

## Critical Values and Power for a Small Sample Test of Difference in Proportions in the Presence of Extra-Binomial Variation

John S. Lawson<sup>1</sup> and Benjamin Ahlstrom<sup>2</sup>  
<sup>1</sup>*Brigham Young University and* <sup>2</sup>*Bandag Inc.*

*Abstract:* We develop a likelihood ratio test statistic, based on the beta-binomial distribution, for comparing a single treated group with dichotomous data to dual control groups. This statistic is useful in cases where there is overdispersion or extra-binomial variation. We apply the statistic to data from a two year rodent carcinogenicity study with dual control groups. The test statistic we developed is similar to others that have been developed for incorporation of historical control groups with rodent carcinogenicity experiments. However, for the small sample case we considered, large sample theory used by the other test statistics did not apply. We determined the critical values of this statistic by enumerating its distribution. A small Monte Carlo study shows the new test statistic controls the significance level much better than Fisher's exact test when there is overdispersion and that it has adequate power.

*Key words:* Beta-binomial distribution, dual control groups, extra-binomial variation, Fisher's exact test, historical controls, likelihood ratio test, overdispersion, rodent carcinogenicity bioassay, significance level.

### 1. Introduction

When testing the null hypothesis of equal binomial parameters with random samples from two populations, Fisher's Exact test, the continuity corrected Chi-square test, or the  $z$ -test based on normal theory are the test statistics generally used. When there are replicate random samples from each population, a  $t$ -test (using the arcsin square root of the proportions as data) is a common method of analysis. However, the variation in proportions among replicate random samples from the same population may exhibit greater variation than would be predicted by the simple binomial model. This is often the case in toxicological studies, where the proportion of litter mates with certain birth defects may vary more from litter to litter within the same treatment than would be predicted by the binomial model. The increase in variation over that predicted by the binomial model is called overdispersion. When overdispersion is present in the data, the analysis based on the  $t$ -test may not be the most appropriate.

Kleinman(1973) proposed a weighted least squares approach for inference on proportions in the presence of overdispersion. Williams(1975) developed a likelihood ratio statistic, for comparing two proportions in teratology studies in cases where there is overdispersion among replicate groups. His statistic was based on the beta-binomial distribution. Kupper and Haseman(1978) developed a similar likelihood ratio statistic based on the correlated-binomial model. Pack(1986) showed, through simulation studies, that the likelihood ratio test based on the beta binomial model was superior to the alternate parametric models. In toxicology-teratology studies there are usually 20-40 litters (or replicate random samples) per group, and the large sample properties of the likelihood ratio statistic guarantee that the asymptotic Chi-square distribution is appropriate for the beta-binomial likelihood ratio statistic.

In addition to applications in toxicology and teratology studies, the beta-binomial likelihood ratio test of proportions has also been applied to animal carcinogenicity studies when the study results are combined with historical controls. Tarone, Chu and Ward (1981) found overdispersion when they examined tumor rates in historical controls, and this will normally be the case due to differences from study to study. Hoel (1983) and Tarone (1982) proposed tests for carcinogenic effect utilizing historical controls. Both of these tests followed Williams' (1975) idea of using the beta-binomial distribution to model the overdispersion (or extra-binomial variation in proportions). With a large number of historical control groups, their likelihood ratio test statistics follow the Chi-square distribution due to large sample properties. Hoel(1983) also considered exact tests. Since Hoel and Tarone's seminal papers on the incorporation of historical controls in tests for carcinogenic effects, additional research has been done to expand and extend their results.

The asymptotic Chi-square null distribution of the beta-binomial likelihood ratio test of proportions with overdispersion is justified in toxicology studies with 20-40 litters per group or in animal carcinogenicity studies where there are 50 or more historical control groups. However, in cases where there are very few replicate samples per treatment group, using the large sample properties will not be valid. In this paper we examine the case where there are only two replicate samples for the control group and one for the treated group. We develop a likelihood ratio statistic with overdispersion, along the lines of Williams(1975), but then find its exact null distribution by enumerating all possible tables of counts. We study the significance level of this test and its power via a simulation study. We found it to be quite acceptable.

## 2. An Illustrative Data Set

Long-term animal experiments for routine testing of chemicals is part of the

procedure for setting public health regulations in many countries (see Gart et al.(1986)). There has been discussion of the use of two identical control groups in these studies as a quality control check for identifying possible biases within the design (see Society for Toxicology(1982)). Table 1 presents data from one such study. This was a two-year study with Swiss crl:CD-1 BR mice that was submitted to the United States Environmental Protection Agency June 29, 1993<sup>1</sup> (see Hansen(1994)). The table shows the number of male mice with liver adenoma/carcinoma in dual control groups and three treatment groups. There was no statistically significant difference in survival between the high dose and control groups in this study. Therefore it should be possible to get a first indication of carcinogenic activity by comparing the crude proportion of tumors between the control and treated groups.

Table 1: Data from rodent carcinogenicity study

Group	animals with tumor	animals without tumor	total
Control 1	1	49	50
Control 2	8	42	50
Treated (low dose)	9	41	50
Treated (mid dose)	10	40	50
Treated (high dose)	12	38	50

The regulatory agency reviewing the data suggested that each treatment group should be compared to each control group and to the combined control groups, using Fisher's Exact test, to determine if there was an increase in tumors. Using this approach there was no significant increase in tumors for the low and mid-dose group, but there was a significant increase for the high-dose group. The data for the last comparison (control groups versus high-dose group) is the data we will use to illustrate the likelihood ratio statistic we develop.

The problem with the data in Table 1 is that there is a significant difference in the proportion of tumors between the two control groups ( $p = 0.0154$ ). Therefore, it appears that there is a highly significant increase in tumors for the high dose when compared to control group 1, but there is no significant increase when compared to control group 2. This situation can arise with dual control groups and is exactly the case that Haseman et. al (1990) describes as difficult to interpret. When the results of a study are difficult to interpret, they cannot simply be ignored. Some way must be found to explain and interpret the data for the regulatory agency where it is submitted.

<sup>1</sup>See Hansen, L. J. (1994). United States Environmental Protection Agency Memorandum from Linnea J. Hansen (Health Effects Division) to Robert Brennis and Joseph Tavano (Registration Division) dated June 22, 1994.

Is it possible that two control groups could have different tumor rates in the same study? There are many reasons cited in the literature why this could be the case (see Haseman et.al. (1989), and Haseman (1995)). The most common cause of differences in tumor rates between supposedly equivalent groups is differences in mortality. If this is the case, the appropriate way to test for a treatment effect would be to do a mortality adjusted analysis rather than an analysis of the crude proportions. However, there was no difference in mortality for the groups in Table 1, and methods for mortality adjusted analysis will not be discussed here.

Other reasons for a difference in tumor rates between equivalent control groups in the same study include things such as differences in preparation of tissue slides, differences among histology technicians reading the slides, and time related diagnostic shifts in reading the slides. Haseman et. al. (1986) said there may also be random differences in tumor rates between equivalent groups and estimated there is a 47-50% chance that some tumor type may show significant differences by chance. We believe the beta-binomial overdispersion model is a good way to model these random or unexplainable differences and provide a reasonable way to explain and interpret the data in Table 1. We believe the use of Fisher's exact test when there is a significant difference in the control groups is inappropriate, and we show in Section 7 that the significance level of Fisher's Exact test is inflated when there is overdispersion.

### 3. Modeling Overdispersion

To clarify the concept of overdispersion we will first describe it for quantitative variables and then show the analogous case for dichotomous data. If the response variable in a study is quantitative, such as animal weight, and there are  $J_i$  independent samples or groups of  $n_{ij}$  animals for each of two treatments, the model for the data can be written as:

$$Y_{ijk} = \mu_i + e_{ijk} \quad (3.1)$$

where  $i = 1, 2, j = 1, \dots, J_i, k = 1, \dots, n_{ij}$ , and  $e_{ijk} \sim N(0, \sigma^2)$ .  $Var(Y_{ijk}) = \sigma_e^2$ . If there is overdispersion an additional term is added to the model

$$Y_{ijk} = \mu_i + G_{ij} + e_{ijk} \quad (3.2)$$

where  $G_{ij} \sim N(0, \sigma_G^2)$ . The random nested group effect,  $G_{ij}$ , represents the overdispersion in the data, since the  $Var(Y_{ijk}) = \sigma^2 + \sigma_G^2 > \sigma^2$  when  $\sigma_G > 0$  in model (3.2).

If the data are binomial tumor counts rather than quantitative measures, a model analogous to model (3.1) is the binomial model

$$P(Y_{ij} = y | p_i) = \binom{n_{ij}}{y} p_i^y (1 - p_i)^{n_{ij} - y}. \quad (3.3)$$

$Y_{ij}$  is the number of tumors in the  $j$ -th sample or group for the  $i$ -th treatment,  $p_i$  is the probability of a tumor in the  $i$ -th treatment group, and  $Var(Y_{ij}) = n_{ij}p_i(1 - p_i)$ .

The beta-binomial model takes overdispersion into account by allowing the Binomial parameter,  $p_i$ , in model (3.3) to be a random variable,  $P_{ij}$ , that varies from sample group to sample group of animals, receiving the same treatment, according to the Beta Distribution:

$$f(P_{ij} | \alpha_i, \beta_i) = \frac{P_{ij}^{\alpha_i-1}(1 - P_{ij})^{\beta_i-1}}{B(\alpha_i, \beta_i)}, \quad 0 < P_{ij} < 1, \alpha_i > 0, \beta_i > 0, \quad (3.4)$$

where  $\alpha_i$  and  $\beta_i$  are the parameters of the beta distribution, and  $B((\alpha_i, \beta_i) = \Gamma(\alpha_i)\Gamma(\beta_i)/\Gamma(\alpha_i + \beta_i)$ . The marginal distribution of  $Y_{ij}$  is then the beta-binomial distribution:

$$P(Y_{ij} = y) = \binom{n_{ij}}{y} \frac{B(\alpha_i + y, n_{ij} + \beta_i - y)}{B(\alpha_i, \beta_i)}. \quad (3.5)$$

Griffiths (1973) reparameterized the beta-binomial parameters such that

$$P(Y_{ij} = y) = \binom{n_{ij}}{y} \frac{B(\mu_i/\theta_i + y, n_{ij} + (1 - \mu_i)/\theta_i - y)}{B(\mu_i/\theta_i, (1 - \mu_i)/\theta_i)}. \quad (3.6)$$

where  $\mu_i = \alpha_i(\alpha_i + \beta_i)^{-1}$  is the mean of the distribution, or average probability of a tumor in the  $i$ -th treatment, and  $\theta_i = (\alpha_i + \beta_i)^{-1}$  is a reasonable measure of overdispersion because  $1/(\alpha_i + \beta_i + 1)$  is the correlation between the binary variates in the beta-binomial setting.

For the case we will study  $i = 1, 2$  where 1 = control and 2 = treated;  $J_1 = 2$  for dual control groups, and  $J_2 = 1$  for one treatment group. Fixing  $i = 1$ , the beta-binomial model was fit to the two control groups in Table 1, and the maximum likelihood estimates of  $\mu_1$  and  $\theta_1$  were 0.089 and 0.056 respectively.

#### 4. A Likelihood Ratio Test Based on the Beta-Binomial Model

A likelihood ratio test for differences in average proportion of tumors in the treatment and control groups of a  $3 \times 2$  table (in the presence of overdispersion) can be constructed under the beta-binomial model given in equation (3.6). A reasonable test would be to compare the probability of a tumor in the treatment group to the probability of a tumor in the two control groups, assuming the extra-binomial variation or overdispersion is the same for both treatment and control groups (since we have no replicate samples to estimate the overdispersion in the treated group).

The log likelihood function for all the tumor counts in a  $3 \times 2$  table, under the beta-binomial model given in equation (3.6), can be expressed as:

$$L = C + \sum_{i=1}^2 \sum_{j=1}^{J_i} [\ln B(\mu_i/\theta_i + y_{ij}, n_{ij} + (1 - \mu_i)/\theta_i - y_{ij}) - \ln B(\mu_i/\theta_i, (1 - \mu_i)/\theta_i)], \quad (4.1)$$

where  $C$  is a constant,  $J_i = 2$  when  $i = 1$  (control) and  $J_i = 1$  when  $i = 2$  (treatment). The likelihood ratio test statistic for testing the hypothesis  $H_0 : \mu_1 = \mu_2$  versus the alternative  $H_a : \mu_1 \neq \mu_2$  with the restriction that  $\theta_1 = \theta_2$ , is  $S = -2 \times (L_0 - L_a)$  where  $L_0$  is the maximized value of (4.1) under the restriction that  $\mu_1 = \mu_2$  and  $\theta_2 = \theta_1$ , and  $L_a$  is the maximized value of (4.1) without the restriction that  $\mu_1 = \mu_2$ .

Since  $\ln B(\alpha, \beta) = \ln \Gamma(\alpha) - \ln \Gamma(\beta) - \ln \Gamma(\alpha + \beta)$ , the terms of the sum of equation (4.1) can be evaluated using the readily available statistical software such as SAS and Splus and the maximization can be accomplished using the programming languages of these packages, or even using simple spreadsheet programs (see Lawson and Meade (1998)). The statistic  $S$  for the  $3 \times 2$  Table 1 was computed to be  $S = 2.15$ , and the parameters were estimated to be  $\hat{\theta} = 0.0802$ ,  $\hat{\mu}_1 = 0.1382$ .

Even though the test statistic can be easily computed, it is questionable whether the chi-square distribution with one degree of freedom would be appropriate for determining the critical region. In the next section the exact distribution of the likelihood ratio test statistic under the null hypothesis is discussed.

## 5. Significance Limits for Likelihood Ratio Test with Small Sample Sizes

To compute the exact distribution of the beta-binomial likelihood ratio statistic,  $S$ , defined in the last section, we used the following procedure. For control group 1, let  $a$  represent the number of animals with tumors, and  $b$  represent the number of animals without tumors. Similarly for control group 2, let  $c$  and  $d$  represent the number of animals with and without tumors, and for the treated group let  $e$  and  $f$  represent the number of animals with and without tumors. The normal situation for carcinogenicity studies is  $a + b = c + d = e + f = 50$ , or fifty animals per group. Since  $a, c$ , and  $e$  can each range between 0 and 50, there are  $51^3 = 132,651$  potential tables of counts for the domain of our study. The test statistic,  $S$ , was computed for all possible 132,651 tables of counts. This was accomplished using the NLPCG subroutine in SAS proc IML. This subroutine will do a nonlinear optimization of a function by the conjugate gradient method. The values of the test statistic, calculated for each potential table of counts, were then sorted and weighted according to the probability of the occurrence of the

counts in the table (i.e.,  $P(Y_{11} = a) \times P(Y_{12} = c) \times P(Y_{21} = e)$  where  $P(Y_{ij} = y)$  were calculated using equation (3.6)). This was done for the case the of no treatment effect, i.e.,  $\mu_1 = \mu_2$ , and  $\theta_1 = \theta_2$ , and for various overdispersion scenarios determined by the value of  $\theta_1 = \theta_2$ .

The 20 scenarios studied were

$$(\mu = 0.0076, 0.01, 0.14, 0.20) \times (\theta = 0.01, 0.04, 0.06, 0.0711, 0.110).$$

Figure 1 shows the cumulative distribution of the test statistic,  $S$ , for five scenarios and compares them to the cumulative distribution function for the Chi-Square Distribution with one degree of freedom. From the graph it can be seen that the distributions are not close to the chi-square with one degree of freedom. Since the chi-square distribution is to the left, using it to calculate critical limits would result in too many false positives and inflated significance limits.

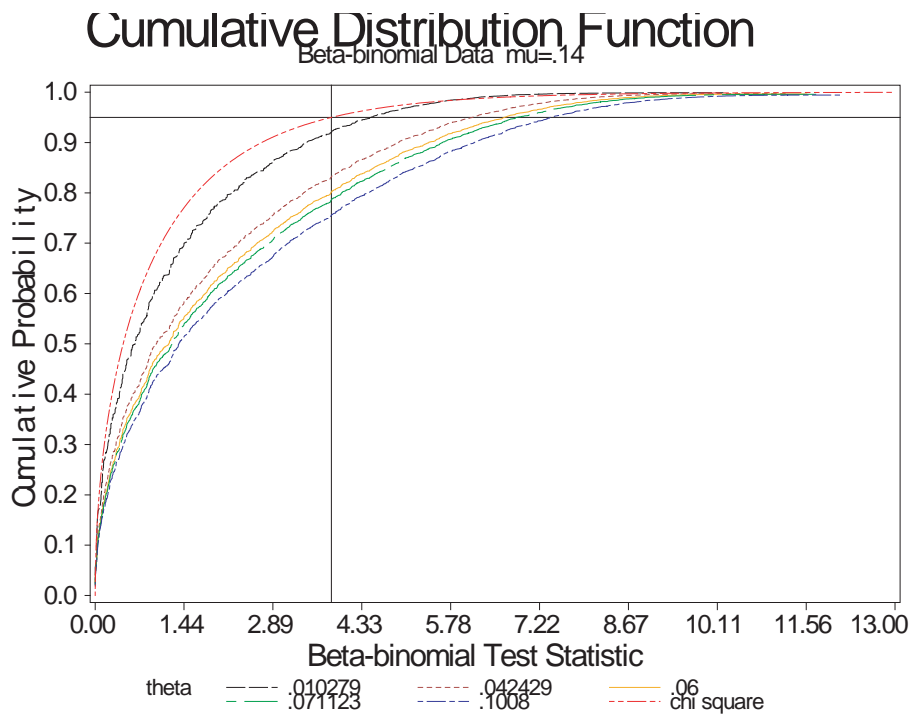


Figure 1: Cumulative distribution – Beta-binomial likelihood ratio statistic

Figure 1 also shows that the likelihood ratio test statistic,  $S$ , is not a pivotal quantity because its distribution (under the null hypothesis) depends upon the unknown population parameters  $\mu$  and  $\theta$ . It can be seen in Figure 1 that as the parameter  $\theta$  increases, the whole cumulative distribution shifts to the right. Because  $\theta$  increases with the variance of the beta-binomial distribution, a larger

value of the test statistic is required to reject the null hypothesis  $H_0 : \mu_1 = \mu_2$  at the same significance level when there is a larger variance.

Table 2: Summary of quadratic curve fits

Statistic	90-th Percentile	95-th Percentile	99-th Percentile
Response Mean	3.9922	5.4740	7.6281
Root MSE	0.1668	0.2080	0.2205
R-Square	0.9920	0.9843	0.9893
Coefficient of Variation	4.18%	3.80%	2.89%

From the cumulative distribution curves generated for the 20 different scenarios, empirical quadratic equations were fit by least-squares so that the extreme percentiles of the distribution can be approximated as a function of the parameters  $\mu$  and  $\theta$ . These are given below as equations (5.1), (5.2), and (5.3). These equations will be useful for predicting the extreme percentiles of  $S$  for any  $2 \times 3$  table containing dichotomous data with one treated and two control groups and with estimated values of  $\mu$  and  $\theta$  within the range of the 20 scenarios studied. The quadratic equations were fit using the desired percentage as the response and the square root of  $\theta$  and  $\mu$  as the independent variables. The fits were excellent, and the summary statistics from the fits are shown in Table 2.

$$\begin{aligned} 90\text{-th percentile} = & 0.49329 + 14.097646\sqrt{\mu} + 5.286806\sqrt{\theta} \\ & - 27.879764\mu + 38.035832\sqrt{\mu\theta} - 20.208964\theta \end{aligned} \quad (5.1)$$

$$\begin{aligned} 95\text{-th percentile} = & 1.791458 + 11.312982\sqrt{\mu} + 8.100139\sqrt{\theta} \\ & - 22.324785\mu + 29.442947\sqrt{\mu\theta} - 15.511246\theta \end{aligned} \quad (5.2)$$

$$\begin{aligned} 99\text{-th percentile} = & 4.988716 + 4.02004\sqrt{\mu} - 1.961369\sqrt{\theta} \\ & - 11.422557\mu + 40.398358\sqrt{\mu\theta} + 8.678507\theta \end{aligned} \quad (5.3)$$

These equations were used to predict the 90-th, 95-th and 99-th percentiles of the distribution of  $S$  for the case where  $\theta = 0.0802$ ,  $\mu = 0.1382$  (the values estimated from Table 1) and resulted in the predicted values 5.762, 7.061, and 9.298. These were very close to the actual percentiles 5.807, 7.084, and 9.265 obtained by enumerating the distribution for this case.

A proposed procedure for using the test statistic,  $S$ , and the significance limits given above is outlined as follows:

- (1) Compute the maximum of the likelihood function  $L_a$  and the maximum likelihood estimates of  $\mu_1$ ,  $\mu_2$ , and  $\theta$  under the alternative hypothesis that  $\mu_1 \neq \mu_2$ .



- (2) Compute the maximum of the likelihood function  $L_0$  and the maximum likelihood estimates of  $\mu_1, \mu_2$ , and  $\theta$  under the null hypothesis that  $\mu_1 = \mu_2$ .
- (3) Compute  $S = -2 \times (L_o - L_a)$ .
- (4) Compare  $S$  to the 90-th, 95-th or 99-th percentile critical limits obtained from equations (5.1), (5.2) or (5.3) using the maximum likelihood estimates of  $\mu_1$  and  $\theta$  obtained in step (1).

Following steps (1) through (4) described above with the data from Table 1, we calculated the estimates of  $\hat{\theta} = 0.0802$ ,  $\hat{\mu}_1 = 0.1382$  and  $S = 2.15$  for comparing the treatment group to the controls. The 90-th percentile of the distribution of  $S$ , approximated from equation (5.1) with these estimates of  $\theta$  and  $\mu$  is 5.762. Since  $S = 2.15 < 5.762$ , it indicates that there is no significant difference between treatment and controls. This is the same conclusion that would have been reached if  $S$  had been compared to the incorrect chi-square reference distribution, but it is different from the conclusion that was reached using Fisher Exact test in Section 2.

## 6. Power Study

In this section we complete our study of the proposed likelihood ratio test statistic,  $S$ , by studying its power properties under various alternatives. This was done by simulation. In the simulation study two control groups were compared to one treatment group in a  $3 \times 2$  table with 50 simulated animals per group.

Generating random data from the beta-binomial distribution is similar to generating data from the binomial distribution. Since a binomial random variable can be expressed as a sum of independent Bernoulli random variables, a beta-binomial random variable can likewise be expressed as a sum of independent Bernoulli random variables where the Bernoulli parameter,  $P_{ij}$ , varies from group  $j$  to  $j$  within treatment  $i$  according to a beta distribution. For the two control groups ( $i = 1, j = 1, 2$ ) and the treated group ( $i = 2, j = 1$ ), 50 independent uniform(0,1) variables  $U_{ij,1}, \dots, U_{ij,50}$  were generated (Ross (1987)). The values of each of these variables represented an individual animal. In addition, a random value of  $P_{ij}$  was generated for each group,  $j$ , from the beta distribution with the values for the parameters  $\mu_i$  and  $\theta_i$  that are constant for each treatment  $i$ . If  $Y_{ijk}$  is the indicator of a tumor for the  $k$ -th animal in the  $j$ -th group for  $i$ -th treatment, then  $Y_{ijk}$  is equal to 1 if  $U_{ijk} < P_{ij}$  and zero otherwise. The simulated count of tumors in the  $j$ -th group for  $i$ -th treatment,  $Y_{ij} = \sum_{k=1}^{60} Y_{ijk}$ , follows the beta-binomial distribution with parameters  $\mu_i$  and  $\theta_i$ .

The beta-binomial likelihood ratio test statistic,  $S$ , was computed for each simulated  $3 \times 2$  table along with the maximum likelihood estimates of the parameters and the computed critical values. For each alternative, this process

was repeated 1000 times and Table 3 shows the power (at the  $\alpha = 0.05$  level of significance) under each alternative. Three values of the control group mean, six values of the treatment group mean, and four overdispersion scenarios are represented in the table. One thousand repetitions reduce the margin of error in the estimated power to less than 0.035.

Table 3: Power of the test\*

Treatment mean ( $\mu$ )	Oversispersion ( $\theta$ )	Control Mean ( $\mu$ )		
		0.01	0.07	0.20
0.01	0.01	0.070	—	—
0.07	0.01	0.522	0.054	—
0.14	0.01	0.842	0.162	—
0.20	0.01	0.956	0.352	0.058
0.40	0.01	0.996	0.893	0.474
0.80	0.01	1.000	0.997	0.996
0.01	0.04	0.047	—	—
0.07	0.04	0.373	0.055	—
0.14	0.04	0.698	0.128	—
0.20	0.04	0.849	0.210	0.040
0.40	0.04	0.963	0.651	0.246
0.80	0.04	0.994	0.918	0.510
0.01	0.08	0.047	—	—
0.07	0.08	0.356	0.064	—
0.14	0.08	0.671	0.138	—
0.20	0.08	0.790	0.166	0.056
0.40	0.08	0.932	0.467	0.154
0.80	0.08	0.970	0.758	0.666
0.01	0.11	0.040	—	—
0.07	0.11	0.356	0.066	—
0.14	0.11	0.595	0.112	—
0.20	0.11	0.758	0.190	0.040
0.40	0.11	0.932	0.443	0.246
0.80	0.11	0.973	0.513	0.510

\* Power values were simulated at the 0.05 level of significance.

It can be seen in the results that the power increases as the mean for the treatment group increases, and decreases as the overdispersion ( $\theta$ ) increases. The significance level, or power when the treatment mean is equal to the control mean, is not inflated and remains near 0.05 when the overdispersion parameter ( $\theta$ ) increases. The power appears reasonable for the alternatives listed.

## 7. Performance of Fisher's Exact Test with Overdispersed Data

Fisher's exact test is one commonly used test statistic for comparing crude proportions. We investigated the type I error rate of this test statistic for comparing one treatment group to dual control groups when extra-binomial variation may be present in the data. To do this, the cumulative distribution was calculated under three different overdispersion scenarios, similar to what was done for the likelihood ratio statistic in section 5. Four Fisher's exact test statistics were calculated for each of the 132,651 possible tables of counts. The first statistic, labeled **cg1** compares the treatment group to control group 1.

$$\mathbf{cg1} = \sum_{i=0}^a \binom{a+b}{i} \binom{e+f}{a+e-i} / \binom{a+b+e+f}{a+e} \quad (7.1)$$

This statistic is a function of the potential values  $a, b, c, d, e$ , and  $f$  in each table of counts as described in Section 5, and represents the probability of  $a$  or fewer tumors in control group 1.

The second statistic, labeled **cg2**, compares the treatment to control group 2, and represents the probability of  $c$  or fewer tumors in control group 2.

$$\mathbf{cg2} = \sum_{i=0}^c \binom{c+d}{i} \binom{e+f}{c+e-i} / \binom{c+d+e+f}{c+e} \quad (7.2)$$

The third statistic, labeled **com**, compares the treatment group to the combined control groups, and represents the probability of  $a + c$  or fewer tumors in the combined control groups.

$$\mathbf{com} = \sum_{i=0}^{a+c} \binom{a+b+c+d}{i} \binom{e+f}{a+c+e-i} / \binom{a+b+c+d+e+f}{a+c+e} \quad (7.3)$$

The fourth statistic, labeled **sma**, compares the treatment to the control group that results in the smallest  $p$ -value (worst case). The cumulative distribution of each Fisher's exact test statistic was calculated under different overdispersion scenarios by weighting the values calculated for each potential table of counts by the joint probability of that table obtained using equation (3.6) under the null hypothesis of no treatment effect, i.e.,  $\mu_1 = \mu_2$ , and  $\theta_1 = \theta_2$ .

For a one sided test of whether the number of tumors in the treated group is large with respect to the control group(s) (or equivalently whether the number of tumors in the control group(s) is small with respect to the treated group), we would reject the null hypothesis when the test statistic is smaller than a critical value. In this case the test statistic represents the probability of observing a

number of tumors less than or equal to what was actually observed. For example, when the chosen level of type I error is  $\alpha = 0.05$ , the critical region for the statistic **cg1** would be  $\{\mathbf{cg1} | \mathbf{cg1} \leq 0.05\}$ .

Table 4: Actual significance levels for 0.05 level Fisher's exact test statistics with beta-binomial data

Beta-binomial parameter		Fisher's exact statistic			
$\mu$	$\theta$	<b>cg1</b>	<b>cg2</b>	<b>com</b>	<b>sma</b>
0.20	0.00	0.029	0.029	0.033	0.053
	0.01	0.060	0.060	0.066	0.104
	0.04	0.134	0.134	0.138	0.216
	0.11	0.221	0.221	0.216	0.333
0.14	0.00	0.028	0.028	0.030	0.051
	0.01	0.057	0.057	0.062	0.100
	0.04	0.127	0.127	0.132	0.204
	0.11	0.210	0.210	0.206	0.314
0.01	0.00	0.024	0.024	0.029	0.044
	0.01	0.051	0.051	0.060	0.089
	0.04	0.118	0.118	0.128	0.187
	0.11	0.208	0.208	0.199	0.311

Table 4 shows the significance levels at critical value 0.05 for binomial data (i.e.  $\theta = 0$ ) and three different overdispersion scenarios for three different values of  $\mu$ . Overdispersion increases with the parameter,  $\theta$ , and the table shows that as the overdispersion increases, the significance levels of all of the test statistics also increase. Therefore it appears that comparing the crude proportion of tumors from one treated group to dual control groups using Fisher's exact test will produce too many false positives if there is overdispersion or extra-binomial variation.

## 8. Summary and Conclusions

We developed approximating equations for the significance limits of a likelihood ratio test statistic for comparing the crude proportion of tumors between a single treated group and dual control groups based on the beta-binomial model. These approximated critical limits do not rely on large sample theory. We show that the significance level of the test statistic we develop does not increase as the level of overdispersion increases. By comparison, we show that Fisher's exact test, which assumes homogeneity of variance, has an inflated type I error

rate when comparing proportions from one treated group to dual control groups where overdispersion is present.

An example of the use of the new statistic was shown with a set of real data where overdispersion appears to be present. In this example no significant difference was found between the treated and control group using the likelihood ratio statistic, but there was a significant difference found using Fisher's Exact test. The difference in conclusion between the likelihood ratio and the Fisher Exact test is due to the inflated significance level of the Fisher Exact statistic.

The likelihood ratio statistic developed in this paper for comparing one treated group to two control groups in the presence of overdispersion could easily be extended to more general comparisons of proportions between treatment groups where overdispersion is present and a small number of replicate samples from one more treatment groups are present. With the speed of modern computers, it would be possible to determine the distribution of such a test statistic by enumerating all possible cases as we did. For example, Williams (1975) developed a large sample test for comparing the number of pups with birth defects in a teratology experiment. If only one or two litters were available for each treatment, Williams' large sample theory would not be accurate. However, the method we used in this paper could be used to develop critical limits for a likelihood ratio test of group differences.

## References

- Gart, J. J., Krewski, P.N., Lee, R. E., Tarone, R. E. and Wahrendorf, J. (1986). *Statistical Methods in Cancer Research: Volume III — The Design and Analysis of Long Term Animal Experiments*. IARC Scientific Publications No. 79, World Health Organization International Agency for Research on Cancer, Oxford University Press.
- Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics* **29**, 637-648.
- Haseman, J. K., Winbush, J. S. and O'Donnell, M. W. Jr. (1986) Use of dual control groups to estimate false positive rates in laboratory animal carcinogenicity studies. *Fundam. Appl. Toxicol.* **7**, 573-584.
- Haseman, J. K., Hajian, G., Crump, K. S., Selwyn, M. R. and Peace, K. E. (1990). Dual control groups in rodent carcinogenicity studies. In *Statistical Issues in Drug Research and Development* (Edited by K.E. Peace), 351-361. Marcell Dekker.
- Haseman, J. K., Huff, J. E., Rao, G. N. and Eustis, S. L. (1989). Sources of variability in rodent carcinogenicity studies. *Fundam. Appl. Toxicol.* **12**, 793-804.
- Haseman, J. K. (1995). Data analysis: Statistical analysis and use of historical control data. *Regulatory Toxicology and Pharmacology* **21**, 52-59.

- Hoel, D. G. (1983). Conditional two sample tests with historical controls. In *Contributions to Statistics: Essays in Honour of Norman L. Johnson* (Edited by P.K. Sen), 229-236. North Holland Publishing Company.
- Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent samples. *Journal of the American Statistical Association* **68**, 46-54.
- 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.
- Lawson, J. and Meade D. (1998). Calculating maximum likelihood estimates of reliability parameters using spreadsheets. *Quality Engineering* **11**, 43-53.
- Pack, S. E.(1986). Hypothesis testing for proportions with overdispersion. *Biometrics* **42**, 967-972.
- Ross, S. (1987). *Introduction to Probability and Statistics for Engineers and Scientists*. Wiley.
- Society for Toxicology (1982). Animal data in hazard evaluation: Paths and pitfalls. *Fundam. Appl. Toxicol.* **2**, 101-107.
- Tarone, R. E. (1981). The use of historical control information in testing for a trend in proportions. *Biometrics* **38**, 215-220.
- Tarone, R. E., Chu, K. C. and Ward, J. M. (1981). Variability in the rates of some common naturally occurring tumors in F344 rats and B6C3F1 mice. *Journal of the National Cancer Institute* **66**, 1175-1181.
- Westfall, P. H. and Soper, K. A. (2001). Using priors to improve multiple animal carcinogenicity tests. *Journal of the American Statistical Association* **96**, 827-834.
- Williams, D. A. (1975). The analysis of binary response data from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**, 949-952.

Received April 19, 2005; accepted June 20, 2005.

John S. Lawson  
Department of Statistics  
Brigham Young University  
Provo, UT 84602, USA  
lawson@byu.edu

Benjamin Ahlstrom  
Bandag Inc.,  
Muscatine, IA 52761, USA  
BAhlstrom@bandag.com