BOOTSTRAP GOODNESS-OF-FIT TESTS FOR THE BETA-BINOMIAL MODEL

S. Garren, R. Smith and W. Piegorsch

January 1994

Mimeo Series #2314

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

MIMEO SERIES #2314

S. Garren, R. Smith and W. Piegorsch

BOOTSTRAP GOODNESS-OF-fit TEST FOR THE BETA-BINOMIAL MODEL

NAME

DATE

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

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

[1]Department of Statistics, University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina 27599-3260, U.S.A.

and

[2]Statistics and Biomathematics Branch, National Institute of Environmental Health Sciences,
Research Triangle Park, North Carolina 27709, U.S.A.

SUMMARY

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 teratological experiments. Although there are many procedures for testing the binomial distribution against overdispersed alternatives, there apparently has been very little consideration of how to test the null hypothesis of a beta-binomial distribution against all other distributions. One difficulty in the construction of such tests is the lack of homogeneity in the data when there is large variability in the individual litter sizes. Herein, we propose a test statistic based on combining Pearson statistics from individual litter sizes, and evaluate its null distribution by both exact calculation and simulation. A randomized form of the test also is proposed, which seems to improve on the nonrandomized version in some cases. A Monte Carlo study confirms the accuracy and power of the test, and the method is applied to real data sets. The test also is extended to the situation of multiple dose levels.

---

† *Corresponding author.*

* *Current address:* Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, U.S.A.

Key words: Beta-binomial distribution; Bootstrap; Kolmogorov-Smirnov goodness-of-fit; Logistic regression; Monte Carlo simulation; Overdispersion; Pearson statistic.

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 bounded above by the positive integer n. When n is assumed fixed and known, a natural sampling model for the distribution of Y is the binomial distribution, $Y \sim \text{Binomial}(n,p)$, where p is an unknown value between 0 and 1 representing the mean proportion response, $\text{E}\left(\frac{Y}{n}\middle| n\right) = p$.

A common characterization that leads to binomial sampling is where Y is the sum of n independent Bernoulli random variables, W_m $(m = 1,\cdots,n)$, each with binary response probability p. The independence assumption here is critical. If there were some correlation 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, where the W_m represent the binary responses of fetuses or embryos within a litter (of size n) from a female rodent exposed to some toxic stimulus (Piegorsch and Haseman, 1991). 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. This has been called a *litter effect* and often is modeled hierarchically. For instance, one might view the litter effect as inducing variability in the per-fetus response parameter p. If p varies according to a beta distribution, and if the conditional distribution of Y given n and p is taken as binomial, then the marginal density of Y (given n) is beta-binomial (Williams, 1975; Haseman and Kupper, 1979). We will describe the construction of this model in greater detail in §2, below.

Testing for departure from the binomial distribution is a well-known problem which may be approached from a number of perspectives. Perhaps the best known is Cochran's (1954) binomial variance test

$$\chi^2_B = \sum_{i=1}^{J} \frac{(Y_i - n_i\widehat{p})^2}{n_i\widehat{p}\,\widehat{q}},$$

in which

$$\widehat{p} = \frac{\sum_{i=1}^{J} Y_i}{\sum_{i=1}^{J} n_i},$$

$\widehat{q} = 1 - \widehat{p}$, and J is the number of litters. Another test is Tarone's (1979) $C(\alpha)$-test for binomial overdispersion

$$\chi^2_C = \frac{\left(\sum_{i=1}^{J} \frac{(Y_i - n_i\widehat{p})^2}{\widehat{p}\,\widehat{q}} - \sum_{i=1}^{J} n_i\right)^2}{2\sum_{i=1}^{J} n_i(n_i - 1)}.$$

The binomial variance statistic is compared to a χ^2 distribution with $J-1$ degrees of freedom (df), while the $C(\alpha)$ statistic is compared to a χ^2 distribution with 1 df. Both test statistics possess good statistical power to detect departures from binomial variability, with the $C(\alpha)$-test being optimal against certain forms of overdispersion. In particular, the $C(\alpha)$-test is optimal against the beta-binomial distribution (Tarone, 1979).

In cases with strong evidence of extra-binomial variability, the binomial distribution is not a reasonable null hypothesis. In view of the popularity of the beta-binomial distribution for modeling extrabinomial variation, it is desirable to test the null hypothesis of a beta-binomial distribution against all other alternatives. In contrast to the binomial distribution, however, there has been very little formal testing of whether the beta-binomial distribution fits observed toxicity data. In a recent study of the statistical features and sources of variability in an assay for heritable mutagenesis (a form of developmental toxicity), Lockhart et al. (1992) questioned whether the beta-binomial is a valid sampling model for the overdispersed binomial data that they encountered. Their results were inconclusive, but identifying departure from this sampling model is important.

In what follows, we will explore the use of a bootstrapped chi-square method for testing goodness-of-fit to the beta-binomial model. The method begins by estimating the two unknown parameters of the (single-sample) beta-binomial distribution via maximum likelihood (MLE). With these estimates, one computes a Pearson-type goodness-of-fit statistic for each litter size observed in the sample. For each test statistic corresponding to an observed litter size, its separate significance level is computed by simulating beta-binomial pseudo-data based on the calculated MLEs for the observed sample, and finding the associated Pearson pseudo-statistics. The beta-binomial goodness-of-fit statistic is then determined by the smallest level of significance among all litter sizes. Details for construction of the test statistic are given in §3. Section 4 gives simulation results illustrating the test's conservative behavior under the null hypothesis and power against mixtures of binomial distributions, while §5 illustrates the method on data from a developmental toxicity experiment. Section 6 shows how the method can be extended to include cases where the pregnant rodents are not identically distributed. Section 6 also includes an application to teratogenicity data. We end in §7 with a short discussion.

2. Description of Beta-Binomial Model

One characterization of the beta-binomial model employs the following hierarchy: Assume that a given experiment consists of J litters of animals, and that n_i is the number of pups in the $i^{\underline{th}}$ litter, for $i = 1, \cdots, J$. The litter sizes n_i are treated as fixed constants. Let Y_i denote the number of responses in the $i^{\underline{th}}$ litter, for $i = 1, \cdots, J$. Conditional on $\{p_i, i = 1, \cdots, J\}$, the $\{Y_i, i = 1, \cdots, J\}$ are independent binomial random variables

$$(Y_i \mid n_i, p_i) \sim \text{Binomial}(n_i, p_i)\ (i = 1, \cdots, J)$$

with mean $n_i p_i$ and variance $n_i p_i(1-p_i)$, $i = 1, \cdots, J$. The random variables $\{p_i, i = 1, \cdots, J\}$ are independent and have the common beta density

$$\mathfrak{B}(p \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha - 1}(1-p)^{\beta - 1} \ (0 < p < 1) \ ,$$

where α and β are unknown positive constants, and the beta function $B(\cdot, \cdot)$ is defined by

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \ .$$

The unconditional distribution of Y_i is expressed by the beta-binomial probability

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

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

$$\mu = \frac{\alpha}{\alpha + \beta} \text{ and } \phi = \frac{1}{\alpha + \beta} \tag{2}$$

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

$$\mathrm{E}(Y_i \mid n_i, \mu, \phi) = n_i \mu$$

and

$$\mathrm{var}(Y_i \mid n_i, \mu, \phi) = n_i \mu (1 - \mu) \frac{1 + n_i \phi}{1 + \phi}$$

for $i = 1, \cdots, J$. The parameter μ may be referred to as the *mean parameter* of the marginal proportions. (Notice that $0 < \mu < 1$.) Also, note that as $\phi \downarrow 0$, the variance of $(Y_i \mid n_i, \mu, \phi)$ monotonically decreases to $n_i \mu(1-\mu)$, and $(Y_i \mid n_i, \mu, \phi)$ converges in distribution to a binomial random variable. Hence, the parameter ϕ is said to be the *dispersion parameter*, and if $\phi > 0$, then data generated by such a beta-binomial distribution are said to be *overdispersed*. The MLE of (μ, ϕ) can be shown to be a consistent estimator of (μ, ϕ) (Lehmann, 1983, pp. 409-413) and is determined numerically.

3. Approaches to Goodness-of-Fit Testing

Suppose the data consist of independent pairs $\{(Y_i, n_i), i = 1, \cdots, J\}$ as described in §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, ϕ)

parametrization (2) to represent a specific member of the beta-binomial family.

Since the distribution is discrete, the often used goodness-of-fit statistics are Pearson's chi-squared statistic and the likelihood ratio statistic. Neither is easy to adapt to this problem, however. Pearson's chi-squared statistic would be a convenient choice if all of the litter sizes were equal, say $n_1 = \cdots = n_J = n \geq 2$, since then there are $n+1$ multinomial cell probabilities and two estimated parameters, so the distribution of the test statistic would converge to χ^2_{n-2} as $J \to \infty$.

If the litter sizes are not all equal, but belong to a finite set $\{n_1, \cdots, n_K\}$ say, then one approach would be to compute a separate Pearson statistic, conditional on the litter size, for each of $n_1, \cdots, n_K$, followed by combining them into an overall test statistic. This approach would be most reasonable if K were small and there were a large number of litters at each of the litter sizes $n_1, \cdots, n_K$. In the problems that motivated this work, however, there are a large number of different litter sizes with relatively few litters at each individual n_i. In this situation it would be much less reasonable to suppose the test statistic to have even approximately a χ^2 distribution.

Another approach (Mantel and Paul, 1987, pp. 169-176) is to assume that the $\{n_i, i = 1, \cdots, J\}$ are themselves random variables from some known distribution, and to base the Pearson statistic on the unconditional probability distribution of the $\{Y_i\}$ using the MLE of (μ, ϕ). This approach, however, loses all information about the individual litter sizes when determining the observed numbers of the $\{Y_i\}$, and thus could conceal large variations in the proportion of responses among litters. For this reason, we find their approach unsatisfactory.

At first sight, it might appear that a likelihood ratio statistic would avoid this difficulty of combining different litter sizes, but a natural choice for the alternative model does not seem to exist. Lockhart et al. (1992) used the likelihood ratio statistic for testing the beta-binomial model against the alternative that the Y_i have independent binomial distributions with parameters (n_i, p_i), where $p_1, \cdots, p_J$ are treated as arbitrary unknown constants. Since the number of parameters p_i to be estimated in the alternative model goes to infinity as $J \to \infty$, then the usual asymptotic theory developed for likelihood ratio tests cannot necessarily be applied to their test statistic. Even if this difficulty could be avoided through a simulated or exact calculation of the significance level, consistency of such a test still is not guaranteed. As an example to illustrate this, consider the model in which $(y_i, n_i) = (1, 2)$ with probability one for all i. It is easy to verify that in this case the maximized likelihood is the same under both the beta-binomial and independent binomial models. Therefore, a likelihood ratio test would in this case erroneously accept the null hypothesis.

Liang and McCullagh (1993) have given a general discussion of estimating equation approaches to overdispersed binary data. Among other things, they proposed a test for whether the mean-variance relationship across different litter sizes is consistent with the beta-binomial model. In this case, unfortunately, one can construct non-beta-binomial distributions which would pass the Liang-

McCullagh test, so their test is not consistent against certain models that are not beta-binomial, and hence does not entirely fulfill our intended purpose.

The difficulties of constructing a simple, consistent test prompt us to return to the Pearson statistic, but with some modifications to combine the different litter sizes and to use bootstrapping or exact calculations to determine the null distribution. This approach is in line with the general approach to bootstrapping goodness-of-fit statistics advocated by Romano (1988).

Initially, we assume that the beta-binomial parameters, μ and ϕ, are known, and that all of the litter sizes are fixed and known. Specifically, 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 each $J_k > 0$ $(k = 1, \cdots, K)$ 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 $n = n_k$ $(k = 1, \cdots, K)$, let $O_{y,n}$ denote the observed number of litters of size n which contain y responses, and let $E_{y,n}$ denote the expected number under a specified beta-binomial model, assuming J_k litters of this size. Then define an individual Pearson statistic as

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

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

2. Determine the lower tail probabilities of the q_n by simulation. Thus, for each litter size $n = n_k$ $(k = 1, \cdots, K)$, generate a large number $J_n \cdot M$ of beta-binomial pseudo-random samples with parameters (μ, ϕ), where M is independent of n. Then, repeat the calculation in step 1 to generate a bootstrap sample $Q^*_{n,1}, \cdots, Q^*_{n,M}$ such that each $Q^*_{n,m}$ is based on J_n litters. Based on this sample compute

$$\rho_n = \frac{\text{Number of times } Q^*_{n,m} < q_n}{M}. \tag{3}$$

so that ρ_n is the estimated lower tail probability associated with q_n. The null distribution of ρ_n is approximately uniform on (0,1); notice, however, that this approximation is not an exact distribution, even in the limit as $M \to \infty$, because the distribution of Q_n is discrete.

3. Combine the tail probabilities ρ_n by computing

$$\tau = 1 - \left[\max_{k=1,\cdots,K} \rho_{n_k}\right]^K. \tag{4}$$

Intuitively, we should reject the null hypothesis if any of the ρ_n's is too large, i.e., if τ is too small. The power transformation in (4) ensures that, if it is indeed the case that each ρ_n has a uniform distribution, then τ also has a uniform distribution on (0,1) and is the appropriate p-value. Thus, an

approximately level τ_0 test is obtained by rejecting the null hypothesis whenever $\tau < \tau_0$.

This method is preferable to the perhaps more obvious strategy of bootstrapping a weighted sum of the Q_n. We would want to use a weighted sum because, for instance, one litter of size 11 should not influence the test statistic as much as ten litters of size 10. However, determining what the weights should be is not so easy. We therefore do not consider this strategy any further.

In practice, there are a number of modifications to this approach:

(a) When J_n is small, it is feasible to compute the exact distribution of Q_n, so this may replace the simulation in step 2. In that case, we define $\rho_n = \Pr(Q_n < q_n)$, but note that this version of ρ_n still is not exactly uniformly distributed because of the discrete nature of the observations.

(b) This test based on steps 1–3 is conservative, because the inequality in (3) means that ρ_n is stochastically smaller than an exact uniform (0,1) random variable, so $\Pr(\tau < \tau_0) \leq \tau_0$. One also may define an exact randomized test, in which the definition $\rho_n = \Pr(Q_n < q_n)$ is replaced by

$$\rho_n = \Pr(Q_n < q_n) + U \cdot \Pr(Q_n = q_n) ,$$

U having an independent uniform (0,1) distribution. This randomized statistic ρ_n, however, is exactly uniform (0,1). There is an obvious bootstrap analog for this in which both $\Pr(Q_n < q_n)$ and $\Pr(Q_n = q_n)$ are estimated by their bootstrap values as in (3). This test has the disadvantage of a randomized outcome, but we expect it to be more powerful than the original test.

(c) In the practical situation where (μ, ϕ) are unknown, we calculate their maximum likelihood estimators $(\hat{\mu}, \hat{\phi})$ and use them to generate the bootstrap samples. The rest of the procedure is as outlined above. Note that we do not re-estimate μ and ϕ in each individual bootstrap sample unless specified otherwise; in principle re-estimation would give a more accurate test, but the computational time of such a procedure renders it impractical when resources are limited. We, nevertheless, did perform a few simulations using re-estimated μ and ϕ, and the results are summarized at the end of §4.

4. Monte Carlo Results

We performed a Monte Carlo study to determine how well the tests proposed in §3 operate in practice. The simulations were carried out for two models, one a beta-binomial distribution, used to examine the level accuracy (size) of the tests, the other a mixture of two binomial distributions, used as an indicator of their power. Both sets of simulations used litter sizes from a toxicological study considered below (trial number 10 in Table 3) in which there were 50 litters of sizes ranging from 6 to 18. Simulations using 100 litters also were carried out by using each litter size from this same trial twice. The parameters of the beta-binomial distribution were allowed to vary over all combinations of $\mu \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$ and $\phi \in \{0, 0.05, 0.1\}$. In the alternative distribution, the litter response

probabilities p_i were random with $\Pr(p_i = \mu_1) = 1 - \Pr(p_i = \mu_2)$ equal to some specified mixing probability. Various values of μ_1, μ_2, and the mixing probability were selected, although not all combinations of these three parameters were used. Whenever $(n+1)^J n > 1000$, which occurs for $11 \leq n \leq 16$ when $J = 50$ and also for $n = 18$ when $J = 100$, the bootstrap method is used to compute ρ_n with $M = 1000$ since the number of litters of size n is considered too large to compute ρ_n exactly. Otherwise, ρ_n is computed exactly. The values of ρ_n were calculated exactly for $n \leq 10$ and $n = 17$, otherwise by bootstrapping with $M = 1000$. Randomized test statistics were calculated for all of these simulations. Nonrandomized test statistics were calculated for two of the beta-binomial distributions to demonstrate that in some cases the nonrandomized test is unnecessarily more conservative than the randomized test. In other cases, however, the randomized and nonrandomized tests are practically identical. The results are based on 2000 or 9000 replications carried out in Fortran 77 on a Sun Sparcstation with the pseudo-random uniform number generator proposed by Marsaglia et al. (1990). Pseudo-random beta-binomial variates were generated by determining the exact cumulative beta-binomial distribution and by comparing the uniform pseudo-variates to this distribution. Pseudo-random variates from the mixture model were generated similarly.

The sizes and the power of the test are summarized in Tables 1 and 2, respectively. Figures 1 through 3 are histograms of τ for the beta-binomial model for $\mu = 0.4$ and $\phi = 0.1$ using the nonrandomized test statistic with 50 litters, the randomized test statistic with 50 litters, and the randomized test statistic with 100 litters, respectively. Figure 4 is a histogram of the randomized τ for the mixture of binomials model using 100 litters, mixing probability $= 0.8$, $\mu_1 = 0.3$, and $\mu_2 = 0.8$. Reasonably accurate estimates of the sizes and power were attainable using only 2000 replications, but 9000 replications were used whenever τ was graphed in Figures 1 through 4 so that these graphs would better reflect the true distributions. The simulated standard error of any of these estimated sizes is approximated by

$$\sqrt{\frac{(\text{estimated size}) \cdot [1 - (\text{estimated size})]}{\#\text{ of replications}}},$$

and a similar expression is obtained for the standard error of an estimated power. Thus, for example, an estimated size of 0.01 from 2000 replications has the standard error approximately equal to 0.0022. Notice that each of Figures 1 through 4 is the graph of simulations summarized in Table 1 or 2.

The lower tail of Figure 1, which uses the nonrandomized test statistic, is too small compared to that of a uniform distribution and thus reveals that the nonrandomized test is too conservative. As shown in Table 1, the simulations determine that the nonrandomized test statistic τ using $\mu = 0.4$, $\phi = 0.1$, and 50 litters has estimated sizes 0.0611, 0.0276, 0.0119, and 0.0063 at nominal significance levels 0.1, 0.05, 0.025, and 0.01, respectively. Moreover, the Kolmogorov-Smirnov goodness-of-fit test

statistic for these simulations based upon the assumption of uniformity is 0.0671, which is significantly large and thus indicates that τ departs strongly from uniformity. Upper tail critical values at the 0.05 nominal significance level for the Kolmogorov-Smirnov goodness-of-fit test statistic are 0.0303 and 0.0143 for 2000 and 9000 replications, respectively. For a discussion of the Kolmogorov-Smirnov goodness-of-fit test statistic, see, for example, Lawless (1982, pp. 431-438).

Figure 2 represents the same simulations of fetal responses as Figure 1, but Figure 2 uses the randomized test statistic instead. This figure is somewhat closer to being uniform than Figure 1 over its entire support, including the lower tail where rejection of the beta-binomial model occurs. As shown in Table 1, the simulations represented by Figure 2 determine that this randomized test has estimated sizes 0.0798, 0.0396, 0.0170, and 0.0081 at nominal levels 0.1, 0.05, 0.025, and 0.01, respectively. This is clearly more satisfactory than the simulations from Figure 1. Yet, Figure 2 still does not fit perfectly a uniform distribution as exemplified by the value of its Kolmogorov-Smirnov goodness-of-fit test statistic of 0.0419. This lack-of-fit to a uniform distribution in Figure 2 is attributed to the fact that (μ, ϕ) is estimated; simulated values (not shown) of τ using (μ, ϕ) instead of $(\widehat{\mu}, \widehat{\phi})$ easily pass uniformity tests.

The fit of τ to a uniform distribution can be improved further by increasing the number of litters. Using a larger number of litters tends to increase the Pearson statistics $\{Q_n\}$ to less conservative levels since each $E_{y,n}$ becomes less influenced by its corresponding value of $O_{y,n}$. Figure 3, again representing simulations from the randomized test, uses the same model and parameters as does Figure 2 except 100 litters are used instead of 50. As shown in Table 1, these simulations determine that this beta-binomial goodness-of-fit test has estimated sizes 0.0879, 0.0416, 0.0183, and 0.0102 at nominal levels 0.1, 0.05, 0.025, and 0.01, respectively. This is an improvement over Figure 2, and Figure 3 fits a uniform distribution much better than does Figure 2 as exemplified by the former's Kolmogorov-Smirnov goodness-of-fit test statistic value of 0.0313, although this still is significantly non-uniform.

Additional values of (μ, ϕ) were used for simulating estimated sizes and Kolmogorov-Smirnov goodness-of-fit test statistics in Table 1. The estimated sizes presented in the table indicate that most of our choices of (μ, ϕ) produce conservative tests for departure from the beta-binomial model. Furthermore, the estimated sizes and Kolmogorov-Smirnov goodness-of-fit test statistics indicate that τ tends to approach a uniform distribution more closely as the number of litters increases from 50 to 100 and as μ increases from 0.05 to 0.4. No general observation concerning how the choice of ϕ affects uniformity of τ is made other than the fact that choosing $\phi = 0$ tends to produce more conservative beta-binomial goodness-of-fit tests and less uniform simulations of τ than does $\phi = 0.05$ or $\phi = 0.1$. This lack of uniformity of τ when using $\phi = 0$ possibly may be explained by the fact that the support of ϕ does not include negative values, but this idea has not been adequately examined.

Simulations for determining power under the mixture of binomials model are summarized in

Table 2, and an illustrative histogram of τ is shown in Figure 4. All of these simulations used the randomized test. Figure 4 is grossly non-uniform, as is corroborated by its Kolmogorov-Smirnov goodness-of-fit test statistic of 0.4327. The simulations determine that this test has estimated power of 0.3557, 0.2162, 0.1206, and 0.0803 at nominal levels 0.1, 0.05, 0.025, and 0.01, respectively.

The simulations of the mixture of binomials model reveal that some choices of parameters and numbers of litters produce very high power while other choices produce very low power. Furthermore, these simulations confirm that increasing the number of litters from 50 to 100 sometimes dramatically increases the power. For example, when the mixing probability is 0.6, $\mu_1 = 0.3$, and $\mu_2 = 0.9$, the estimated power increases from 0.2580 to 0.5250 at the 0.1 nominal level as the number of litters increases from 50 to 100, as shown in Table 2. Perhaps the most striking result from Table 2 is the fact that the power against mixture of binomials models increases but then decreases as the mixing probability increases for fixed μ_1 and μ_2. Large mixing probabilities ($\simeq 0.95$) not too close to one tend to produce at least one relatively extreme value of y, which increases its corresponding statistics $O_{y,n}$ and ρ_n and decreases τ, since the $E_{y,n}$ usually are not greatly influenced by a small number of extreme values of y. Mixing probabilities which are somewhat larger (≥ 0.99, but < 1) do not produce extreme values of y as often and thus lead to less power, since the simulated data tend to closely fit a binomial distribution with parameter μ_1 when no extreme values of y exist. Using a mixing probability near 0.5 often produces simulated data which fit a beta-binomial model better than using a large mixing probability and thus results in less power. For example, Table 2 shows that when using $\mu_1 = 0.3$, $\mu_2 = 0.9$, and 100 litters, the estimated power at the 0.1 nominal level are 0.5250, 0.5670, 0.9760, and 0.6095 for mixing probabilities 0.6, 0.7, 0.95, and 0.99, respectively. In addition, further simulations (not shown) reveal that data consisting of 10 and 20 litters result in more conservative and less powerful beta-binomial goodness-of-fit tests than do data consisting of 50 and 100 litters.

As mentioned at the end of §3, we did perform some simulations using re-estimated values of μ and ϕ. Under this scenario, each value of ρ_n is calculated from 1000 bootstrap samples, never from the exact distribution, and the randomized test statistic τ is replicated 2000 times. Under the beta-binomial model with $\mu = 0.05$, $\phi = 0$ and 50 litters, the estimated sizes are 0.0755, 0.0415, 0.0215, and 0.0135 for nominal levels 0.1, 0.05, 0.025, and 0.01, respectively. This method clearly is an improvement over the method where μ and ϕ are not re-estimated as shown in Table 1, since the estimated sizes for the latter case are 0.0300, 0.0100, 0.0035, and 0.0015 at the same nominal levels. Under the beta-binomial model with $\mu = 0.4$, $\phi = 0.1$, and 50 litters, the estimated sizes using the re-estimated μ and ϕ are 0.1130, 0.0580, 0.0275, and 0.0195 at the same nominal levels. Observe, however, that 0.0195 is significantly larger than 0.01. Again, these estimated sizes are larger than the corresponding estimated sizes where μ and ϕ are not re-estimated, as shown in Table 1, where these

latter estimated sizes are 0.0798, 0.0396, 0.0170, and 0.0081, respectively. Although these simulations indicate that using re-estimated μ and ϕ might produce nonconservative goodness-of-fit tests, we have not investigated this further because of the large amounts of computational time required. In practice, however, re-estimation may be a more useful approach if sufficient computer resources are available. The rest of this paper discusses only situations where μ and ϕ are not re-estimated.

5. An Example

As an illustration, we applied our goodness-of-fit test statistic to data provided by Dr. W. M. Generoso from sixteen experiments involving pregnant mice. The experiments involve mating between a male and a female mouse to examine damage in the resulting embryos based on so-called dominant lethal mutations. To assess such damage, approximately two weeks after mating, the pregnant females are sacrificed and their uterine contents are examined. For each litter the number of viable implants and the number of non-viable implants are determined, where a strict definition of *viable* is adopted before the experiment begins (Lockhart et al., 1991). None of these parent mice are exposed to any chemicals before or during the experiment, except for data set #16, where each dam is given a constant dose of a chemical prior to mating. A more detailed description of these data sets can be found in Lockhart et al. (1992). Data sets #1–10, excluding data set #1*, were analyzed originally by Lockhart et al. (1992). Data sets #11–15 were obtained similarly. Data set #1* contains the same data as data set #1 except for the removal of three (y,n) pairs: (7,7), (9,9) and (5,8). These appear to be outliers since they occur with very small probabilities under $\widehat{\mu} = 0.059780$ and $\widehat{\phi} = 0.022420$, as produced by data set #1*.

As in §4, M bootstrap samples are generated when and only when $(n+1)^{J_n} > G$ (to be defined) to calculate ρ_n; otherwise, ρ_n is calculated exactly. For each data set the observed level of significance is determined by two different methods. The first method involves computing the test statistic τ from nonrandomized values of ρ_n and letting $G = 160{,}000$ and $M = 1{,}000{,}000$. These large values of G and M were chosen since much computer time was available but were not necessary in order to obtain accurate results. The second method involves estimating the average value of τ from 250 randomized tests and letting $G = 1000$ and $M = 1000$. We do not claim that this average value of τ is a valid significance level; rather it is average of valid significance levels. The standard error of this average τ using the randomized method is determined by the product of the sample deviation of τ and $1/\sqrt{250}$. Although standard errors of the nonrandomized τ are not listed in Table 3, upper bounds on these errors can be determined. The standard error on ρ_n using the bootstrap is

$$\text{s.e.}(\rho_n) \simeq \sqrt{\rho_n(1-\rho_n)/M}\ ,$$

and s.e.$(\rho_n) = 0$ when exact probabilities are used rather than bootstrap estimates. Therefore, using a Taylor expansion, the approximate standard error in estimating the nonrandomized τ can be shown to be

$$\text{s.e.}(\tau_{\text{nonrandom}}) \leq K \left[\text{s.e.}(\max_m \rho_m)\right] \left[\max_n \rho_n\right]^{K-1}$$

$$= (1 - \tau_{\text{nonrandom}})\, K \sqrt{\frac{1 - \max_m \rho_m}{M \max_n \rho_n}},$$

which is quite small for $M = 1{,}000{,}000$ and large τ.

For all of the data sets in Table 3 except data sets #7 and #11, the randomized test produces a smaller observed significance level than does the nonrandomized test. Based on the standard errors of the sample mean $\bar{\tau}_{\text{random}}$, there is, however, no significant difference between the observed significance levels from the randomized and the nonrandomized tests for data sets #7 and #11. Use of a randomized test decreases the observed significance level in data set #13 from 0.1490 to 0.0966 and decreases the observed significance level in data set #14 from 0.0583 to 0.0318. According to the randomized test, data sets #1, #2, #7, #8, and #14 are significant at the 0.05 nominal level. Note, however, that removing the three rare data values from data set #1, which produces data set #1*, increases the observed significance level from 0.0000 to 0.0821 using the randomized test.

6. An Extension of Tests to Multiple Doses

Our goodness-of-fit test statistic can be extended to data where the pregnant mothers are not treated homogeneously. Suppose in reproductive toxicology experiments that the pregnant animals are administered particular doses of a chemical before mating. The response probability of a fetus is now a function of the dose of the chemical. Williams (1982) suggested modeling the mean response probability μ_i by a logistic model via

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = a + b d_i \; (i = 1, \cdots, L),$$

where d_i is the dose administered to the $i^{\underline{th}}$ mother, L is the number of dose levels, and a and b are constants.

The goal is to test whether or not data were generated by a family of beta-binomial distributions, where the mean response probability follows logistic regression and the dispersion parameter ϕ is constant across dose levels. Notice that there are now three unknown parameters, a, b, and ϕ, and these can be estimated by maximum likelihood. The test statistic becomes

$$\tau^{\dagger} = 1 - \left[\max_{k,i} \rho_{n_k, d_i}\right]^{K^{\dagger}},$$

where $K^\dagger$ is the number of litter sizes represented by the data; i.e.,

$$K^\dagger = \sum_{n=1}^{\infty} \sum_{\forall d} I(\exists \text{ at least one litter of size } n \text{ at dose } d) ,$$

and $\rho_{n,d}$ is similar to ρ_n but restricted to dose d.

This approach for determining goodness-of-fit of the beta-binomial model using logistic regression to link different dose groups is applied to data provided by Dr. Keith Soper and published by Liang and McCullagh (1993). These data represent fetal deaths from a standard teratological experiment, where dosages of a chemical of sizes 0, 2, 4, and 8 mg/kg were administered to a total of 82 pregnant mice. Both randomized and nonrandomized versions of the test statistic are performed on these data, where ρ_n is determined exactly or by M bootstrap samples according to the same rule and values of G and M presented in §5.

The results from applying our goodness-of-fit test statistic to these data involving multiple doses can be summarized as follows: The MLE of (a,b,ϕ), denoted by $(\widehat{a},\widehat{b},\widehat{\phi})$, is (–4.313602, 0.131012, 0.043400). The value of the nonrandomized τ is 0.3891, and the sample mean of the randomized τ is 0.2351, where the standard error of this sample mean is 0.0077. Since this sample mean is many standard errors away from 0.05, departure from the beta-binomial model with logistic regression is not significant for this data set.

7. Discussion

The beta-binomial model frequently is used in the literature when analyzing data having binomial response within litters. This model is quite rich, has some intuitive appeal, and is relatively simple to use since its probability distribution is tractable and can be generated easily from uniform random variables. Although the model is popular for the above reasons, it has not undergone much goodness-of-fit analysis. As noted above, standard goodness-of-fit approaches are difficult to employ. The method considered herein 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 mixtures of binomial random variables or when a small number of rare values exist in the data. The effect of using estimated values of (μ,ϕ), however, is the greatest weakness in our test and should be examined thoroughly before concluding significance. Furthermore, using a logit link when multiple doses are available allows a natural extension of our test.

ACKNOWLEDGEMENTS

Special thanks are due to Drs. Beth Gladen, Norman Kaplan, Barry Margolin, and David Umbach for their helpful suggestions and Drs. W. M. Generoso and Keith Soper for providing data leading to the

results in §§5 and 6.

REFERENCES

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

Haseman, J. K., and Kupper, L. L. (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics* **35**, 281-293.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, New York: John Wiley and Sons.

Lehmann, E. L. (1983). *Theory of Point Estimation*, New York: John Wiley and Sons.

Liang, K.-Y., and McCullagh, P. (1993). Case studies in binary dispersion. *Biometrics* **49**, 623-630.

Lockhart, A.-M., Bishop, J. B., and Piegorsch, W. W. (1991). Issues regarding data acquisition and analysis in the dominant lethal assay. *Proceedings of the American Statistical Association, Biopharmaceutical Section*, 234-237.

Lockhart, A.-M., Piegorsch, W. W., and Bishop, J. B. (1992). Assessing overdispersion and dose response in the male dominant lethal assay. *Mutation Research* **272**, 35-58.

Mantel, N., and Paul, S. R. (1987). Goodness-of-fit issues in toxicological experiments involving litters of varying size. In *Biostatistics*. New York: Reidel Publishing Company.

Marsaglia, G., Zaman, A., and Tsang, W. W. (1990). Toward a universal random number generator. *Statistics and Probability Letters* **8**, 35-39.

Piegorsch, W. W., and Haseman, J. K. (1991). Statistical methods for analyzing developmental toxicity data. *Teratogenesis, Carcinogenesis, and Mutagenesis* **11**, 115-133.

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

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

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

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics* **31**, 144-148.

Table 1
Estimated sizes of proposed test, based on simulations from beta-binomial distributions

			Estimated size at nominal level τ_0				Kolmogorov-Smirnov
μ	ϕ	Number of litters	τ_0=0.1	τ_0=0.05	τ_0=0.025	τ_0=0.01	statistic
0.05	0.00	50	0.0300	0.0100	0.0035	0.0015	0.1143
0.05	0.00	100	0.0420	0.0160	0.0050	0.0020	0.1132
0.05	0.05	50	0.0520	0.0230	0.0075	0.0035	0.0821
0.05	0.05	100	0.0745	0.0340	0.0125	0.0035	0.0688
0.05	0.10	50	0.0505	0.0155	0.0040	0.0010	0.0850
0.05	0.10	100	0.0780	0.0360	0.0150	0.0075	0.0773
0.10	0.00	50	0.0385	0.0170	0.0085	0.0030	0.1087
0.10	0.00	100	0.0510	0.0235	0.0085	0.0040	0.1066
0.10	0.05	50	0.0640	0.0300	0.0110	0.0015	0.0667
0.10	0.05	100	0.0775	0.0380	0.0155	0.0080	0.0423
0.10	0.10	50	0.0700	0.0305	0.0140	0.0080	0.0842
0.10	0.10	100	0.0765	0.0390	0.0175	0.0085	0.0528
0.20	0.00	50	0.0525	0.0250	0.0085	0.0030	0.0925
0.20	0.00	100	0.0580	0.0220	0.0110	0.0070	0.0996
0.20	0.05	50	0.0730	0.0365	0.0170	0.0075	0.0608
0.20	0.05	100	0.0860	0.0445	0.0190	0.0080	0.0544
0.20	0.10	50	0.0795	0.0380	0.0200	0.0130	0.0725
0.20	0.10	100	0.0850	0.0360	0.0195	0.0130	0.0528
0.30	0.00	50	0.0575	0.0275	0.0090	0.0035	0.1001
0.30	0.00	100	0.0665	0.0265	0.0145	0.0090	0.1050
0.30	0.05	50	0.0810	0.0440	0.0215	0.0090	0.0518
0.30	0.05	100	0.0875	0.0460	0.0205	0.0125	0.0453
0.30	0.10	50	0.0720	0.0365	0.0170	0.0080	0.0502
0.30	0.10	100	0.0830	0.0400	0.0160	0.0095	0.0504
0.40	0.00	50	0.0575	0.0215	0.0080	0.0025	0.1152
0.40	0.00	100	0.0610	0.0260	0.0115	0.0085	0.1057
0.40	0.05	50	0.0740	0.0275	0.0145	0.0055	0.0486
0.40	0.05	100	0.0915	0.0460	0.0230	0.0140	0.0292
†* 0.40	0.10	50	0.0611	0.0276	0.0119	0.0063	0.0671
* 0.40	0.10	50	0.0798	0.0396	0.0170	0.0081	0.0419
†* 0.40	0.10	100	0.0830	0.0391	0.0163	0.0098	0.0329
* 0.40	0.10	100	0.0879	0.0416	0.0183	0.0102	0.0313

†Indicates that the nonrandomized test is used; all other tests are randomized.
*Indicates that 9000 replications are used; otherwise, 2000 replications are used.
Note: The estimated size is the proportion of times that 0.1, 0.05, 0.25, and 0.01 exceed the test statistic τ, respectively. The Kolmogorov-Smirnov statistic tests fit of τ to a uniform random variable, where upper tail critical values at the 0.05 nominal significance level are 0.0303 and 0.0143 for 2000 and 9000 replications, respectively.

Table 2
Estimated power of proposed test against mixtures of binomial distributions, based on simulated samples

	Mixing probability	μ_1	μ_2	Number of litters	Estimated power at nominal level τ_0				Kolmogorov-Smirnov statistic
					τ_0=0.1	τ_0=0.05	τ_0=0.025	τ_0=0.01	
	0.40	0.30	0.80	50	0.0830	0.0390	0.0170	0.0075	0.0854
	0.40	0.30	0.80	100	0.1535	0.0780	0.0375	0.0245	0.1731
	0.40	0.30	0.90	50	0.1750	0.0985	0.0490	0.0340	0.1770
	0.40	0.30	0.90	100	0.3945	0.2580	0.1595	0.1155	0.4234
	0.50	0.10	0.40	50	0.0100	0.0035	0.0015	0.0005	0.2266
	0.50	0.10	0.40	100	0.0110	0.0030	0.0010	0.0000	0.2641
	0.50	0.10	0.50	50	0.0140	0.0065	0.0020	0.0005	0.2174
	0.50	0.10	0.50	100	0.0250	0.0100	0.0010	0.0005	0.1857
	0.50	0.10	0.60	50	0.0750	0.0385	0.0145	0.0095	0.0678
	0.50	0.10	0.60	100	0.1745	0.1070	0.0530	0.0320	0.2126
	0.60	0.30	0.80	50	0.0965	0.0510	0.0290	0.0175	0.0657
	0.60	0.30	0.80	100	0.2015	0.1185	0.0630	0.0465	0.2214
	0.60	0.30	0.90	50	0.2580	0.1665	0.0925	0.0625	0.2719
	0.60	0.30	0.90	100	0.5250	0.3795	0.2495	0.1915	0.5281
	0.70	0.30	0.90	50	0.2740	0.1595	0.0875	0.0615	0.3113
	0.70	0.30	0.90	100	0.5670	0.4260	0.2765	0.1995	0.5609
	0.80	0.10	0.15	50	0.0500	0.0215	0.0090	0.0045	0.1026
	0.80	0.10	0.15	100	0.0740	0.0340	0.0170	0.0075	0.0615
	0.80	0.30	0.80	50	0.2650	0.1500	0.0825	0.0545	0.3285
*	0.80	0.30	0.80	100	0.3557	0.2162	0.1206	0.0803	0.4327
	0.95	0.30	0.90	50	0.8575	0.7695	0.6205	0.5130	0.7619
	0.95	0.30	0.90	100	0.9760	0.9310	0.8395	0.7535	0.8893
	0.98	0.05	0.80	50	0.5895	0.5370	0.4570	0.3945	0.5000
	0.98	0.05	0.80	100	0.8585	0.8305	0.7830	0.7295	0.7816
	0.98	0.20	0.80	50	0.5665	0.5030	0.4170	0.3585	0.4718
	0.98	0.20	0.80	100	0.8335	0.7950	0.7175	0.6455	0.7486
	0.99	0.10	0.70	50	0.4035	0.3510	0.2875	0.2400	0.3104
	0.99	0.10	0.70	100	0.6175	0.5815	0.5245	0.4815	0.5326
	0.99	0.30	0.90	50	0.3840	0.3420	0.2920	0.2520	0.2948
	0.99	0.30	0.90	100	0.6095	0.5745	0.5300	0.4845	0.5256

*Indicates that 9000 replications are used; otherwise, 2000 replications are used.

Note: A binomial random variate is generated having parameter μ_1 with probability (mixing probability) and having parameter μ_2 with probability (1 − mixing probability). The estimated power is the proportion of times that 0.1, 0.05, 0.25, and 0.01 exceed the randomized test statistic τ, respectively. The Kolmogorov-Smirnov statistic tests fit of τ to a uniform random variable, where upper tail critical values at the 0.05 nominal significance level are 0.0303 and 0.0143 for 2000 and 9000 replications, respectively.

Table 3
Analysis of dominant lethal data

Data set number	Number of litters	$\widehat{\mu}$	$\widehat{\phi}$	$\tau_{\text{nonrandom}}$	$\overline{\tau}_{\text{random}}$	Standard error of $\overline{\tau}_{\text{random}}$
1	263	0.067936	0.064312	0.0000	0.0000	0.0000
1*	260	0.059780	0.022420	0.0939	0.0821	0.0020
2	207	0.064637	0.000000	0.0288	0.0233	0.0010
3	12	0.097226	0.254085	0.4957	0.4589	0.0014
4	123	0.091277	0.060090	0.0853	0.0612	0.0010
5	34	0.052009	0.000000	0.2694	0.2526	0.0027
6	78	0.034574	0.000000	0.2094	0.2002	0.0024
7	169	0.047676	0.041390	0.0003	0.0004	0.0001
8	201	0.051231	0.040513	0.0289	0.0148	0.0005
9	57	0.042254	0.000000	0.9733	0.9005	0.0050
10	50	0.074736	0.020870	0.2617	0.2509	0.0023
11	525	0.042917	0.003602	0.1389	0.1392	0.0024
12	303	0.026688	0.019521	0.8970	0.8426	0.0045
13	93	0.124870	0.078674	0.1490	0.0966	0.0018
14	51	0.076423	0.000000	0.0583	0.0318	0.0010
15	97	0.030290	0.024136	0.3282	0.3080	0.0026
16	36	0.334024	0.150256	0.3447	0.3151	0.0011

Note: $\tau_{\text{nonrandom}}$ is the nonrandomized goodness-of-fit test statistic, and $\overline{\tau}_{\text{random}}$ is the sample mean of 250 randomized goodness-of-fit test statistics.

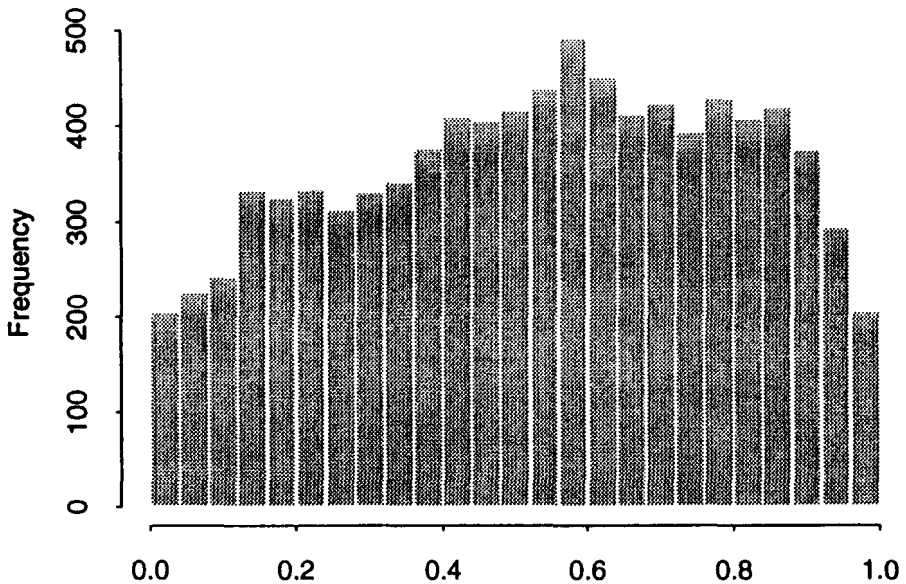

Figure 1. Nonrandomized test statistic under null model & 50 litters.

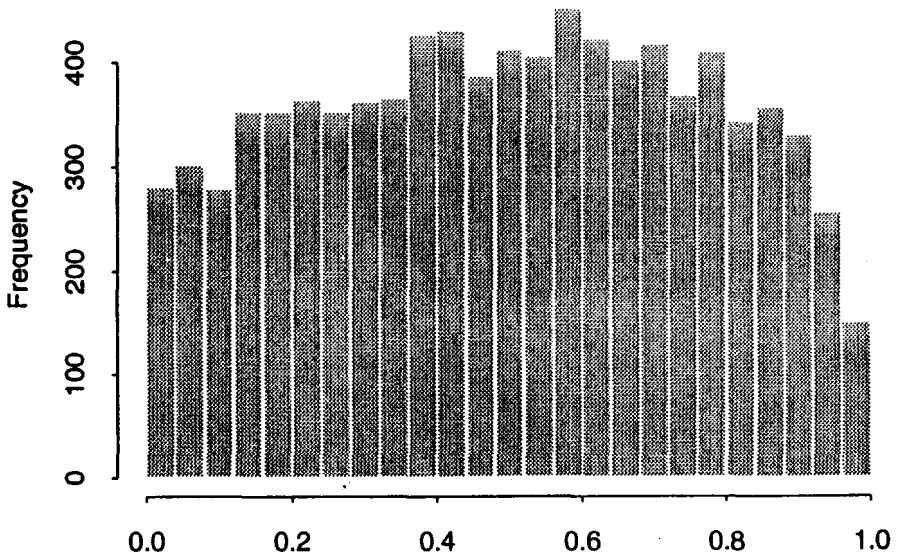

Figure 2. Randomized test statistic under null model & 50 litters.

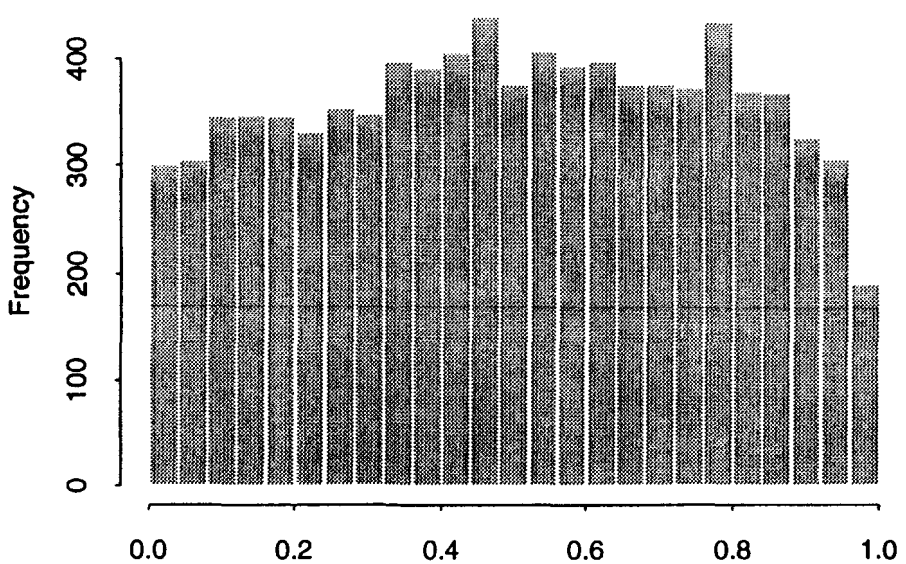

Figure 3. Randomized test statistic under null model & 100 litters.

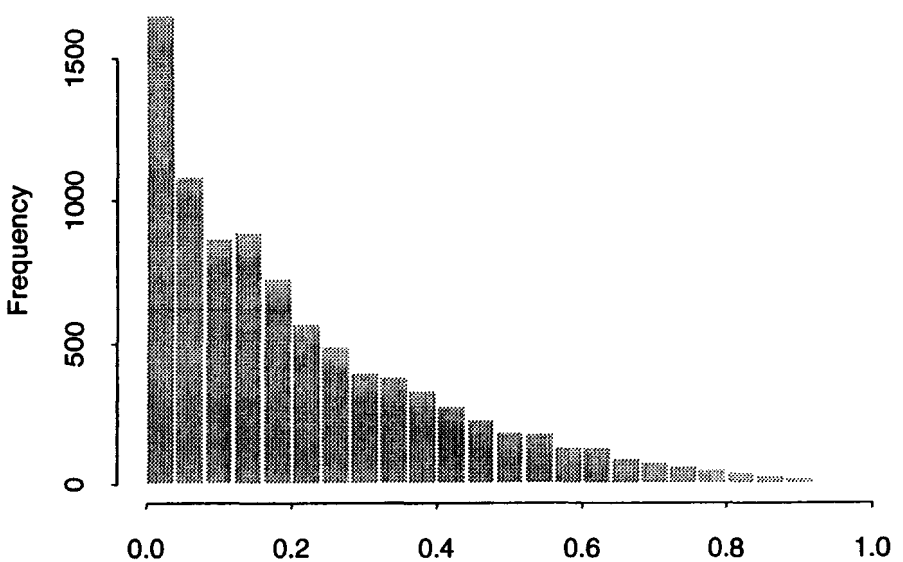

Figure 4. Randomized test statistic under alternative model & 100 litters.