Implementation of a Simple Method for the Analysis of Over-Dispersed Counts and Proportions

Chul Ahn and Jennifer Lee
Department of Biostatistics, City of Hope Medical Center, Duarte, CA

ABSTRACT

Over-dispersed binary and count data occur frequently in many fields of application. Examples include occurrence of cavity in one or more teeth, and development of tumors of a litter. Methods of statistical analysis that ignore over-dispersion underestimate standard errors. Consequently, coverage proportions of confidence intervals and significance levels of tests are distorted. To analyze the over-dispersed data, one could postulate a specific statistical model and use maximum likelihood methods for the estimation of parameters. However, it may be preferable to employ an approach that does not rely on modeling because the true model is hard to know with confidence. Rao and Scott (1992) and Scott and Rao (1995) proposed a simple method for analyzing correlated binary and count data exhibiting over-dispersion relative to a binomial and Poisson model. This method assumes no specific models for over-dispersion. Furthermore, it can be implemented easily using the standard procedures in SAS/STAT software designed to handle independent binary and Poisson counts after a small amount of preprocessing.

INTRODUCTION

Binary or count data often display greater variability than that would be predicted by simply fitting binomial or Poisson models. In such cases this over-dispersion is referred to as extra-binomial or extra-Poisson variation, respectively. Rao and Scott (1992) and Scott and Rao (1995) proposed a simple method for analyzing correlated binary and count data exhibiting over-dispersion relative to a binomial and Poisson model. This method compares groups of over-dispersed binary and count data with group-specific covariates based on the concept of "design effect" widely used in the sample survey literature. This method assumes no specific models for over-dispersion. The paper illustrates its use with examples involving sensitivity of monoclonal antibody and the number of mammary tumors developing in rats.

Although much literature on sample size determination for clinical trials with independent observations is being published, there is a lack of literature or methodology to obtain sample size for correlated binary or count data exhibiting over-dispersion relative to a binomial and Poisson model. This paper presents a simple way to calculate sample size requirement for over-dispersed binary and count data.

STATISTICAL METHODS

(1) Correlated or over-dispersed binary data

The method of Rao and Scott (1992) provides a simple way of comparing over-dispersed binary data among groups. Suppose there are I groups of measurements, where measurements are binary. Let \( x_i \) be the number of affected lesions among the \( n_i \) lesions of the \( j \)th patient (\( j = 1, 2, ..., m_i \)) in the \( i \)th group (\( i = 1, 2, ..., I \)). The natural estimator of \( p_i \) is the overall observed proportion for the \( i \)th group

\[
\hat{p}_i = \frac{x_i}{n_i},
\]

where \( x_i = \sum x_{ij} \) and \( n_i = \sum n_{ij} \) in the \( i \)th group.

Cochran (1977) derived the variance, \( \hat{v}_i \), of \( p_i \) under cluster sampling. For large \( m_i \), the variance estimate is

\[
\hat{v}_i = \frac{m_i \sum (x_{ij} - \hat{p}_i)^2}{(m_i - 1) \hat{p}_i (1 - \hat{p}_i)}
\]

The ratio of \( \hat{v}_i \) to the estimated binomial variance \( \hat{v}_i (1 - \hat{p}_i)/\hat{p}_i \) under the independence assumption, denoted by

\[
d_i = \frac{\hat{v}_i}{\hat{v}_i (1 - \hat{p}_i)}
\]

represents the variance inflation due to clustering. We set \( d_i = 1 \) if \( x_i = 0 \) or \( x_i = n_i \) since \( d_i \) is not defined in these situations. Fung et al. (1994) suggested using \( d_i = \max (d_i, 1) \) since the intraclass correlation is expected to be non-negative. We transform the data \( (x_{ij}, n_{ij}) \) to \( (\tilde{x}_{ij}, \tilde{n}_{ij}) \), where \( \tilde{x}_{ij} = x_{ij}/d_i \) and \( \tilde{n}_{ij} = n_{ij}/d_i \). These transformed data can be analyzed using any standard statistical software for the analysis of independent binary data. This method can be easily applied to stratified analysis. Ahn and Maryon (in press) showed that Mantel-Haenszel estimator based on Rao-Scott transformation performed well for correlated observations in their simulation study.

(2) Over-dispersed count data

Scott and Rao (1995) proposed the method for analyzing count data exhibiting over-dispersion relative to a Poisson model, which is similar to one for the analysis of clustered binary data in Rao and Scott (1992). Suppose there are \( I \) groups of measurements, where measurements are counts. Let \( x_i \) be the number of affected lesions taken over time \( t_i \) in the \( j \)th patient (\( j = 1, 2, ..., m_i \)) in the \( i \)th group (\( i = 1, 2, ..., I \)). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \). Let \( \lambda_i \) be the rate of affected lesion in the \( i \)th group. The natural estimator of \( \lambda_i \) is the overall observed rate for the \( i \)th group \( \hat{\lambda}_i = x_i/t_i \).
is a ratio estimator. The variance of the ratio estimator is given by

\[ \hat{\sigma}_i^2 = \frac{\sum (x_{ij} - \hat{\lambda}_i t_{ij})^2}{(m_i - 1)} \]

The design effect \( d_i \), which represents the estimated inflation in the variance of \( \hat{\lambda}_i \) due to over-dispersion compared to homogeneous Poisson counts, is given by

\[ d_i = \frac{\hat{\sigma}_i^2}{\hat{\lambda}_i^2} \]

We set \( d_i = \max(d_i, 1) \) since the intracluster correlation is expected to be non-negative. Having estimated \( d_i \), the aggregate data \( \{x_{i, t_i}, \hat{\lambda}_i\} \) are transformed to \( \{\bar{x}_{i, \hat{\lambda}_i}, \hat{\tau}_i\} \), where \( \bar{x}_{i, \hat{\lambda}_i} = \frac{x_{i, \hat{\lambda}_i}}{\hat{\lambda}_i} \) and \( \hat{\tau}_i = \frac{t_{i, \hat{\lambda}_i}}{\hat{\lambda}_i} \). The transformation is referred to as "pre-processing" of the aggregate data. These pre-processed data can be analyzed using any standard statistical software for the analysis of independent count data.

**EXAMPLES**

**Example 1 (Correlated binary data)**

Clustered samples of binary data arise frequently in radiology. For example, one to many sites of lesions may be identified using techniques such as CT scans, MRI scans, or monoclonal antibody scans. Often, the fundamental unit for statistical analysis in radiologic studies is the lesion rather than the patient. The results for lesions from the same subject can be highly correlated. Therefore, statistical methods that account for the dependence of observations within a subject are appropriate.

A hypothetical data was constructed based on a current ongoing study, where the patients are infused with an antibody that has been labelled with an isotope [e.g., \(^{111}\)Indium] such that sites of suspicious tissue can be identified by serial nuclear images. All patients underwent surgery following their antibody scan such that pathologic confirmation of all sites of suspicious tissue identified by the antibody scans is obtained. In this data set, the number of lesions per patient ranges from 2 to 9 with a median of 3. The total number of patients is 27 with the total number of lesions = 85 and the total number of affected lesions = 31. Table 1 shows the numbers of lesions and affected lesions.

**Determination of sample size**

It is widely recognized among statisticians that the evaluation of sample size and power is a crucial element in the planning of any research project. We will show how to compute sample size requirements in designs of studies for correlated binary data. Sample size requirements can be similarly computed for over-dispersed count data.

In one-sample problem that yielded a single proportion, the hypothesis \( H_0: \mu = \mu_n \) is tested where one wished to detect a clinically relevant alternative \( H_1: \mu < \mu_n \). As an example, in a monoclonal antibody scan data, one might test that the sensitivity of one antibody (antibody A) equals that obtained in a traditionally used antibody (antibody B). Suppose that the variance obtained from a traditionally used antibody B is twice the variance obtained under independence assumption. That is, the design effect \( d = 2 \) in an antibody B. We will assume that the similar design effect is expected in an antibody A. Often, the fundamental unit for statistical analysis in radiologic studies is the lesion rather than the patient. The results for lesions from
the same subject can be highly correlated.

The sample size requirement under the assumption of independence among lesions is given by (Lachin, 1981)

\[ N = \left[ \frac{Z_a(1-\pi_0) + Z_b(1-\pi_1)}{\pi_1 - \pi_0} \right]^2 \]

It can be easily seen that the sample size needed for correlated binary observations is \( N' = d \times N = 2N \). In general, the sample size needed for correlated binary observations is estimate of design effect \( d \) times the sample size needed under the assumption of independence. In the similar way, we can show that the sample size requirement for over-dispersed count data is \( d \) times the sample size needed under the assumption of homogeneous Poisson distribution.

DISCUSSION

The method we used here is flexible and very easy to implement in SAS® software. Space restrictions preclude the inclusion of our SAS® program which computes design effect and pre-processes the data. It can be obtained through e-mail from the author. Sample size and power of the study can be easily computed once we get an estimate of design effect.

ACKNOWLEDGMENT

SAS® is registered trademark of SAS Institute Inc. in the USA.

REFERENCES


Table 1

<table>
<thead>
<tr>
<th>Affected and Total Lesions Identified by Serial Nuclear Images in 27 Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>the number of affected lesions</td>
</tr>
<tr>
<td>the number of lesions examined</td>
</tr>
<tr>
<td>0 0 3 0 3 2 0 0 0 1 3 2 1 2 0 1 2 0 0 0 0 3 1 0 0 2 5</td>
</tr>
<tr>
<td>3 4 3 4 3 2 2 3 2 3 5 2 4 9 2 3 3 2 2 2 2 3 5</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Number of Mammary Tumors Developing in Female Rats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
</tr>
<tr>
<td>Treatment (n=23)</td>
</tr>
<tr>
<td>Control  (n=25)</td>
</tr>
<tr>
<td>Number of tumors</td>
</tr>
<tr>
<td>1 0 2 1 4 3 6 1 0 1 5 2 1 5 2 3 4 5 5 1 2 6 1</td>
</tr>
<tr>
<td>7 1 1 9 2 9 4 6 7 6 1 3 2 1 1 0 4 5 1 1 1 9 1 2 1 2 1 3 3</td>
</tr>
</tbody>
</table>