Published in final edited form as:

J Agric Biol Environ Stat. 2016 December ; 21(4): 698–712. doi:10.1007/s13253-016-0261-6.

Exact-Permutation Based Sign Tests for Clustered Binary Data via Weighted and Unweighted Test Statistics

Janie McDonald¹, Patrick D. Gerard¹, Christopher S. McMahan^{1,*}, and William R. Schucany²

¹Department of Mathematical Sciences, Clemson University, Clemson, SC 29634

²Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332

Summary

Clustered binary data occur frequently in many application areas. When analyzing data of this form, ignoring key features, such as the intracluster correlation, may lead to inaccurate inference; e.g., inflated Type I error rates. For clustered binary data, Gerard and Schucany (2007) proposed an exact test for examining whether the marginal probability of a response differs from 0.5, which is the null hypothesis considered in the classic sign test. This new test maintains the specified Type I error rate and has more power, when compared to both the classic sign and permutation tests. The test statistic proposed by these authors equally weights the observed data from each cluster, regardless of whether the clusters are of equal size. To further improve the performance of the Gerard and Schucany test, a weighted test statistic is proposed and two weighting schemes are investigated. Seeking to further improve the performance of the proposed test, empirical Bayes estimates of the cluster level success probabilities are utilized. These adaptations lead to 5 new tests, each of which are shown through simulation studies to be superior to the Gerard and Schucany (2007) test. The proposed tests are further illustrated using data from a chemical repellency trial.

Keywords

Binomial; Clustered binary data; Exact test; Permutation test; Power; Sign test

1 Introduction

The analysis of dependent binary data often pose nontrivial challenges, and, further, data of this form commonly arise in a variety of application areas. For example, consider laboratory trials which are designed to assess the effectiveness of chemical insect repellents. In trials of this nature, a relatively small number of Petri dishes are used with one side of each dish being treated with the chemical in question while the other is left untreated. Insects are then placed in the center of each dish, and after a period of time, the number of insects on the treated and untreated sides are recorded. The primary goal of these trials is to determine whether the chemical in question is an effective repellent, which can be statistically evaluated by testing $H_0 : p = 0.5$ versus $H_1 : p < 0.5$, where p denotes the marginal

* mcmaha2@clemson.edu.

probability that an insect is located on the treated side. Under the assumption that the study units (i.e., insects) are independent, these hypotheses are easy to evaluate. However, in many applications this assumption naturally comes under scrutiny; e.g., many insects (e.g., termites) are known to be social, consequently it is unreasonable to assume that the insects which share a common Petri dish are independent. In order to ensure reliable and valid inference, methods for analyzing clustered binary data should be used in order to evaluate the aforementioned hypotheses; for further discussion see Gastwirth and Rubin (1971).

Many techniques for analyzing clustered binary data have been proposed, which include summary aggregation of the cluster level information, random effects regression models, and asymptotic based hypothesis testing procedures. For example, Larocque, Nevalainen, and Oja (2007) proposed a weighted multivariate sign test for clustered data, which extended the earlier work of Larocque (2003). This proposal made by these authors, like many other existing techniques, is an asymptotic test and is appropriate for analyzing data arising from studies which consider a moderate to large number of clusters. Even though asymptotic tests perform well, these procedures often suffer from inflated Type I error rates when used to analyze a relatively small number of clusters (e.g., 5–10). For further discussion see Follmann and Proschan (1999). In such small sample venues, exact and permutation based hypothesis testing procedures applied to the summary aggregation of the cluster level information are natural alternatives which maintain specified Type I error rates, but may suffer from severely reduced power.

Gerard and Schucany (2007) proposed a test, hereafter referred to as the EPB-test, which is appropriate for analyzing clustered binary data even when considering a small number of clusters; i.e., the EPB-test maintains specified Type I error rates. Moreover, the EPB-test was shown to be uniformly superior to the classic sign test based on aggregated cluster level data and as, if not more, powerful than the standard permutation test. The gains in power over these traditional techniques are attributable to the fact that the EPB-test utilizes both permutation arguments and an exact parametric bootstrap to calculate p -values. Further, unlike standard permutation tests which advocate permuting the observed units (e.g., see Follmann and Proschan, 1999) the EPB-test permutes estimated random effects. In doing so, the EPB-test is able to exploit the binomial nature of the observed data.

Motivated by the work of Gerard and Schucany (2007), herein adaptations of the EPB-test are proposed in an effort to improve its performance. In particular, a weighted test statistic is proposed and two weighting schemes are suggested: (1) weights based on the cluster sizes and (2) weights that minimize the variance of the test statistic. Further, the use of empirical Bayes estimates of the cluster level success probabilities is investigated. The combinations of these proposed adaptations lead to 5 new hypothesis tests that are appropriate for clustered binary data, maintain their specified Type I error rate, and are, in general, more powerful than the EPB-test.

Subsequent sections of this article are organized as follows. In Section 2, the salient features of the EPB-test are presented. Section 3 provides the proposed modifications to the EPB-test. In Section 4, the performance of the proposed methodology is evaluated through a

simulation study, and Section 5 provides the motivating application. Section 6 concludes with a summary discussion.

2 EPB-test

Motivated by a chemical termite repellency trial, Gerard and Schucany (2007) developed the EPB-test for evaluating $H_0 : p = 0.5$ versus $H_1 : p < 0.5$, where p is the marginal probability that a termite is on the treated side of a Petri dish. In this setting, the former hypothesis indicates that the chemical appears to be ineffective at repelling termites, while the latter suggests that the chemical is effective. In order to statistically frame this problem, let the binary random variable X_{ij} denote the location of the j th termite within the i th Petri dish, for $j = 1, \dots, n_i$ and $i = 1, \dots, m$, where $X_{ij} = 1$ denotes that the termite is on the treated side, and $X_{ij} = 0$ otherwise. Consequently, a small number of total “successes” (i.e., a relatively small value of $T = \sum_{i=1}^m \sum_{j=1}^{n_i} X_{ij}$) provides evidence that the chemical is effective at repelling termites (i.e., H_1).

To acknowledge the potential dependence within clusters (i.e., Petri dishes), it is assumed that the study units within clusters are dependent ($X_{ij} \not\perp X_{ij'}, \forall j \neq j'$), but the units across clusters are independent ($X_{ij} \perp X_{i'j'}, \forall j \text{ and } j', \text{ where } i \neq i'$), where $\perp$ denotes statistical independence. To account for the within cluster dependence, a random effects model is adopted. Thus, given the random effects, the X_{ij} are (conditionally) independent and they each follow a Bernoulli distribution; i.e.,

$$X_{ij}|p_i \sim \text{Bernoulli}(p_i) \text{ and } \log\left(\frac{p_i}{1-p_i}\right) = \theta + \beta_i, \quad (1)$$

where β_i is a random effect associated with the i th cluster, and θ is the log-odds parameter associated with the marginal probability of success. Further, it is assumed that the β_i follow a symmetric probability distribution with mean 0 and variance σ_{β}^2 . Consequently, the X_{ij} across clusters have different success probabilities, and unconditionally, the X_{ij} within the same cluster would be dependent. Under the mixed effects model, testing $H_0 : p = 0.5$ versus $H_1 : p < 0.5$ is equivalent to testing

$$H_0: \theta = 0 \text{ versus } H_1: \theta < 0, \quad (2)$$

as it pertains to the log-odds.

In order to evaluate the hypotheses in (2), the EPB-test uses both permutation arguments and an exact parametric bootstrap to calculate p -values, through the process described herein. Under the null hypothesis and conditional on the random effects, the X_{ij} are independent Bernoulli random variables with success probability $p_i = \exp(\beta_i) / \{1 + \exp(\beta_i)\}$, for $j = 1, \dots, n_i$ and $i = 1, \dots, m$. Consequently, $T = \sum_{i=1}^m \sum_{j=1}^{n_i} X_{ij}$ can be viewed as the sum of independent, but non-identically distributed, binomial random variables; i.e.,

$\sum_{j=1}^{n_i} X_{ij} \sim \text{binomial}(n_i, p_i)$. Further, considering the assumptions regarding the distribution of the cluster level random effects, it follows that β_i and $-\beta_i$ are exchangeable. Therefore, permuting the signs of the β_i 's yields $N = 2^m$ configurations of the random effects. Denote the collection of all such permutations as

$$\mathcal{B} = \{\beta_k : \beta_k = (\beta_1^k, \dots, \beta_m^k), \text{ where } \beta_i^k = \beta_i \text{ or } \beta_i^k = -\beta_i\}.$$

For each $\beta_k \in \mathcal{B}$, define $T_k = \sum_{i=1}^m \sum_{j=1}^{n_i} X_{ij}^k$, where $\sum_{j=1}^{n_i} X_{ij}^k \stackrel{\text{ind.}}{\sim} \text{binomial}(n_i, p_i^k)$ and $p_i^k = \exp(\beta_i^k) / \{1 + \exp(\beta_i^k)\}$, for $i = 1, \dots, m$. Using the recursive algorithm proposed by Thomas and Taub (1982), one can easily compute probabilities associated with the random variable T_k , for $k = 1, \dots, N$. Consequently, motivated by the principles of the standard permutation test, one can evaluate the hypotheses in (2) by determining how unusual the observed number of “successes” (i.e., $T = \sum_{i=1}^m \sum_{j=1}^{n_i} X_{ij} = t$) are when compared to T_k , for $k = 1, \dots, N$. More specifically, a p -value for testing (2) can be constructed as the $P(T \leq t)$ over all N of the equally likely combinations of the random effects provided in $\mathcal{B}$; i.e.,

$$P(T \leq t) = \sum_{k=1}^N P(T_k \leq t | \beta_k) P(\beta_k),$$

where $P(\beta_k) = \frac{1}{N}$ and t is the observed number of successes. Proceeding in this fashion retains the general principles of the permutation test, but utilizes the binomial nature of the data which the permutation test ignores. Motivated by the data application, the current presentation focuses solely on a lower-tail test, but minor adaptations can allow for the calculation of the p -value associated with an upper-tail or two-tailed test.

In practice, the β_i are never known, but they can be estimated/predicted based on the proportion of successes within each cluster. Under the null hypothesis, the model described in (1) reduces to $X_{ij}/p_i \sim \text{Bernoulli}(p_i)$, where $\log\{p_i/(1-p_i)\} = \beta_i$. Thus, a natural estimator of β_i is given by $\hat{\beta}_i = \log\{\hat{p}_i/(1-\hat{p}_i)\}$, where $\hat{p}_i = n^{-1} \sum_{j=1}^{n_i} X_{ij}$. Note, $\hat{\beta}_i = -\infty(\infty)$ when $\hat{p}_i = 0(1)$, but this is of no consequence in the implementation of the EPB-test. Subsequently, define

$$\hat{\mathcal{B}} = \{\hat{\beta}_k : \hat{\beta}_k = (\hat{\beta}_1^k, \dots, \hat{\beta}_m^k), \text{ where } \hat{\beta}_i^k = \hat{\beta}_i \text{ or } \hat{\beta}_i^k = -\hat{\beta}_i\}.$$

to be the collection of the N configurations of the estimated random effects obtained from permuting the signs of the $\hat{\beta}_i$'s. Consequently, the p -value for testing the set of hypotheses provided in (2) can be calculated as

$$P(T \leq t) = \sum_{k=1}^N P(T_k \leq t | \hat{\beta}_k) P(\hat{\beta}_k),$$

where $P(\hat{\beta}_k) = 1/N$ and the probabilities associated with T_k are calculated, using

$\hat{p}_i^k = \exp(\hat{\beta}_i^k) / \{1 + \exp(\hat{\beta}_i^k)\}$ rather than p_i^k , as was previously described. Note, if $\hat{\beta}_i^k = -\infty(\infty)$ then $\hat{p}_i^k = 0(1)$.

It is worthwhile to note that for the purpose of mathematical tractability the EPB-test was developed for the set of hypotheses provided in (2). Through simulation studies (results not shown) it appears that the EPB-test, as well as the adaptations proposed herein, would be appropriate for testing the more general set of hypotheses given by $H_0 : p \geq 0.5$ versus $H_1 : p < 0.5$. More specifically, these additional studies indicate that the proposed tests maintain the specified Type I error rate for all values of p under the more general null hypothesis.

However, formally establishing this could be mathematically challenging and is beyond the scope of this work, but could be a topic of future research. Thus, caution should be exercised if one desires to interpret the results of the test under the more general null hypothesis.

3 Proposed adaptations

In this section, adaptations to the methodology presented in Gerard and Schucany (2007) are proposed in order to develop hypothesis testing procedures which maintain their Type I error rate and are more powerful than the EPB-test. In particular, a weighted test statistic is proposed and the use of empirical Bayes estimates for the cluster level success probabilities is investigated.

3.1 Weighted test statistic

In many studies, such as the chemical replency trial considered by Gerard and Schucany (2007), the number of study units within each cluster may vary. To exploit the cluster sizes, with respect to developing a more powerful test, a weighted test statistic can be used. In particular, define the following test statistic

$$T^w = \sum_{i=1}^m w_i \sum_{j=1}^{n_i} X_{ij},$$

where w_i is the weight assigned to each cluster. Motivated by the data application, the specification of the weights in T^w will be chosen to exploit the heterogeneity of the cluster sizes, but other weighting schemes could be adopted. Using the weighted test statistic, the p -value for testing the set of hypotheses provided in (2) can be calculated as

$$P(T^w \leq t^w) = \sum_{k=1}^N P(T_k^w \leq t^w | \hat{\beta}_k) P(\hat{\beta}_k), \quad (3)$$

where t^w is the observed realization of T^w , $T_k^w = \sum_{i=1}^m w_i \sum_{j=1}^{n_i} X_{ij}^k$, and $\sum_{j=1}^{n_i} X_{ij}^k \stackrel{\text{ind.}}{\sim} \text{binomial}(n_i, \hat{p}_i^k)$, with $\hat{p}_i^k$ again being defined as in Section 2. That is, in this context one can view T^w and T_k^w as the weighted sum of independent but non-identically distributed binomial random variables. Consequently, the recursive algorithm developed by Thomas and Taub (1982) is not applicable for calculating probabilities associated with these random variables, and standard recursive techniques are computationally infeasible. To obviate this difficulty, a lookup table algorithm was developed to calculate the necessary probabilities in (3); for further details see the Appendix. Note, a key feature of this algorithm requires that the weights (i.e., the w_i) be integer, but after rounding and rescaling this poses no limitations to the proposed approach. Using this new algorithm, once the weights are specified the approach to calculating the p -value for testing the set of hypotheses provided in (2) is analogous to the approach described in Section 2.

In what follows, two weighting schemes that utilize the varying cluster sizes are proposed. Directly acknowledging the cluster sizes, the first scheme assigns weights so that each cluster's contribution to T^w is equal. In particular, this weighting scheme specifies $w_i = (mn_i)^{-1}$. The testing procedure which makes use of these weights is referred to as the WB-test. The second weighting scheme specifies the w_i to minimize the variance of the test statistic. Following the work of Ahn, Hu, and Schucany (2011), it can be shown that the variance of T^w is given by

$$\sigma_0^2(T^w) = p(1-p) \sum_{i=1}^m w_i^2 n_i \{1 + (n_i - 1)\rho\},$$

where p is the marginal success probability and ρ is the common intracluster correlation coefficient; i.e., $\rho = \text{corr}(X_{ij}, X_{ij'})$ for $j \neq j'$. Under the constraints that the $w_i \geq 0$ and $m^{-1} \sum_{i=1}^m w_i n_i = 1$, it can be shown that the set of weights that minimize the variance of T^w are given by

$$w_i = \frac{m \{1 + (n_i - 1)\rho\}^{-1}}{\sum_{i=1}^m n_i \{1 + (n_i - 1)\rho\}^{-1}},$$

for $i = 1, \dots, m$. The testing procedure which makes use of these weights is referred to as the MV-test.

Note, all of the information required to implement the former weighting scheme is readily available in the observed data, while the latter depends on an unknown parameter. In order to implement the MV-test an estimate of ρ has to be obtained. Following the recommendations of Ridout, Demetrio, and Firth (1999), an estimator of ρ can be obtained as

$$\hat{\rho}_{AOV} = \frac{MS_b - MS_w}{MS_b + (n_0 - 1)MS_w},$$

where $n_0 = (m-1)^{-1} \left\{ \sum_{i=1}^m n_i - \left(\sum_{i=1}^m n_i \right)^{-1} \sum_{i=1}^m n_i^2 \right\}$ and $MS_b (MS_w)$ is the between-group (within-group) mean squares from a one way analysis of variance of the X_{ij} ; i.e.,

$$MS_b = \frac{1}{m-1} \left\{ \sum_{i=1}^m n_i^{-1} S_i^2 - \left(\sum_{i=1}^m n_i \right)^{-1} \left(\sum_{i=1}^m S_i \right)^2 \right\}$$

and

$$MS_w = \frac{1}{\sum_{i=1}^m n_i - m} \left(\sum_{i=1}^m S_i - \sum_{i=1}^m n_i^{-1} S_i^2 \right),$$

where $S_i = \sum_{j=1}^{n_i} X_{ij}$ are the observed number of successes within the i th cluster. Note, the estimator $\hat{\rho}_{AOV}$ has the propensity to both exceed 1 and fall below 0. In the event that $\hat{\rho}_{AOV} > 1$, the estimator of ρ should be specified to be $\hat{\rho} = 1$. Likewise, in the context of the motivating data application, values of $\rho < 0$ are nonsensical. In the event that $\hat{\rho}_{AOV} < 0$, the estimator of ρ should be specified to be $\hat{\rho} = 0$. Further, the estimator $\hat{\rho}_{AOV}$ may be undefined if the number of successes (or failures) equals the cluster size in all clusters, in which case the estimator of ρ should be specified to be $\hat{\rho} = 1$. Note, other techniques for estimating ρ are available and can be directly adopted for the purpose of identifying the weights in the MV-test; for further discussion see Ridout, Demetrio, and Firth (1999) and the references therein. It is worthwhile to point out that when the clusters are of equal size (i.e., $n_i = n$ for all i) the proposed test statistic (3) under each aforementioned weighting scheme is equivalent to the test statistic proposed by Gerard and Schucany (2007).

3.2 Proportion estimates

Gerard and Schucany (2007) noted that the EPB-test seemed to be conservative, but this feature was mitigated when the true random effects were used in the p -value calculations. In practice, this approach is infeasible, but this finding tends to suggest that a more reliable test can be developed through adopting a more precise estimator of the cluster level random effects, or equivalently the p_i . To facilitate this goal, an empirical Bayes estimator of the cluster level success probabilities is proposed as an alternative to the estimator $\hat{p}_i$ utilized in the original EPB-test. Following the development given in Tamura and Young (1987) and Carlin and Louis (2000), it is assumed that $p_i \sim \text{beta}\{M\mu, M(1-\mu)\}$. Hence, the posterior distribution of p_i , given the observed number of successes in the i th cluster, is beta

$\left\{ M\mu + \sum_{j=1}^{n_i} X_{ij}, M(1-\mu) + n_i - \sum_{j=1}^{n_i} X_{ij} \right\}$. An empirical Bayes estimator of p_i can subsequently be constructed as

$$\tilde{p}_i = \frac{\hat{M}}{\hat{M} + n_i} \hat{\mu} + \frac{n_i}{\hat{M} + n_i} \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i},$$

where $\hat{M}$ and $\hat{\mu}$ are method of moments estimates obtained from available data. When the n_i are not equal, Tamura and Young (1987) suggest the following estimators of M and μ

$$\hat{\mu} = \frac{\sum_{i=1}^m n_i \hat{p}_i}{\sum_{i=1}^m n_i} \quad \text{and} \quad \hat{M} = \frac{\hat{\mu}(1 - \hat{\mu})(\sum_{i=1}^m n_i z_i - \sum_{i=1}^m z_i)}{\sum_{i=1}^m n_i (\hat{p}_i - \hat{\mu})^2 - \hat{\mu}(1 - \hat{\mu})(\sum_{i=1}^m z_i)} - 1,$$

where $z_i = 1 - \frac{n_i}{\sum_{i=1}^m n_i}$. Alternatively, in the case in which the n_i are equal (i.e., $n_i = n$, for $i = 1, \dots, m$) the following estimators are suggested

$$\hat{\mu} = \frac{\sum_{i=1}^m \hat{p}_i}{m} \quad \text{and} \quad \hat{M} = \frac{\hat{\mu}(1 - \hat{\mu}) - (m - 1)^{-1} \sum_{i=1}^m (\hat{p}_i - \hat{\mu})^2}{(m - 1)^{-1} \sum_{i=1}^m (\hat{p}_i - \hat{\mu})^2 - \hat{\mu}(1 - \hat{\mu})n^{-1}}.$$

This adaptation can be used to calculate a p -value for testing (2) in the exact same fashion as was described in Sections 2 and 3.1 after replacing $\hat{p}_i$ by $\tilde{p}_i$ for $i = 1, \dots, m$. One may note that $\tilde{p}_i$ is a “shrinkage” estimator of p_i and the testing procedures that make use of these estimates are therefore aptly referred to as the EPBS-, WBS-, and MVS-tests, where the p -values for these tests are calculated in the same fashion as was described for the EPB-, WB-, and MV-tests, respectively, after replacing $\hat{p}_i$ by $\tilde{p}_i$.

4 Simulation study

In order to evaluate the finite sample performance of the proposed adaptations to the EPB-test, the following simulation study was conducted. This study considers testing the null hypothesis $H_0 : \theta = 0$ versus the one-sided alternative $H_1 : \theta < 0$, or equivalently $H_0 : p = 0.5$ versus $H_1 : p < 0.5$. Correlated binary data were generated using the algorithm developed in Lunn and Davies (1998). In particular, this study considers values of the true marginal probability of success $p \in \{0.5, 0.3, 0.2, 0.1\}$, where $p = 0.5$ corresponds to H_0 , while the other values provide settings under which H_1 is true. For each of the considered marginal probabilities, 5 different clustering coefficients (ϕ) were considered, where $\phi \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The clustering coefficient controls the strength of the intracluster correlation (i.e., $\phi = \sqrt{\rho}$) and these specifications lead to correlations ranging from 0.01 to 0.81.

In order to compare the performance of the EPB-test with the testing procedures that make use of the proposed adaptations, it is first noted that the weighted tests (i.e., the WB- and MV-test) are equivalent to the EPB-test for equally sized clusters. Consequently, the study described considers clusters of unequal size. In particular, the cluster sizes were generated as 3 plus a negative binomial random variable having mean parameter μ and dispersion parameter r , i.e., $n_i \sim 3 + \text{nb}(\mu, r)$, for $i = 1, \dots, m$. At each considered setting of (p, ϕ, m, μ, r) 1000 independent data sets were generated and for each the EPB-, EPBS-, WB-, WBS-, MV-, and MVS-tests were implemented. For the purposes of this study, four configurations of (m, μ, r) were considered; i.e., $(m, \mu, r) = (7, 7, 40)$, $(m, \mu, r) = (6, 8, 4)$, $(m, \mu, r) = (5, 20, 40)$, and $(m, \mu, r) = (5, 10, 2)$ with the corresponding results being provided in Tables 1–

4, respectively. Each of these tables provides the percentage of the times for which H_0 was rejected at the $\alpha = 0.05$ level of significance, for the different combinations of (p, ϕ) .

Recall, the lookup table algorithm developed for calculating probabilities associated with the sum of weighted binomial random variables requires that the weights be integer. Thus, the weights based on the cluster sizes were adjusted to be $w_i^* = \lfloor 1000w_i \rfloor + 1$, where $\lfloor \cdot \rfloor$ is the nearest integer function, and the weights which minimize the variance of the test statistic were adjusted to be $w_i^* = \lfloor 100w_i \rfloor + 1$. Note, rescaling the weights in (3) in this fashion still retains the principles on which the different schemes were developed, but allows for the use of the lookup table algorithm. Further, to implement the MV- and MVS-tests, ρ was estimated using the technique described in Section 3.1. In order to assess the impact of having to estimate this parameter, the MV- and MVS-tests were also implemented using the true intraclass correlation value and these tests are referred to as the MV^* - and MVS^* -tests, respectively.

From the results presented in Tables 1–4, it can be inferred that the proposed testing procedures are appropriate for evaluating the hypotheses in (2) in the presence of clustered binary data; i.e., all of the testing procedures maintained their specified Type I error rate. Further, the proposed testing procedures performed almost uniformly better than the EPB-test, both with respect to power and being less conservative. In particular, the use of the empirical Bayes estimates to replace $\hat{p}_i$ provided for a significant gain in power, especially for slight departures from the null (i.e., $p = 0.3$) under moderate levels of the intraclass correlation (e.g., $\phi = 0.5$). This adaptation also provided for a less conservative testing procedure, in general. Alternatively, when considering the use of a weighted test statistic (i.e., the WB- and MV-tests) less dramatic gains in power were realized when compared to the original EPB-test, and these tests still appeared to retain the conservative nature of the EPB-test. Overall, the performance of the EPB-test was enhanced the most by adopting both the use of empirical Bayes estimates of the cluster level success probabilities and the weighted test statistic; i.e., the WBS- and MVS-tests would generally be preferred to the other considered testing procedures. Further, the WBS-test tended to perform as well if not better than the MVS-test. Recall, in order to implement the MV- and MVS-tests an estimate of ρ had to be obtained, and having to estimate this additional parameter could explain why these tests did not perform as well as the WB- and WBS-tests, respectively. Interestingly, the versions of these two tests which make use of the true intraclass correlation value (i.e., the MV^* - and MVS^* -tests) tended to perform worse than the versions that were left to estimate ρ . This finding tends to suggest that the MV- and MVS-tests are relatively sensitive to the specification of the value of ρ . Consequently, given this and the findings presented above it is generally recommended that the WBS-test be adopted in practice.

5 Motivating example

To further illustrate the advantages of the proposed adaptations to the EPB-test, data from a termite repellency trial was considered. This study investigated the potential use of catnip oil as a termite repellent, and the investigators considered several dilutions of the oil; i.e., a control level (no catnip oil) and three increasing dilution levels. For each level, a series of Petri dishes were treated, with the catnip oil dilution being applied to one side of the dish,

with the other being left untreated. Ten termites were then placed in the middle of each of the treated Petri dishes, and after a six hour period of time, the number of termites on the treated and untreated sides were recorded, but the location of the individual termites within the dish were not recorded; these data are provided in Table 5. Note, during the course of the six hour period several of the termites expired, resulting in the unequal sample sizes within each petri dish reported in Table 5. Further, the cause of death during the course of the study was not thought to be treatment-related, but rather was due to the tendency of the termites to cannibalize one another. From the data reported in Table 5, one can see that the proportion of termites on the treated side of the Petri dishes for the control and the three increasing dilution levels are 0.32, 0.25, 0.214, and 0.315, respectively, providing moderate evidence that catnip oil might be an effective termite repellent. To formally assess the effectiveness of catnip oil as a termite repellent, the EPB-, EPBS-, WB-, WBS-, MV-, and MVS-tests were implemented in the same fashion as was described in Section 4.

Table 5 presents the results of this analysis. In particular, at each dilution level a p-value for each of the six hypothesis testing procedures is reported. Before interpreting these results, it is noted that catnip oil has, through other more recent studies, been shown to be an effective termite repellent; e.g., see Verma et al. (2009) and the references therein. From these results, one will first note that for the control level all six hypothesis testing procedures are in agreement and result in p-value which provides little evidence of the alternative hypothesis; i.e., the control level exhibits no evidence of being an effective termite repellent, which intuitively suggests that all six hypothesis testing procedures came to the correct conclusion. For the results pertaining to the three catnip oil dilution levels, all of the testing procedures came to concordant results, with the five new hypothesis testing procedures generally obtaining p-values which were smaller than the p-values obtained by the EPB-test; e.g., for Dilution 1, the EPB-test obtained a p-value of 0.051, where as the EPBS-, WB-, WBS-, MV-, and MVS-tests resulted in p-values of 0.049, 0.030, 0.029, 0.032, and 0.030, respectively. Assuming that catnip oil is a useful repellent, these findings reinforce the main findings presented in Section 4, that is, the proposed adaptations of the EPB-test have more power to detect departures from the null hypothesis than the original EPB-test.

6 Discussion

Adaptations to the EPB-test have been proposed that led to the development of five new hypothesis testing procedures, which are appropriate for analyzing clustered binary data. The considered adaptations include using empirical Bayes estimates of the cluster level success probabilities and a weighted test statistic, with the weights being chosen either based on cluster size or to minimize the variance of the test statistic. These new tests were evaluated through simulation and were found to generally perform better than the EPB-test, across all considered simulation configurations. The proposed testing procedures were further illustrated using data from a termite repellency trial.

From the findings of this study the testing procedure which makes use of both the empirical Bayes estimates of the cluster level success probabilities and the weighted test statistic, with weights being chosen based on cluster sizes, (i.e., the WBS-test) appears to be preferable when compared to the other considered procedures. Thus from an application point of view,

it is suggested that this test be adopted. Further, even though the testing procedures proposed herein would be appropriate for analyzing data consisting of a large number of clusters, it is generally advisable, due to ease of implementation, to use asymptotic tests such as the one proposed in Larocque, Nevalainen, and Oja (2007) in these scenarios.

The aforementioned adaptations of the test originally proposed by Gerard and Schucany (2007) provide for better power while maintaining the specified Type I error rate. Given these improvements, one could consider several other extensions which could be used to address issues commonly encountered in the analysis of clustered binary data. Motivated by the improvements observed in the MV- and MVS-tests one could develop weighting schemes which more efficiently use information about the intracluster correlation structure. For example, in our motivating data application if information about whether termites within a Petri dish grouped together or acted independently was collected, this additional data could be used to better evaluate the within-group correlation on a cluster-by-cluster basis, thus potentially leading to a more efficient weighting scheme. A primary challenge with developing this approach would be associated with developing reliable small sample estimators of the within cluster correlation structure, a feature which the MV- and MVS-tests are not as susceptible to since the intracluster correlation is estimated across all clusters.

Other generalizations could include adapting the EPB-test to allow for comparisons between different treatment levels and/or to account for other confounding variables, but these adaptations do not appear to be straightforward. Further, it is common in repellency trials for study units to be examined over time, thus adaptations to the EPB-test that accommodate this data structure would be useful. These generalizations could follow the work of Ridout et al. (1998) which made use of Markov models to investigate β -fraction, a by-product of hop processing, as a repellent of the two-spotted spider mite.

Appendix

In order to compute the p-values for the WB-, WBS-, MV-, and MVS-tests, one must compute the probability associated with a weighted sum of independent, but non-identically distributed random variables; i.e., one must be able to calculate probabilities of the form $P(T \leq t)$, where $T = \sum_{i=1}^m w_i Y_i$ and $Y_i \stackrel{\text{ind.}}{\sim} \text{binomial}(n_i, p_i)$. When the weights are all unity or are equal (i.e., $w_i = w_j$ for all i, j), the recursive algorithm presented in Thomas and Taub (1982) can be used to calculate such probabilities. When the weights are not equal, the algorithm presented in Thomas and Taub (1982) is not applicable. To circumvent this issue, a lookup table algorithm was developed for the purposes of computing these probabilities for integer weights. In what follows, an R program is provided that can be used to calculate $P(T \leq t)$. The arguments of the function and its output are:

Arguments:

- t : Value at which the cumulative distribution function of T should be evaluated.
- p : Vector of length m consisting of the p_i ; i.e., $p = (p_1, \dots, p_m)'$.
- w : Vector of length m consisting of the w_i ; i.e., $w = (w_1, \dots, w_m)'$.

- N : Vector of length m consisting of the n_j ; i.e., $N = (n_1, \dots, n_m)'$.

Returns:

- The value of $P(T \leq t)$.

R Function:

```
pweighted.sum.binom<-function(t,p,w,N){
  m<-length(N)
  Mnj<-matrix(0, nrow=m, ncol=(t+1))
  sup<-floor((0:t)/w[m])
  Mnj[m,1:(t+1)]<-pbinom(sup,N[m],p[m])
  for(i in (m-1):1){
    for(r in 0:t){
      sup<-unique(floor((0:r)/w[i]))
      pull<-r-w[i]*sup
      pull<-pmax(pull,0)+1
      Mnj[i,r+1]<- sum(dbinom(sup,N[i],p[i])*Mnj[i+1,pull])
    }
  }
  return(Mnj[1,t+1])
}
```

Acknowledgments

The authors wish to thank the Associate Editor and two anonymous referees for their insightful comments on earlier versions of this manuscript. Implementing their suggestions significantly improved the manuscript. This research was partially supported by Grant R01 AI121351 from the National Institutes of Health.

References

- Ahn C, Hu F, Schucany W. Sample size calculation for clustered binary data with sign tests using different weighting schemes. *Statistics in Biopharmaceutical Research*. 2011; 3:65–72. [PubMed: 21339864]
- Carlin, B., Louis, T. Bayes and empirical Bayes methods for data analysis. 2nd. Chapman and Hall; 2000.
- Follmann D, Proschan M. Valid inference in random effects meta-analysis. *Biometrics*. 1999; 55:732–737. [PubMed: 11315000]
- Gerard P, Schucany W. An enhanced sign test for dependent binary data with small numbers of clusters. *Computational Statistics and Data Analysis*. 2007; 51:4622–4632.
- Gastwirth J, Rubin H. Effect of dependence on the level of some one-sample tests. *Journal of the American Statistical Association*. 1971; 66:816–820.
- Larocque D. An affine-invariant multivariate sign test for cluster correlated data. *Canadian Journal of Statistics*. 2003; 31:437–455.
- Larocque D, Nevalainen J, Oja H. A weighted multivariate sign test for cluster-correlated data. *Biometrika*. 2007; 94:267–283.
- Lunn A, Davies S. A note on generating correlated binary variables. *Biometrika*. 1998; 85:487–490.
- Ridout M, Demetrio C, Firth D. Estimating intraclass correlation for binary data. *Biometrics*. 1999; 55:137–148. [PubMed: 11318148]

- Ridout M, Faddy M, Jones G, Campbell C. Markov models of mite movements. *Journal of the Royal Statistical Society: Series C*. 1998; 47:265–277.
- Tamura R, Young S. A stabilized moment estimator for the beta-binomial distribution. *Biometrics*. 1987; 43:813–824. [PubMed: 3427166]
- Thomas M, Taub A. Calculating binomial probabilities when the trial probabilities are unequal. *Journal of Statistical Computation and Simulation*. 1982; 14:125–131.
- Verma M, Sharma S, Prasad R. Biological alternatives for termite control: A review. *International Biodeterioration and Biodegradation*. 2009; 63:959–972.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Percentage of 1000 replications for which $H_0 : p = 0.5$ was rejected at the $\alpha = 0.05$ level of significance, when $(m, \mu, r) = (7, 7, 40)$.

p	ϕ	EPB	EPBS	WB	WBS	MV	MVS	MV*	MVS*
0.1	0.1	0.6	0.7	0.6	1.0	0.6	0.7	0.3	1.4
	0.3	1.4	3.5	1.4	4.0	1.3	3.6	0.7	4.2
	0.5	2.7	4.3	2.4	4.2	2.4	4.3	2.7	4.5
0.3	0.1	0.6	0.7	0.6	1.0	0.6	0.7	0.3	1.4
	0.3	1.4	3.5	1.4	4.0	1.3	3.6	0.7	4.2
	0.5	2.7	4.3	2.4	4.2	2.4	4.3	2.7	4.5
0.5	0.1	0.6	0.7	0.6	1.0	0.6	0.7	0.3	1.4
	0.3	1.4	3.5	1.4	4.0	1.3	3.6	0.7	4.2
	0.5	2.7	4.3	2.4	4.2	2.4	4.3	2.7	4.5
0.7	0.1	0.6	0.7	0.6	1.0	0.6	0.7	0.3	1.4
	0.3	1.4	3.5	1.4	4.0	1.3	3.6	0.7	4.2
	0.5	2.7	4.3	2.4	4.2	2.4	4.3	2.7	4.5
0.9	0.1	0.6	0.7	0.6	1.0	0.6	0.7	0.3	1.4
	0.3	1.4	3.5	1.4	4.0	1.3	3.6	0.7	4.2
	0.5	2.7	4.3	2.4	4.2	2.4	4.3	2.7	4.5
0.1	0.1	68.1	75.9	66.8	74.9	67.6	75.4	55.1	61.7
	0.3	52.7	67.3	51.6	68.2	52.0	68.0	46.1	56.6
	0.5	39.3	48.8	40.2	51.0	40.9	50.9	32.9	42.7
0.3	0.1	68.1	75.9	66.8	74.9	67.6	75.4	55.1	61.7
	0.3	52.7	67.3	51.6	68.2	52.0	68.0	46.1	56.6
	0.5	39.3	48.8	40.2	51.0	40.9	50.9	32.9	42.7
0.5	0.1	68.1	75.9	66.8	74.9	67.6	75.4	55.1	61.7
	0.3	52.7	67.3	51.6	68.2	52.0	68.0	46.1	56.6
	0.5	39.3	48.8	40.2	51.0	40.9	50.9	32.9	42.7
0.7	0.1	68.1	75.9	66.8	74.9	67.6	75.4	55.1	61.7
	0.3	52.7	67.3	51.6	68.2	52.0	68.0	46.1	56.6
	0.5	39.3	48.8	40.2	51.0	40.9	50.9	32.9	42.7
0.9	0.1	68.1	75.9	66.8	74.9	67.6	75.4	55.1	61.7
	0.3	52.7	67.3	51.6	68.2	52.0	68.0	46.1	56.6
	0.5	39.3	48.8	40.2	51.0	40.9	50.9	32.9	42.7
0.1	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.2	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.3	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.4	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.5	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.6	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.7	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.8	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.9	0.1	98.3	97.8	98.1	98.7	98.3	97.6	94.6	91.7
	0.3	87.9	93.9	87.8	93.2	88.1	94.1	81.4	86.4
	0.5	73.3	79.8	75.9	81.9	75.3	81.8	66.1	73.7
0.1	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4
0.2	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4
0.3	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4
0.4	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4
0.5	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4
0.6	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4
0.7	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4
0.8	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4
0.9	0.1	100.0	100.0	99.8	100.0	100.0	100.0	99.8	99.6
	0.3	99.7	99.9	99.8	99.8	99.8	99.9	99.1	99.6
	0.5	95.9	97.2	95.5	97.2	95.6	96.9	90.8	93.4

Table 2

Percentage of 1000 replications for which $H_0 : \rho = 0.5$ was rejected at the $\alpha = 0.05$ level of significance, when $(m, \mu, r) = (6, 8, 4)$.

ρ	ϕ	EPB	EPBS	WB	WBS	MV	MVS	MV*	MVS*
0.1	0.1	0.3	0.7	0.5	2.3	0.3	0.7	0.0	2.0
	0.3	0.6	1.6	0.8	2.7	0.6	2.1	0.4	1.9
	0.5	2.1	3.4	2.4	4.4	2.3	4.4	2.1	3.5
0.5	0.7	3.3	4.7	3.4	5.1	3.8	4.7	3.7	5.0
	0.9	3.5	3.6	2.9	3.4	3.0	3.7	4.7	4.7
0.3	0.1	52.8	57.0	56.3	60.4	53.8	57.5	34.6	40.3
	0.3	38.9	52.4	42.1	57.0	40.3	55.2	28.2	41.6
	0.5	31.4	38.7	33.8	42.5	33.3	41.2	26.8	32.4
0.7	0.7	26.6	29.0	30.3	32.9	30.2	33.5	25.5	27.4
	0.9	22.1	22.8	22.9	23.7	23.1	23.6	20.1	20.2
0.2	0.1	93.7	91.2	93.2	94.2	93.7	91.0	83.8	75.1
	0.3	82.1	88.2	83.9	90.1	84.4	89.1	72.1	73.9
	0.5	62.5	67.6	66.9	73.5	66.1	72.5	57.3	62.4
0.9	0.7	50.3	52.7	55.3	58.1	54.9	57.9	45.7	48.3
	0.9	43.0	43.4	43.6	43.9	43.6	44.2	40.6	40.9
0.1	0.1	100.0	99.6	100.0	99.9	100.0	99.6	100.0	97.5
	0.3	98.6	99.0	97.9	98.7	98.7	99.0	95.2	94.9
	0.5	88.7	90.8	91.9	94.0	91.8	94.1	82.0	85.0
0.7	0.7	78.0	79.1	84.2	85.0	83.5	84.2	73.5	74.7
	0.9	69.5	69.9	73.1	73.2	73.0	73.1	67.9	68.0

Table 3

Percentage of 1000 replications for which $H_0 : \rho = 0.5$ was rejected at the $\alpha = 0.05$ level of significance, when $(m, \mu, r) = (5, 20, 40)$.

ρ	ϕ	EPB	EPBS	WB	WBS	MV	MVS	MV*	MVS*
0.1	0.1	0.2	0.6	0.2	0.7	0.2	0.6	0.0	1.1
	0.3	2.1	3.6	2.2	3.6	2.1	3.6	0.6	2.8
	0.5	3.6	4.5	3.2	4.5	3.3	4.5	2.7	2.9
0.7	0.1	1.9	1.9	1.8	1.8	1.8	1.8	2.0	2.0
	0.3	4.1	3.9	4.3	4.3	4.4	4.3	4.3	4.4
	0.5	76.3	78.9	76.3	77.9	76.4	78.9	34.9	43.5
0.3	0.1	49.2	62.2	50.6	63.0	49.3	62.1	30.8	39.9
	0.3	30.7	35.0	32.1	36.8	32.1	36.8	23.3	27.0
	0.5	23.5	23.6	24.5	25.9	25.5	26.6	22.0	22.6
0.9	0.1	17.6	17.6	18.0	18.5	18.0	18.5	19.1	19.7
	0.3	99.7	96.5	99.6	97.8	99.6	96.5	95.2	75.4
	0.5	83.5	90.1	83.7	90.4	83.8	90.9	71.2	70.3
0.2	0.1	59.5	62.9	62.3	65.6	61.4	64.8	49.6	52.8
	0.3	45.0	45.0	47.8	49.4	47.2	48.6	40.7	42.6
	0.5	31.0	31.4	31.4	32.7	32.3	33.7	31.4	32.8
0.1	0.1	100.0	99.8	100.0	99.8	100.0	99.8	100.0	93.9
	0.3	99.2	99.5	99.3	99.5	99.2	99.5	97.0	95.0
	0.5	89.6	90.4	90.0	90.9	89.4	90.6	82.2	84.1
0.7	0.1	73.9	73.2	78.0	78.7	77.9	78.4	67.3	68.1
	0.3	60.2	61.9	62.0	64.9	62.4	64.3	61.2	62.9
	0.5	60.2	61.9	62.0	64.9	62.4	64.3	61.2	62.9

Table 4

Percentage of 1000 replications for which $H_0 : \rho = 0.5$ was rejected at the $\alpha = 0.05$ level of significance, when $(m, \mu, r) = (5, 10, 2)$.

p	ϕ	EPB	EPBS	WB	WBS	MV	MVS	MV*	MVS*
0.1	0.1	0.0	0.2	0.1	1.2	0.0	0.2	0.0	1.4
	0.3	0.4	0.8	0.7	2.2	0.5	1.3	0.2	2.1
	0.5	2.1	3.9	2.7	3.9	2.3	4.3	1.4	3.3
0.5	0.7	4.4	4.9	4.5	4.9	4.4	4.9	3.9	3.3
	0.9	3.2	3.5	3.0	3.1	3.2	3.4	3.4	3.4
0.3	0.1	33.7	38.5	39.1	48.2	34.6	40.8	16.3	30.7
	0.3	29.3	36.8	34.1	43.0	31.5	40.5	19.2	29.1
	0.5	23.5	27.3	28.3	35.1	26.8	32.1	22.3	24.0
0.7	0.7	21.6	22.2	24.6	28.5	24.1	26.7	22.5	22.0
	0.9	16.5	16.5	17.6	18.2	17.3	17.8	17.1	17.4
0.2	0.1	82.9	80.5	85.7	86.4	84.2	80.9	67.6	59.7
	0.3	65.6	70.4	72.3	78.5	70.1	75.3	55.6	55.5
	0.5	48.3	52.5	55.9	62.1	55.9	59.7	46.9	47.3
0.9	0.7	39.0	40.2	45.2	47.9	44.2	46.4	39.7	41.2
	0.9	34.0	34.5	35.6	37.2	35.5	36.9	34.8	35.7
0.1	0.1	98.8	96.7	98.7	98.9	99.3	96.9	97.1	85.0
	0.3	93.4	93.5	94.8	96.9	95.2	95.1	90.3	86.4
	0.5	77.2	80.4	85.8	88.2	84.9	87.4	78.7	80.6
0.7	0.7	65.7	66.0	72.0	73.1	71.5	72.4	66.9	67.9
	0.9	61.9	62.0	63.4	65.9	63.4	65.5	62.6	64.2

Table 5

Results from data application: The upper portion of the table presents the number of termites recorded on the treated(untreated) side of the Petri dishes, for the different dilution levels. Note, Dilution 1 was the low dose setting and Dilution 3 was the high dose setting. The lower portion of the table provides the p-values obtained by the EPB-, EPBS-, WB-, WBS-, MV-, and MVS-tests.

Observed data					
Dish	Control	Dilution 1	Dilution 2	Dilution 3	
1	0(9)	0(7)	0(10)	7(2)	
2	0(9)	0(8)	0(10)	1(8)	
3	8(2)	0(9)	0(10)	0(9)	
4	5(5)	0(9)	0(8)	0(8)	
5	0(10)	0(4)	2(8)	7(2)	
6	6(4)	8(2)	0(10)	0(10)	
7		8(1)	2(8)	7(3)	
8		3(7)	5(5)	0(10)	
9		0(10)	5(5)	7(2)	
10			7(3)	0(9)	

Hypothesis testing results						
	EPB	EPBS	WB	WBS	MV	MVS
Control	0.155	0.145	0.141	0.130	0.141	0.131
Dilution 1	0.051	0.049	0.030	0.029	0.032	0.030
Dilution 2	0.010	0.009	0.009	0.009	0.009	0.009
Dilution 3	0.091	0.085	0.084	0.078	0.084	0.078