An Enhanced Sign Test for Dependent Binary Data with Small Numbers of Clusters

Patrick D. Gerard

Experimental Statistics Unit, Mississippi State University

Box 9653, 151 Dorman Hall, Mississippi State, Mississippi 39762, USA

Email: pdg1@ra.msstate.edu

March 10, 2005

William R. Schucany
Department of Statistical Science, Southern Methodist University
Dallas, Texas 75275-0332, USA

Summary. We propose a test, like the classical sign test, of whether the probability of an event differs from 0.5 that is appropriate with clustered binary data. It combines a permutation approach and an exact parametric bootstrap calculation. Simulation studies show it to be superior to a sign test based on aggregated cluster level data. The new test is more powerful than or comparable to a standard permutation test whenever 1) the number of clusters is small or 2) for larger cluster numbers under strong clustering resulting from within cluster correlations of greater than .80. The results from a chemical repellency trial are used to illustrate the three test methods.

Keywords and phrases: binomial, bootstrap, correlation, permutation, power, simulation

1. Introduction

Clustered binary data are commonly encountered in a variety of areas. Situations such as survival after exposure to a contaminant often yield results that are similar for animals from the same litter. Members of a household may hold similar Yes/No opinions on political issues. Trees within the same orchard may tend to have contracted a disease in such a way that knowledge of one tree's status may be useful in predicting the disease status of other trees in the same orchard. In each of these cases the response of interest may be dichotomous and these binary responses within the same cluster (litter, household, orchard) are typically not independent.

Methods for handling clustered binary data range from the simplistic summary aggregation of the clusters to the more sophisticated fitting of generalized linear models with random effects that account for clustering. In some cases, however, these techniques may not be totally satisfactory. Consider an experiment with clustered binary data in which the objective is to test whether the marginal probability of a response differs from 0.5. This is the null hypothesis tested with the classical sign test, but here we have clustered binary data. In some settings the number of clusters may be small (≈ 5), which would severely reduce the power of simplistic aggregation techniques. A generalized linear model with a random cluster effect can be fit and tests conducted but there may be concerns about maintaining the Type I error rate with small numbers of clusters. We propose a modification of the sign test for use with clustered binary data. It involves both permutation and parametric bootstrap techniques. It has improved power compared to simplistic aggregation and is easily carried out for problems with as few as 4 or 5 clusters.

In Section 2 a real example motivates the problem. In Section 3 the new test is introduced. Section 4 describes simulation results, and Section 5 returns to the motivating application.

2. Correlated Binary Responses Example

Laboratory trials are conducted to determine the effectiveness of chemical insecticide compounds. Specifically, the trial design outlined here involves evaluating repellency. These trials are conducted with a small number of Petri dishes (5-10), which are treated with the chemical of interest on one half of the Petri dish, with the other half left untreated. A small number of insects (10-25) are then placed in the middle of each Petri dish and after a specified time period, the numbers of insects on the treated and untreated sides are determined for each dish.

One relevant test is whether the chemical of interest performs differently from what is expected by chance, i.e., 50/50. A specific alternative is whether the marginal probability of being on the treated side is less than 0.5. This would indicate that the chemical has some repellency. One complicating factor here is that many insects (e.g. termites) are social creatures that tend to follow one another. Various reasons for this exist, but the primary one seems to be the secretion of pheromones that draw the insects to one another. Hence, the insects within each Petri dish are not independent, but those in different dishes can be assumed to be.

In one such test, catnip oil was being evaluated as a potential repellant of termites. The data for a single dose of catnip involved 5 Petri dishes and 10 termites per dish. Figure 1 depicts a typical Petri dish. The data for 5 dishes in a single trial are 4/10, 0/10, 0/10, 1/10, and 5/10 termites on the treated side. Obviously there is a general tendency for termites to gravitate to the untreated side of the dishes. The results of a test of the null hypothesis mentioned in the previous paragraph would be of interest here.

(Figure 1 about here)

One simple way to analyze data of this type is to declare the entire Petri dish as a success or failure regarding repellency by whether less than 50% of the termites are on the treated side (majority rule). Because the dishes are independent, the classical sign test can be used to test the hypothesis of interest. This test relies upon the proportion of repelled dishes to assess whether that observed proportion is significantly different from 0.5. Obviously, with only 5 dishes, this test will have limited power. On the other hand if these data were fit to a logistic regression model with dish as a random effect, controlling Type I error rate may be a concern.

In some other design settings it may be possible to assign treatments randomly to items within a cluster, rather than to the entire cluster. See Moerbeek (2005) for a comparison of the robustness of these two randomizations. In the next section we propose a new method for testing this hypothesis with clustered binary data. Like the classical sign test, it has a permutation basis, but with an added bootstrap component.

3. Hypothesis Tests for No Effect

There are several mechanisms used to model clustered binary data. These include random cluster effects (McCulloch and Searle, 2001) or allowing the probability of success to vary according to some probability distribution, such as the beta distribution, which results in a marginal beta-binomial distribution (MuCullagh and Nelder, 1989; Agresti, 2002). Williams (1975) advocates the use of likelihood ratio tests in the latter of these two settings. Kupper and Haseman (1978) proposed a correlated binomial model employing a correction factor to adjust the standard binomial probabilities from independence. Rao and Scott (1992) suggested a more general correction to binomial probabilities based on a deviation of the variance from that expected under independence. This paper will adopt a random effects model for clustering. For example, the random variables associated with the responses of each unit have Bernoulli distributions. Specifically,

$$X_{ij} \mid p_i \sim bin(1, p_i),$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \theta + \beta_i, \tag{3.1}$$

and

$$\beta_i \sim F,$$

where X_{ij} is the binary response for the jth unit within the ith cluster, $i = 1,\ldots, m$, the number of clusters, $j = 1,\ldots, n_i$, β_i is a random effect associated with the ith cluster, and θ is a log-odds parameter related to the marginal probability of success. Another link function such as the probit function could be used here. Typically the β_i are assumed to follow some symmetric probability distribution with mean 0 and variance σ_β^2.

As in the traditional sign test, the null hypothesis pertains to the log odds

$$\mathrm{H_o} : \theta = 0, \tag{3.2}$$

which corresponds to a probability of success equal to 0.5, versus one- or two-tailed alternatives. In the application described in Section 2, a one-sided alternative, $\mathrm{H_a} : \theta < 0$, would be appropriate indicating that a termite tends to avoid the treated side.

As is commonly assumed in random effects models, the X_{ij} are independent, conditional on the random effects, but with different probabilities of success. Unconditionally, however, observations within the same cluster, X_{ij} and X_{ik}, are correlated. It is well known (Gastwirth and Rubin, 1971) that to naively ignore this correlation can have dramatic effects on hypothesis tests. One simple way of removing the effects of the correlation that does not require parametric modeling is to classify each cluster as a "success" or "failure" based on the preponderance of the outcomes within the cluster. For example, for $\mathrm{H_a} : \theta < 0$, a cluster can be deemed a "success" if less than 50% of the units within a cluster achieve the outcome of interest. In our example, a dish is considered a "success" if less than 50% of the termites are on the treated side. Under the null, the probability that a cluster will be deemed a success is 0.5 for an odd number of items in a cluster and less than 0.5 for an even number. Now a sign test can be carried out at the cluster level. If the number of clusters is small, this procedure will suffer from low power, but this is avoidable only if one is discarding useful information.

Under the null hypothesis (3.2), (3.1) reduces to $\log\left(\frac{p_i}{1-p_i}\right) = \beta_i$. Hence, the p_i are determined solely by unobserved random variables, β_i. Thus, $X_{ij}|\beta_i$ are independent binomial random variables with probability of success p_i. Even though the β_i are not observable, they may be estimated, or more conventionally for random effects models, predicted from the data. However, the sole reason that we would want to predict β_i would be to estimate p_i, which can be done directly from $\hat{p}_i = \sum_{j=1}^{n_i} \frac{X_{ij}}{n_i}$. This estimated quantity corresponds to a (null) predicted value of β_i, $\hat{\beta}_i = \ln[\hat{p}_i/(1-\hat{p}_i)]$ for $\hat{p}_i \in (0,1)$. To carry out our hypothesis test, we replace β_i and p_i with predicted and estimated quantities $\hat{\beta}_i$ and $\hat{p}_i$, respectively. Thus for each i, we treat $X_{ij}|\hat{\beta}_i$, or equivalently $X_{ij}|\hat{p}_i$, as independent Bernoulli random variables with estimated probability of success $\hat{p}_i$. Our test statistic is

$T = \sum_i \sum_j X_{ij}$ = total number of successes in all clusters. We will use a one-sided, lower tail critical region for illustration.

Now under the null hypothesis and the assumption of symmetry, β_i and - β_i are equally likely and hence p_i is as likely as 1-p_i. In the spirit of permutation tests, we evaluate every permutation of the signs of $\hat{\beta}_i$ (and hence permute $\hat{p}_i$ with 1- $\hat{p}_i$). For each of the 2^m permutations conditional on predicted random effects, we again treat the dichotomous outcomes as independent Bernoulli random variables with estimated probability of success $\hat{p}_i$ or 1- $\hat{p}_i$. Next it is necessary to determine how likely the value of our test statistic is under the null hypothesis. For each permutation let $\hat{p}_{ri}$ * be the estimated probability of success for the rth permutation and the ith cluster. Additionally, let $\hat{\underline{\mathbf{p}}}_r * = (\hat{p}_{r1}*,...,\hat{p}_{rm}*)$ denote the vector of estimated probabilities for the rth permutation. Designate $\hat{\boldsymbol{\beta}}_r$ * analogously. Consider Bernoulli random variables $X_{ij}* \mid \hat{\boldsymbol{\beta}}_{\mathrm{r}}*$, which are viewed as independent with probabilities of success given in $\hat{\underline{\mathbf{p}}}_r$ *. Define also $T_r* = \sum_i \sum_j X_{ij}* \mid \hat{\boldsymbol{\beta}}_r*$, which is the test statistic computed for the rth permutation.

For each permutation the probability that T_r * is less than or equal the observed value $T = t$ is required. Within each of the 2^m permutations, a parametric bootstrap might be used to estimate this probability. This would yield an estimated p-value of the form

$$\hat{p}_{Boot} = \sum_{r=1}^{2^m} \sum_{b=1}^{B} \frac{I\left(T_{rb}^* \le t\right)}{2^m B} ,$$

where $I(\cdot)$ is the indicator function, B is the total number of bootstrap resamples, and T_{rb}^* is the test statistic from the bth bootstrap resampling of the rth permutation.

Note, however, that

$$\sum_{b=1}^{B} \frac{I\left(T_{rb}^* \le t\right)}{B} \tag{3.3}$$

is an estimate of the probability that a sum of independent Bernoulli random variables with nonconstant probabilities of success, defined by $\hat{\underline{\mathbf{p}}}_r$ *, is less than or equal to the observed test statistic. An algorithm of Thomas and Taub (1982) for computing probability distributions for sums of independent Bernoullis with unequal probabilities of success can be used to evaluate the population value of (3.3). This eliminates the need for any resampling. Denote by p_{Tr} this exact (conditional) probability that is estimated in (3.3). Thus our proposed p-value is

$$p\text{-value} = \sum_{r=1}^{2^{n_c}} \frac{p_{Tr}}{2^m},$$

which may be compared to any desired significance level to carry out the hypothesis test.

To further illustrate, let $C_1, C_2, \ldots, C_N$, $N = 2^m$, be the configurations generated by the permutations under the null hypothesis. For the proposed test we are computing

$$P(T \le t) = \sum_{r=1}^{N} P(T_r^* \le t \mid C_r) P(C_r). \qquad (3.4)$$

Each configuration, C_r, is associated with estimated success probabilities $\hat{\underline{\mathbf{p}}}_r^*$. Now, $P(T_r^* \le t \mid C_i)$ is computed as the sum of Bernoulli random variables with unequal probabilities of success given in $\hat{\underline{\mathbf{p}}}_r^*$. Here $P(C_r) = 1/N$ because the configurations are equally likely under the null. For comparison purposes, we include a permutation test (*CP*) for which (3.4) also applies, but now $P(T_r^* \le t \mid C_r)$ is 1 if $T_r^* \le t$ and 0 otherwise.

For example consider a data set with only m= 3 clusters and 10 observations per cluster. Suppose that one cluster has 1 success, the second has 3 successes, and the third has 8 successes. Hence the success proportions are 0.1, 0.3, and 0.8. The observed number of successes, t = 12, in this case. The $N = 2^m = 8$ configurations are given in Table 1. For each configuration, the number of successes associated with $\hat{\underline{\mathbf{p}}}_r^*$ are summed and compared with t. The count of configurations with total number of successes less than or equal to t divided by the number of configurations is the p-value for *CP*. This is the average of the estimated probabilities in the Permutation column, which in this case is 3/8. For the new parametric bootstrap test, the probabilities of success for each permutation are used to compute (3.4). Parametric bootstrap resamples (B =5000) are used to compute (3.3) and the exact calculation done using the algorithm of Thomas and Taub (1982). The two columns give virtually identical results and the latter can be averaged to obtain the proposed p-value for our new test. The actual p-values are 2.7652/8 for the resampled bootstrap and 2.7761/8 for the exact calculation. Each is smaller than the p-value obtained from *PT*. In the next section we use simulation to compare these three tests.

(Table 1 about here)

4. Simulation Results

The performance of the new exact parametric bootstrap test introduced in Section 3 (*EPB*) was evaluated using Monte Carlo simulation methods. The null hypothesis (3.2) is tested against a one-sided alternative $H_a : \theta < 0$. We found, as expected, that the p-values in (3.3) and (3.4) differ only slightly due to sampling error, so only the exact ones (3.4) are reported here. In addition the cluster level sign test (*CLS*) and another permutation test (*CP*) described in Section 3 were also used for comparison.

In *CP* each permutation involves exchanging the number of successes in a cluster with the number of failures. The test statistic used in each permutation is the number of successes and the *p*-value is the number of permutations for which the total number of successes is less than or equal to the observed number of successes, divided by the total number of permutations. As mentioned previously, in *CLS* a cluster is deemed a success whenever less than half of the units in a cluster achieve the outcome of interest. The *p*-value is computed from the probability that a binomial random variable with number of trials equal the number of clusters and success probability .5 is greater than or equal to the number of cluster successes. The hypothesis of interest could also be tested using the NLMixed procedure of SAS® (Agresti 2002). However, some preliminary simulations resulted in greatly inflated Type I error rates in situations with small numbers of clusters and moderate to large numbers of units per cluster, most notably with weak to moderate clustering. Because this test did not consistently maintain its size, it is not part of the overall simulation study. All simulations were conducted using SAS software.

Consider four scenarios with a fixed number of clusters and cluster size and 1000 independent replications each. Clustered binary random variables are generated using the technique described in Lunn and Davies (1998), which allows specification of the marginal probability of success as well as a clustering value whose square is the correlation between units within a cluster, $\phi^2 = \rho$. In each scenario, the marginal probabilities, 0.5 (null of (3.2) is true), 0.3, 0.2, and 0.1, are combined with clustering values, ϕ, 0.1, 0.3, 0.5, 0.7, and .9. Hence the within cluster correlations range from .01, essentially independent, to .81, which is strong clustering. The four scenarios investigated were a) 5 clusters with 10 units/cluster, b) 5 clusters with 20 units/cluster (Table 2), c) 10 clusters with 5 units/cluster (Table 3), and d) 10 clusters with 10 units/cluster (Table 3).

We also investigated the performance of a binomial test that essentially ignores clustering information. We found that for the smallest clustering value, $\phi = 0.1$, this test maintained its size. This is not surprising because the smallest clustering value yields data that are practically independent. However, the Type I error rate was larger than the nominal level for all other clustering values and so this test did not deserve to be in our simulations.

Each table gives the percentage of times that $H_o : \theta = 0$ (3.2) is rejected. In all simulations one-sided tests are at the 5% level. The standard errors of the table entries are 0.7% at the null and bounded by 1.6% at the alternatives. The significance of the differences among the three tests (bold font for the winner) is assessed with an exact conditional version of McNemar's test, Lehmann (1998), page 268 and Bonferroni corrections for the multiple comparisons within each table. The significance of the levels below the nominal 5%, obtained with the binomial (denoted by italics), are also Bonferroni corrected. Figure 2 displays the estimated powers of the three tests for all four scenarios at a marginal probability of 0.2.

(Figure 2 about here)

Results for 5 clusters with $n_i \equiv 10$ units per cluster (Table 2) show that all three tests control size, at the cost of being conservative, especially for weak clustering. *CLS* and *CP* are identical here. *EPB* has significantly higher power, except for essentially independent data ($\phi = .1$) and marginal probability of success 0.3. The

results for 5 clusters and 20 units per cluster are very similar. Rejection proportions under the null are closer to 5%, power values are greater, and *EPB* has uniformly higher power.

(Table 2 about here)

For 10 clusters and $n_i \equiv 5$ units per cluster (Table 3), *CP* provides empirical levels close to the nominal 5%. *CP* also enjoys a power advantage at small clustering values, with *EPB* having nearly equivalent, or slightly greater, power for large clustering values. For 10 clusters and 10 units per cluster (Table 3), the results are similar to those in Table 3. Again, *CP* has greater power for small clustering values ($\phi = 0.1, 0.3$) and *EPB* performs as well or better for large clustering values, ($\phi = 0.9$).

(Tables3 about here)

CLS is equivalent to *CP* for scenarios a) and b) and performs significantly worse than both *EPB* and *CP* in the other two scenarios. The results of *CLS* and *CP* are equivalent in the special case of $m = 5$ clusters. With exactly 5 clusters, *CLS* will reject H_o with significance level .05 only if all 5 clusters have less than 50% of units have the attribute of interest. This yields the smallest total possible among all permutations in the permutation test, *CP*, which is also significant. Conversely, at $m = 5$ the *CP* is significant only if the smallest total possible among all permutations is observed, which in turn implies that less than 50% of units in each cluster achieve the outcome of interest which results in significance for *CLS*.

5. Example

The results of a repellency trial for termites involving catnip are given in Section 2. In this particular trial, $m = 5$ Petri dishes were evaluated, with $n_i = 10$ termites on each dish. Half of the area of each Petri dish was treated and the other half left untreated. After a specified period of time the numbers of termites were recorded, yielding 4/10, 0/10, 0/10, 1/10, and 5/10 termites on the treated side for the five dishes. One can see that in 4 of the 5 dishes, less that half of the termites were found on the treated side. Hence, the *p*-value for a one-sided sign test, *CLS,* of hypothesis (3.2) is .1875. The result of the permutation test, *CP,* which simply permutes the treated and untreated labels, yields a *p*-value of .0625. The *p*-value of the new test introduced in Section 2 *EPB* is .0534. So in this example, we obtain a smaller *p*-value using the new test than with either *CLS* or *CP*.

6. Discussion

. In somewhat related work Miao and Gastwirth (2004) examine confidence intervals from dependent binary data. In a different context Antolini, Nam, and D'Agostino (2004) consider nonparametric inference from a set of correlated indicators.

It should also be pointed out that this new test allows one to reject hypotheses in situations in which it is simply not possible with *CLS* or *CP*. One example is an experiment with 4 clusters. The smallest *p*-value possible for *CLS* and *CP* with a one-sided alternative is .0625, which does not permit rejection at the .05 significance level. Our simulations show that the new test maintains size in this setting with 10

units per cluster and does allow one to reject the null hypothesis, achieving power of greater than .50 in some cases. However, as the strength of the effect increases, with marginal probabilities of the outcome of interest less than 0.1, the power tends to decrease rather than increase. The reason is that as the strength of the effect increases, the behavior of *EPB* approaches that of *CP*, which does not allow rejection.

Therefore, we have identified some situations in which the new test for clustered binary data is superior to simple cluster level aggregation followed by a sign test. Additionally, when the number of clusters is small, it outperforms the standard permutation test. For larger numbers of clusters and with strong clustering it performs similarly or better than the standard permutation test.

We emphasize that this new test tends to perform better than the other tests considered in practical applications such as our example of chemical repellency testing. In these settings, the number of clusters tend to be small and strong clustering is present, precisely where the new test is the superior methodology.

Acknowledgements

The authors thank Chris Peterson of the USDA Forest Service, Wood Products Insect Research Unit, for bringing this problem to our attention and for providing useful insight as well as the data in the example.

References

Agresti, A. (2002). *Categorical Data Analysis*. 2nd Edition. John Wiley and Sons, Inc. Hoboken, New Jersey.

Antolini, L., Nam, B-H., and D'Agostino, R.B. (2004). Inference on Correlated Discrimination Measures in Survival Analysis: a Nonparametric Approach. *Communications in Statistics, Theory and Methods* **33**, 2117-2135.

Gastwirth, J.L. and Rubin, H. (1971) Effect of Dependence on the Level of Some One- Sample Tests. *Journal of the American Statistical Association* **66**, 816-820.

Kupper, L.L. and Haseman, J.K. (1978). The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments. *Biometrics* **34**, 69-76.

Lehmann, E.L. (1998) *Nonparametrics : Statistical Methods Based on Ranks*. Prentice Hall.

Lunn, A.D. and Davies, S.J. (1998). A Note on Generating Correlated Binary Variables. *Biometrika* **85**, 487-490.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. 2nd Edition. Chapman and Hall. New York, New York.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley and Sons, Inc. New York, New York.

Miao, W. and Gastwirth, J.L. (2004). The Effect of Dependence on Confidence Intervals for a Population Proportion. *The American Statistican* **58**, 124-130.

Moerbeek, M. (2005). Randomization of Clusters Versus Randomization of Persons Within Clusters: Which is Preferable? *The American Statistican* **59**, 72-78.

Rao, J.N.K. and Scott, A.J. (1992). A Simple Method for the Analysis of Clustered Binary Data. *Biometrics* **48**, 577-585.

Thomas, M.A. and Taub, A.E. (1982). Calculating Binomial Probabilities When the Trial Probabilities are Unequal. *Journal of Statistical Computation and Simulation* **14**, 125- 131.

Williams, D.A. (1975). The Analysis of Binary Responses From Toxicological Experiments Involving Reproduction and Teratogenicity. *Biometrics* **31**, 949-952.

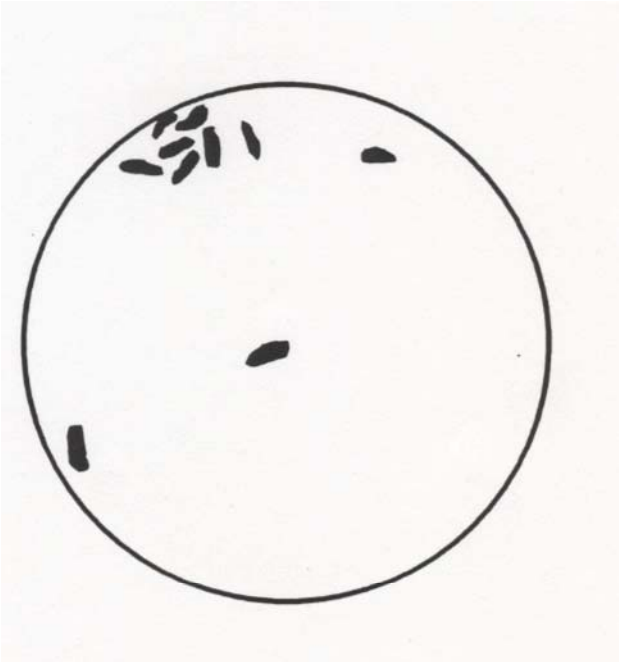

Figure 1. Petri dish in termite repellency test.

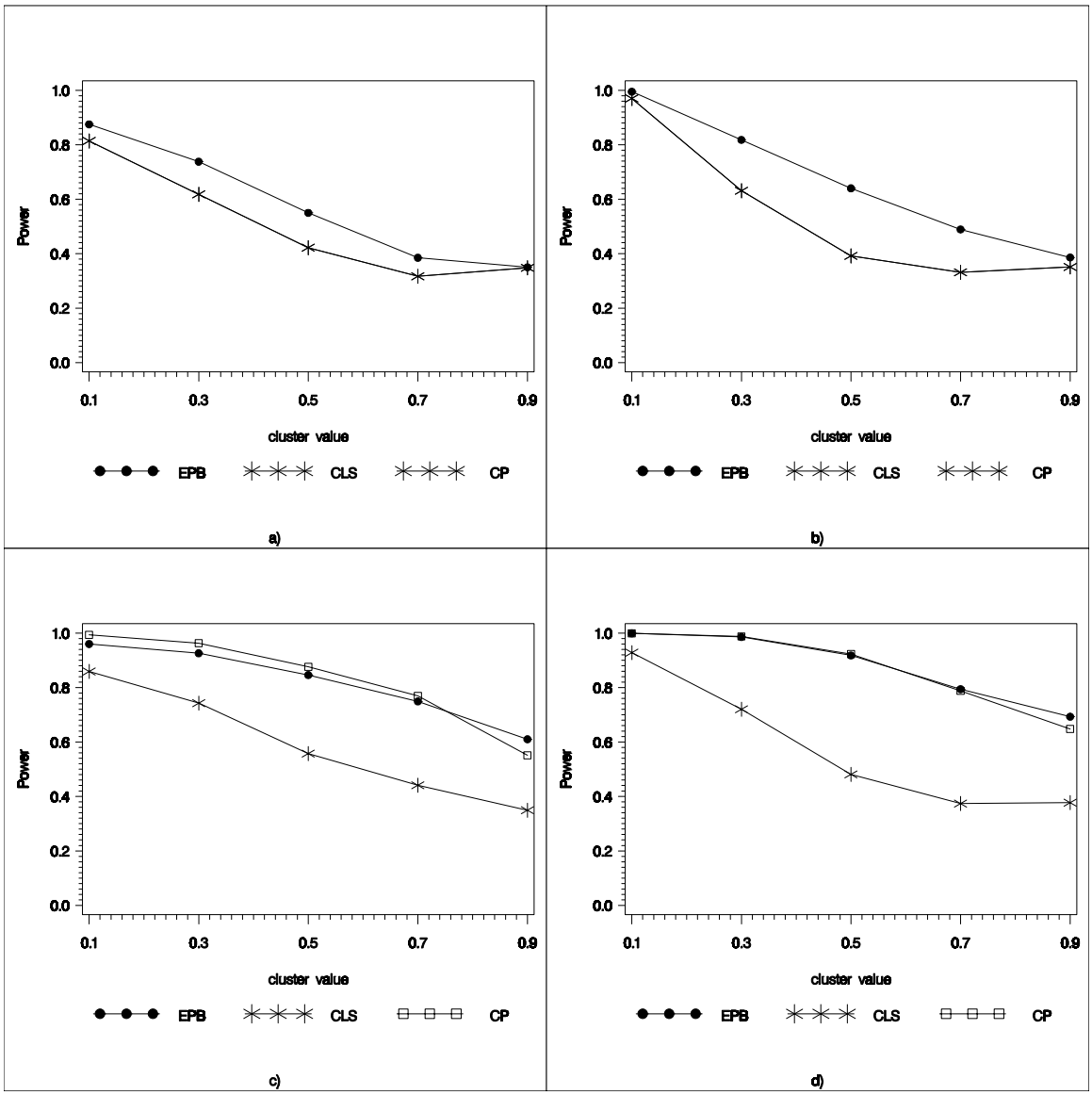

Figure 2. Power results with marginal probability .2 for Scenarios a) through d) for the proposed test (EPB), the sign test on aggregated cluster data (CLS), and the permutation test (CP).

Table 1. Illustration of *p*-value calculations for 3 clusters with 10 units per cluster, having 1, 3, and 8 successes, respectively.

Configuration (r)	$\hat{p}_{r1}^{*}$	$\hat{p}_{r2}^{*}$	$\hat{p}_{r3}^{*}$	$P(T_r^* \le t \mid C_r)$ (Permutation)	$P(T_r^* \le t \mid C_r)$ (Bootstrap) simulated	exact
1	.1	.3	.8	1	.5968	.5968
2	.1	.3	.2	1	.9978	.9975
3	.1	.7	.2	1	.8688	.8791
4	.1	.7	.8	0	.0528	.0532
5	.9	.3	.2	0	.2436	.2436
6	.9	.3	.8	0	.0002	.0003
7	.9	.7	.2	0	.0052	.0056
8	.9	.7	.8	0	.0000	.0000
Average				.375		.3470

Table 2. Percentages of 1000 replications rejected with the nominal α= 5 % tests for twenty combinations of marginal probability and clustering value when there are 5 clusters all with either 10 or 20 units per cluster.

Marginal Probability	Clustering Value, ϕ	$n_i = 10$ *EPB*	$n_i = 10$ *CLS≡CP*	$n_i = 20$ *EPB*	$n_i = 20$ *CLS≡CP*
.5(null)	.1	*0.1*	***0.9***	*0.2*	*1.3*
.5	.3	*0.4*	*1.1*	*2.4*	*2.7*
.5	.5	*2.4*	*2.8*	**4.3**	**4.1**
.5	.7	*2.8*	*3.1*	**4.1**	**3.7**
.5	.9	**3.3**	**3.3**	*2.4*	*2.4*
.3	.1	36.2	**41.4**	**72.8**	**71.4**
.3	.3	30.0	**33.1**	**47.5**	37.3
.3	.5	**27.7**	20.3	**30.9**	18.8
.3	.7	**18.2**	14.7	**22.8**	17.2
.3	.9	**16.5**	**16.5**	**19.4**	18.2
.2	.1	**87.5**	81.4	**99.5**	97.0
.2	.3	**73.8**	61.8	**81.8**	63.1
.2	.5	**55.0**	42.2	**64.0**	39.2
.2	.7	**38.5**	31.7	**48.9**	33.1
.2	.9	**35.0**	**34.8**	**38.6**	35.1
.1	.1	**99.8**	98.0	**100**	**99.9**
.1	.3	**95.3**	86.3	**99.0**	92.7
.1	.5	**84.7**	69.3	**91.8**	68.9
.1	.7	**65.2**	57.1	**81.0**	59.1
.1	.9	**59.6**	**59.4**	**62.4**	57.7

Table 3. Percentages of 1000 replications rejected with the nominal α= 5 % tests for twenty combinations of marginal probability and clustering value when there are 10 clusters and 5 or 10 units per cluster.

Marginal Probability	Clustering Value, ϕ	5 units per cluster			10 units per cluster		
		EPB	*CLS*	*CP*	*EPB*	*CLS*	*CP*
.5(null)	.1	*0.9*	*1.8*	**4.9**	*0.4*	*0.2*	***3.1***
.5	.3	*1.1*	*0.7*	**4.0**	*1.2*	*0.2*	***2.5***
.5	.5	*2.3*	*0.3*	**4.1**	3.7	*0.9*	**4.9**
.5	.7	*2.1*	*0.9*	*2.8*	**5.3**	*1.6*	**5.4**
.5	.9	*4.0*	*1.4*	**2.8**	**5.5**	*1.4*	3.6
.3	.1	62.5	47.7	**81.5**	92.3	53.6	**96.7**
.3	.3	57.2	36.5	**72.3**	77.9	31.9	**82.8**
.3	.5	47.3	24.8	**55.8**	58.3	17.0	**61.5**
.3	.7	41.9	17.2	**45.8**	**46.9**	17.1	**47.3**
.3	.9	**34.9**	15.4	29.1	**38.0**	14.0	34.6
.2	.1	96.0	.85.9	**99.4**	**99.9**	92.8	**99.9**
.2	.3	92.6	74.3	**96.3**	**98.6**	72.0	**98.8**
.2	.5	84.6	55.7	**87.6**	**91.8**	48.1	**92.3**
.2	.7	**74.9**	44.1	**76.9**	**79.4**	37.4	**78.8**
.2	.9	**61.0**	34.9	55.1	**69.3**	37.7	64.8
.1	.1	**100**	**99.6**	**100**	**100**	**99.9**	**100**
.1	.3	**99.9**	96.7	**100**	**100**	96.4	**100**
.1	.5	**99.2**	89.1	**99.5**	**100**	82.8	**100**
.1	.7	**96.5**	77.7	**96.3**	**97.9**	75.8	**97.8**
.1	.9	**89.0**	76.3	87.0	**91.8**	71.2	89.2