determining simple significance levels for testing fit to the beta-binomial model, and the test seems to be powerful when the data are generated by certain mixture models. In particular, our method is powerful when the alternative model is beta-binomial contaminated with a small proportion of outliers.

Our test statistic, τ, can be extended easily to other models when testing goodness-of-fit. For example, one may wish to model toxicity data by a beta-binomial distribution, where the population mean, μ, is a function of the dose of a chemical given to the dam (Catalano and Ryan, 1994, Section 4). Additional parameters may need to be estimated by maximum likelihood or some other approach, although the basic technique for computing τ, the observed significance level, remains the same. Likewise, goodness-of-fit to other forms of extra-binomial model such as the correlated-binomial model or the beta-correlated-binomial model (Brooks et al., 1997) may be tested using this same basic technique. These models might be considered reasonable alternatives to beta-binomial.

Computation time was the greatest drawback when producing the simulation results in Tables 1 and 2; hence, only a modest simulation study was performed. Re-estimating μ and θ using maximum likelihood for each bootstrap sample is rather time consuming. To save computing time simulations were performed without re-estimating μ and θ, but the estimated sizes were far below the nominal levels and are not shown herein.

We note in closing that a sample of 50 litters may appear large, encouraging use of the χ^2 approximation of Q_n, the individual Pearson statistics, rather than the bootstrap. But in fact, a sample of 50 litters is quite small, relative to asymptotic approximations. For example, if $(\mu, \theta) = (0.05, 0)$ and $J = 50$, then $E_{y,n}$ is much less than 1 for most values of (y, n). Thus, in most practical situations $J = 50$ may not be a large enough number of litters to validate replacing the bootstrap with some asymptotic approximation.

Acknowledgements

Special thanks are due to Drs. Beth Gladen and David Umbach for their helpful suggestions. We also thank Dr. Steve Brooks for verifying that there were some typographical errors in

the tables of Brooks et al. (1997). This research was partially supported by NIMH grant MH53259-01A2 (STG), by NSF grants DMS-9205112 and DMS-9705166 (RLS), and by NCI grant CA76031 (WWP).

REFERENCES

AESCHBACHER, H. U., VUATAZ, L., SOTEK, J. & STALDER, R. (1977) The use of the beta-binomial distribution in dominant-lethal testing for "weak mutagenic activity" (Part 1), *Mutation Research*, 44, pp. 369–390.

BROOKS, S. P., MORGAN, B. J. T., RIDOUT, M. S. & PACK, S. E. (1997) Finite mixture models for proportions, *Biometrics*, 53, pp. 1097–1115.

CATALANO, P. J. & RYAN, L. M. (1994) Statistical methods in developmental toxicology. *In*: G. P. PATIL & C. R. RAO (Eds.), *Handbook of Statistics Volume 12: Environmental Statistics*, pp. 507–534 (New York, North-Holland/Elsevier).

COCHRAN, W. G. (1954) Some methods for strengthening the common χ^2 tests, *Biometrics*, 10, pp. 417–451.

ENGEL, B. & TE BRAKE, J. (1993) Analysis of embryonic development with a model for underdispersion or overdispersion relative to binomial variation, *Biometrics*, 49, pp. 269–279.

GARREN, S. T., SMITH, R. L. & PIEGORSCH, W. W. (2000) On a likelihood-based goodness-of-fit test of the beta-binomial model, *Biometrics*, 56, pp. 947–949.

HASEMAN, J. K. & KUPPER, L. L. (1979) Analysis of dichotomous response data from certain toxicological experiments, *Biometrics*, 35, pp. 281–293.

HASEMAN, J. K. & PIEGORSCH, W. W. (1994) Statistical analysis of developmental toxicity data. *In*: C. KIMMEL & J. BUELKE-SAM (Eds.), *Developmental Toxicology*, 2nd ed., pp. 349–361 (New York, Raven Press).

HASEMAN, J. K. & SOARES, E. R. (1976) The distribution of fetal death in control mice and its implications on statistical tests for dominant lethal effects, *Mutation Research*, 41, pp. 277–288.

HOGG, R. V. & CRAIG, A. T. (1995) *Introduction to Mathematical Statistics*, 5th ed. (Englewood Cliffs, NJ, Prentice Hall).

LEHMANN, E. L. (1983) *Theory of Point Estimation* (New York, Wiley).

LIANG, K.-Y. & MCCULLAGH, P. (1993) Case studies in binary dispersion, *Biometrics*, 49, pp. 623–630.

LOCKHART, A.-M., BISHOP, J. B. & PIEGORSCH, W. W. (1991) Issues regarding data acquisition and analysis in the dominant lethal assay, *Proceedings of the American Statistical Association, Biopharmaceutical Section*, pp. 234–237.

LOCKHART, A.-M., PIEGORSCH, W. W. & BISHOP, J. B. (1992) Assessing overdispersion and dose response in the male dominant lethal assay, *Mutation Research*, 272, pp. 35–58.

MANTEL, N. & PAUL, S. R. (1987) Goodness-of-fit issues in toxicological experiments involving litters of varying size. *In*: I. B. MACNEILL & G. J. UMPHREY (Eds.), *Biostatistics*, pp. 169–176 (New York, Reidel Publishing Company).

PACK, S. E. (1986) Hypothesis testing for proportions with overdispersion, *Biometrics*, 42, pp. 967–972.

PRENTICE, R. L. (1986) Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors, *Journal of the American Statistical Association*, 81, pp. 321–327.

RISKO, K. J. & MARGOLIN, B. H. (1996) Some observations on detecting extra-binomial variability within the beta-binomial model. *In*: B. J. T. MORGAN (Ed.), *Statistics in Toxicology*, pp. 57–65 (Oxford, Clarendon Press).

ROMANO, J. P. (1988) A bootstrap revival of some nonparametric distance tests, *Journal of the American Statistical Association*, 83, pp. 698–708.

TARONE, R. E. (1979) Testing the goodness-of-fit of the binomial distribution, *Biometrika*, 66, pp. 585–590.

WILLIAMS, D. A. (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics*, 31, pp. 949–952.

ZHANG, P. (1999) Omnibus test of normality using the Q statistic, *Journal of Applied Statistics*, 26, pp. 519–528.

TABLE 1a. Frequency of litter sizes with number of non-viable implants in Swiss CD-1 mice, from Lockhart et al. (1992)

		y, number of non-viable implants				
		0	1	2	3	4
	6	-	1	-	-	-
	7	-	-	-	-	-
	8	-	-	-	-	-
	9	1	-	-	-	-
	10	-	-	-	-	-
n,	11	1	4	-	-	1
litter	12	4	2	1	-	-
size	13	-	2	1	1	-
	14	4	2	2	-	1
	15	4	6	2	1	-
	16	3	2	-	-	1
	17	-	1	-	-	-
	18	1	1	-	-	-

Note: This data set consists of $J = 50$ litters.

TABLE 1b. Frequency of litter sizes with number of non-viable implants in Swiss CD-1 mice after sham intraperitoneal injection, from Lockhart et al. (1992)

		y, number of non-viable implants				
		0	1	2	3	4
	2	1	-	-		
	3	2	1	-	-	
	4	-	-	-	1	-
	5	2	-	-	-	-
	6	-	1	-	-	-
	7	3	2	-	-	-
n,	8	3	-	-	1	-
litter	9	9	3	-	-	-
size	10	20	9	1	1	-
	11	42	15	4	-	-
	12	26	15	4	-	-
	13	10	5	4	-	2
	14	5	2	2	-	1
	15	-	1	-	-	2
	16	1	-	-	-	-

Note: This data set consists of $J = 201$ litters.

TABLE 1c. Frequency of litter sizes with number of non-viable implants in $(SEC \times C57L)F_1 \times [(SEC \times C57L)F_1 \times C3H \times 101]$ mice, from Lockhart et al. (1992)

		y, number of non-viable implants									
		0	1	2	3	4	5	6	7	8	9
	4	1	-	-	-	-					
	5	-	1	-	-	-	-				
	6	1	2	-	-	-	-	-			
	7	3	2	-	-	1	-	-	1		
	8	3	-	-	-	-	1	-	-	-	
	9	4	2	1	-	-	-	-	-	-	1
n,	10	12	4	1	1	-	1	-	-	-	-
litter	11	12	9	4	-	1	-	-	-	-	-
size	12	20	11	8	-	-	-	-	-	-	-
	13	38	20	12	2	-	-	-	-	-	-
	14	20	17	5	1	1	-	-	-	-	-
	15	12	10	2	-	-	-	-	-	-	-
	16	4	3	3	1	-	-	-	-	-	-
	17	1	-	1	-	-	1	-	-	-	-
	18	-	-	1	-	-	-	-	-	-	-

Note: This data set consists of $J = 263$ litters.

TABLE 2a. Estimated sizes of proposed test, based on simulations from beta-binomial distributions using 50 litters

		Estimated size at nominal level τ_0			
μ	θ	$\tau_0 = 0.1$	$\tau_0 = 0.05$	$\tau_0 = 0.025$	$\tau_0 = 0.01$
0.05	0.00	0.0765	0.0420	0.0180	0.0060
0.05	0.05	0.1155	0.0675	0.0305	0.0210
0.05	0.10	0.1225	0.0710	0.0395	0.0250
0.10	0.00	0.0855	0.0450	0.0265	0.0160
0.10	0.05	0.1100	0.0655	0.0330	0.0195
0.10	0.10	0.1180	0.0635	0.0260	0.0190
0.15	0.00	0.0835	0.0465	0.0230	0.0155
0.15	0.05	0.1165	0.0625	0.0285	0.0160
0.15	0.10	0.1025	0.0585	0.0300	0.0210

TABLE 2b. Estimated sizes of proposed test, based on simulations from beta-binomial distributions using 100 litters

		Estimated size at nominal level τ_0			
μ	θ	$\tau_0 = 0.1$	$\tau_0 = 0.05$	$\tau_0 = 0.025$	$\tau_0 = 0.01$
0.05	0.00	0.0775	0.0365	0.0185	0.0125
0.05	0.05	0.1270	0.0770	0.0395	0.0250
0.05	0.10	0.1155	0.0615	0.0295	0.0195
0.10	0.00	0.0925	0.0505	0.0285	0.0180
0.10	0.05	0.0980	0.0570	0.0295	0.0220
0.10	0.10	0.1170	0.0600	0.0280	0.0190
0.15	0.00	0.0860	0.0500	0.0255	0.0160
0.15	0.05	0.1110	0.0610	0.0335	0.0275
0.15	0.10	0.0985	0.0605	0.0310	0.0220

TABLE 3a. Estimated power of proposed test, based on simulations from mixture models using 50 litters

beta-binomial		binomial	mixing	Estimated power at nominal level τ_0			
μ	θ	p	probability	$\tau_0 = 0.1$	$\tau_0 = 0.05$	$\tau_0 = 0.025$	$\tau_0 = 0.01$
0.05	0.00	0.7	0.85	0.2530	0.1495	0.0785	0.0545
0.05	0.05	0.7	0.90	0.2510	0.1445	0.0790	0.0595
0.05	0.10	0.7	0.95	0.3475	0.2420	0.1470	0.1070
0.10	0.00	0.8	0.85	0.4200	0.2550	0.1465	0.1115
0.10	0.05	0.8	0.90	0.4635	0.3185	0.1990	0.1360
0.10	0.10	0.8	0.95	0.5215	0.3845	0.2500	0.1875
0.15	0.00	0.9	0.85	0.5465	0.3915	0.2555	0.1970
0.15	0.05	0.9	0.90	0.5720	0.4355	0.2855	0.2225
0.15	0.10	0.9	0.95	0.6600	0.5295	0.3675	0.2910

TABLE 3b. Estimated power of proposed test, based on simulations from mixture models using 100 litters

beta-binomial		binomial	mixing	Estimated power at nominal level τ_0			
μ	θ	p	probability	$\tau_0 = 0.1$	$\tau_0 = 0.05$	$\tau_0 = 0.025$	$\tau_0 = 0.01$
0.05	0.00	0.7	0.85	0.2850	0.1410	0.0705	0.0445
0.05	0.05	0.7	0.90	0.2370	0.1300	0.0605	0.0430
0.05	0.10	0.7	0.95	0.3195	0.1960	0.1145	0.0725
0.10	0.00	0.8	0.85	0.5735	0.3850	0.2005	0.1340
0.10	0.05	0.8	0.90	0.4880	0.3070	0.1685	0.1065
0.10	0.10	0.8	0.95	0.5685	0.4025	0.2475	0.1775
0.15	0.00	0.9	0.85	0.8050	0.6615	0.4825	0.3800
0.15	0.05	0.9	0.90	0.7045	0.5700	0.3975	0.3040
0.15	0.10	0.9	0.95	0.7850	0.6350	0.4500	0.3520

TABLE 4. Analysis of dominant lethal data

Table number	Number of litters	MLE $\hat{\mu}$	MLE $\hat{\theta}$	p-value τ
1a	50	0.075	0.021	0.333
1b	201	0.051	0.041	0.010
1c	263	0.068	0.064	0.000

TABLE 5. Analysis of data sets published by Brooks et al. (1997)

Data set number	Number of litters	MLE $\hat{\mu}$	MLE $\hat{\theta}$	p-value τ
1	205	0.090	0.074	0.144
2	211	0.112	0.111	0.231
3	524	0.090	0.073	0.000
4	1328	0.109	0.045	0.000
5	554	0.074	0.081	0.009
6	127	0.069	0.063	0.375