Negative binomial estimation and testing : comparison to minimum disparity methods
by Wendy Lee Swanson

a thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics
Montana State University

Abstract:
Various methods have been proposed for comparing the means of independent samples from two negative binomial distributions, but no method is recognized as the standard. The t-test, after log-transforming the data, is often used. But the Mest is unreliable, especially for small means (i.e., one of the means $\mu < 5$). In this dissertation, a new test procedure, called the Disparity Difference Test (DDT) is derived and compared to existing methods. The new method is based on an idea of Lindsay (1994 Annals of Statistics) who introduced a general approach for estimation and testing based on the Negative Exponential Disparity (NED) measure. The DDT is compared to the t-test, the generalized likelihood ratio test, and some generalized score tests. Because all the tests, except the t-test, are asymptotically equivalent, the comparison is based on a simulation study that used small means and realistic sample sizes.

Estimation is embedded in the significance testing methodology because each method requires an estimate of the common negative binomial variance parameter, as well as estimates of the means. A derivation of the NED estimator is provided. The statistical properties of the NED estimator of the variance parameter is compared to the maximum likelihood estimator and to some robust estimators, including the extended quasi-likelihood estimator, the pseudolikelihood estimator, and a conditional maximum likelihood estimator. The comparisons are based on simulation studies.

The results are that the NED estimator performs well, and the DDT not so well, compared to the other methods. There are no practical differences among the empirical average errors for the various estimators. The DDT has smaller power than the likelihood ratio and scores tests for a majority of the parameter settings. There are no practical differences among the score and likelihood ratio tests. Recommendations are provided.

NEGATIVE BINOMIAL ESTIMATION AND TESTING: COMPARISON TO MINIMUM DISPARITY METHODS

by

Wendy Lee Swanson

a thesis submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY-BOZEMAN
Bozeman, Montana

May 1997

APPROVAL

of a thesis submitted by

Wendy Lee Swanson

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Martin A. Hamilton — Martin A. Hamilton (Signature) — 5/19/97 Date

Approved for the Department of Mathematical Sciences

John R. Lund — John R. Lund (Signature) — 5/19/97 Date

Approved for the College of Graduate Studies

Robert L. Brown — R L Brown (Signature) — 6/3/97 Date

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a doctoral degree at Montana State University-Bozeman, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this thesis should be referred to University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature Wendy Lee Swanson

Date May 8, 1997

ACKNOWLEDGMENT

Partial support was provided by Cooperative Agreement No. CR818324 between the USEPA and Montana State University and by Cooperative Agreement No. EEC-8907039 between the National Science Foundation and Montana State University. This dissertation has not been subjected to peer or administrative review by the USEPA and therefore may not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

ABSTRACT

Various methods have been proposed for comparing the means of independent samples from two negative binomial distributions, but no method is recognized as the standard. The t-test, after log-transforming the data, is often used. But the t-test is unreliable, especially for small means (i.e., one of the means $\mu \leq 5$). In this dissertation, a new test procedure, called the Disparity Difference Test (DDT) is derived and compared to existing methods. The new method is based on an idea of Lindsay (1994 *Annals of Statistics*) who introduced a general approach for estimation and testing based on the Negative Exponential Disparity (NED) measure. The DDT is compared to the t-test, the generalized likelihood ratio test, and some generalized score tests. Because all the tests, except the t-test, are asymptotically equivalent, the comparison is based on a simulation study that used small means and realistic sample sizes.

Estimation is embedded in the significance testing methodology because each method requires an estimate of the common negative binomial variance parameter, as well as estimates of the means. A derivation of the NED estimator is provided. The statistical properties of the NED estimator of the variance parameter is compared to the maximum likelihood estimator and to some robust estimators, including the extended quasi-likelihood estimator, the pseudolikelihood estimator, and a conditional maximum likelihood estimator. The comparisons are based on simulation studies.

The results are that the NED estimator performs well, and the DDT not so well, compared to the other methods. There are no practical differences among the empirical average errors for the various estimators. The DDT has smaller power than the likelihood ratio and scores tests for a majority of the parameter settings. There are no practical differences among the score and likelihood ratio tests. Recommendations are provided.

CHAPTER 1

INTRODUCTION

Background

The negative binomial distribution has provided a representation for count data in many areas of research. As noted by Bliss and Fisher (1953), the earliest fit (empirical) of count data to a negative binomial was applied to microscopic counts of yeast cells by Student (1907). Some of the early uses of the negative binomial model in statistical analyses were on counts of insects (Anscombe, 1949), microbes (Jones, Mollison and Quenouille, 1948) and accidents (Greenwood and Yule, 1920). In more recent years, negative binomial analysis of count data has been applied by researchers in a variety of disciplines. It has been applied to market research (Chatfield, 1975), purchasing (Schmittlein et al., 1985; Ramaswamy et al,1994) and reliability (Bain and Wright, 1982). Jones et al (1991) analyzed negative binomial models for counts of parasites on a host species; Hubbard and Allan (1991) used a sequential negative binomial analysis for an insect pest management strategy, and Morton (1987) used a negative binomial model to analyze insect trap catches in a nested block design. Gold et al (1996) compare sampling protocols for weeds clustered in fields in a spatial distribution that can be described by the negative binomial. Manton et. al. (1981) applied hierarchical negative binomial models to an epidemiological study of lung cancer mortality rates; Maul, El-Shaarawi and Ferard (1991) analyzed a chronic toxicity response using a negative binomial model.

In ecology, the NB is used in the analysis of a pollution impact study on fish abundance (Ramakrishnan and Meeter, 1993). Counts of species richness-area relationship were analyzed using competing models based on the negative binomial distribution (Stein, 1988). Seber (1973) applied negative binomial models arising from several different sampling strategies to the estimation of animal abundance.

Barnwal and Paul (1988) derived two C(α) statistics (Neyman, 1959) for testing equality of means in a one-way layout for negative binomial data and applied them to field and laboratory research counts. Collings and Margolin (1985) compared tests for departure from the Poisson assumption using a negative binomial alternative for data in a one-way layout.

Microbiological applications included analyses of the counts of revertant colonies in the Ames salmonella microsome assay (Margolin et al, 1981; Breslow, 1984; Krewski et al, 1993), spatial and temporal variation of bacterial counts (Maul and El-Shaarawi, 1991) and the estimation of coliform density in drinking water (Pipes et al. 1977).

There are a large number of chance mechanisms which give rise to the negative binomial and have plausible physical applications. Some of them are described below with an outline of their derivation.

A common representation of the negative binomial random variable, Y, uses parameters p, $(0 < p < 1)$ and $k > 0$, denoted NB(p,k) and having discrete density or probability mass function (pmf) $P(Y = y) = \binom{y+k-1}{y} p^k (1-p)^y$ $y = 0,1,\ldots$. This representation has associated moment generating function

(mgf) $M_Y(t) = E\left(e^{tY}\right) = \left(\frac{p}{1-(1-p)e^t}\right)^k$ and probability generating function (pgf) $G_Y(s) = E\left(s^Y\right) = \left(\frac{p}{1-(1-p)s}\right)^k$.

In the above representation k is not limited to the integers. When k is limited to the integers, the distribution is sometimes called the Pascal distribution [Pascal (1679)]. In this spirit, Ross (1989) motivated the NB distribution as "a coin having probability p of coming up heads (being) successively flipped until the kth head appears". This example of the NB derivation as the distribution of the number of tosses of a coin required to achieve a fixed number of successes is due to Montmort (1714). The presentation of NB for integer k is common in statistics textbooks. A geometric random variable is the special case of a negative binomial random variable for $k=1$. One way to generate the NB with integer k is as the sum of n independent and identically distributed (iid) geometric random variables with parameter $p, (0<p<1)$ and mgf $M_{X_i}(t) = \frac{p}{1-qe^t}$. It follows from independence of the X_i's that the mgf of their sum, $Y = X_1 + \cdots + X_n$, is the product of their mgfs, i.e.: $M_Y(t) = M_{X_1+\ldots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(\frac{p}{1-qe^t}\right)^n$ which is the mgf of the $NB(p, k=n)$. This model was used by Seber (1973, p.174) for recapture data from multiple traps (acting independently) within an animal's home range.

The negative binomial can be derived as a Poisson-stopped sum of logarithmic random variables (Luders, 1934; see also Quenouille, 1949). Let $Y = X_1 + \cdots + X_N$, where the X_i's are iid logarithmic random variables with parameter $\eta > 0$, and have pmf, $P(X=x) = \frac{-\eta^x}{x\ln(1-\eta)}$, $x = 1,2,\ldots$, and probability generating function (pgf), $G_{X_i}(s) = \ln(1-\eta s)/\ln(1-\eta)$. Let N be independently

distributed Poisson with parameter $\lambda > 0$ and pgf $G_N(s) = \exp(\lambda(s-1))$. Then the pgf of the random sum $Y = X_1 + \cdots + X_N$ is $G_Y(s) = G_N\left(G_{X_i}(s)\right)$

$$= \exp\left(\lambda\left(\frac{\ln(1-\eta s)}{\ln(1-\eta)} - 1\right)\right) = \exp\left(\ln\left(\frac{1-\eta s}{1-\eta}\right)^{\lambda/\ln(1-\lambda)}\right) = \left[(1-\eta)/(1-\eta s)\right]^{-\lambda/\ln(1-\lambda)}$$ which is

the NB pgf with parameters $p = 1-\eta$ and $k = -\lambda/\ln(1-\lambda)$. Quenouille's derivation of this model stemmed from research counts of soil microbes (Jones, Mollison and Quenouille, 1948) in which the "colony counts followed a Poisson's distribution, the numbers of bacteria per colony were logarithmically distributed, and that, consequently, the bacterial counts were distributed in the negative binomial form."

Another derivation of the negative binomial (due to Greenwood and Yule, 1920) is as a gamma mixture of Poisson random variables. Suppose X, given λ, is distributed Poisson(λ) with (conditional) pmf $P(X = x|\lambda) = e^{-\lambda}\lambda^x/x!$, for $\lambda > 0, x = 0,1,2,\ldots$; and that independently, λ is distributed as gamma(ν,τ) with probability density function (pdf) $f(\lambda;\nu,\tau) = \left(\lambda^{\nu-1}e^{-\lambda/\tau}\right)/\left(\Gamma(\nu)\tau^{\nu}\right)$ for $\lambda,\nu,\tau > 0$. For the gamma distribution ν is referred to as the index or *shape parameter* and τ is the *scale parameter*. Then unconditionally, X is distributed as NB with parameters $p = 1/(\tau+1)$ and $k = \nu$. That is,

$$P(X = x) = \int_0^{\infty} P(X = x|\lambda) f(\lambda;\nu,\tau)\,d\lambda = \left\{\Gamma(\nu)\tau^{\nu}x!\right\}^{-1} \times \int_0^{\infty} e^{-\lambda}\lambda^x\lambda^{\nu-1}e^{-\lambda/\tau}\,d\lambda$$

$$= \frac{\Gamma(\nu+x)\left(\tau/(\tau+1)\right)^{\nu+x}}{\Gamma(\nu)\tau^{\nu}x!} \times \int_0^{\infty} \frac{e^{-\lambda/(\tau/(\tau+1))}\lambda^{(\nu+x)-1}\,d\lambda}{\Gamma(\nu+x)\left(\tau/(\tau+1)\right)^{\nu+x}} = \frac{\Gamma(\nu+x)}{\Gamma(\nu)x!}\left(\frac{\tau}{\tau+1}\right)^x\left(\frac{1}{\tau+1}\right)^{\nu} \times 1,$$ the NB

pmf with mean, $E(X) = \nu\tau = \mu$ and variance $Var(X) = \nu\tau + \nu\tau^2 = \mu + \frac{\mu^2}{\nu} = \mu(1+\tau)$.

Greenwood and Yule (1920; see also Arbous and Kerrich, 1951) used this derivation as a model for accident statistics in which individuals differed in accident-proneness.

Boswell and Patil (1970) give a list of processes which produce the NB distribution. Included are population growth models and other stochastic processes and their derivations. Johnson, Kotz, and Kemp (1992, chapter 5) provide an excellent review of the literature on the negative binomial distribution and a large list of references. They listed many parameterizations for the negative binomial. The forms that I find most useful are denoted NB(μ,k) [Anscombe (1950)] and NB(μ,a) [Bliss and Owen (1958)]. The Anscombe parameterization pmf is $\Pr(Y=y;\mu,k)=\frac{\Gamma(y+k)}{y!\Gamma(k)}\left(\frac{\mu}{\mu+k}\right)^{y}\left(\frac{k}{\mu+k}\right)^{k}$ $y=0,1,\ldots.$ Bliss and Owen used $a=k^{-1}$ and $\Pr(Y=y;\mu,a)=\frac{\Gamma\left(y+a^{-1}\right)}{y!\,\Gamma\left(a^{-1}\right)}\left(\frac{a\mu}{1+a\mu}\right)^{y}\left(\frac{1}{1+a\mu}\right)^{a^{-1}}$ $y=0,1,\ldots.$ Both these forms parameterize using the mean, $E(Y)=\mu$ and an additional parameter in the variance, $\mathrm{var}(Y)=V=\mu+\frac{\mu^2}{k}=\mu+a\mu^2$. Recent literature has made use of the NB(μ,a) because it yields the Poisson distribution as the limiting case when a goes to 0. [Lawless (1987), Piegorsch (1990)]. Clark and Perry (1989) suggest the use of a because it "eliminates the problems of infinite values of $\hat{k}=\frac{\bar{y}^2}{S^2-\bar{y}}$ (method-of-moments estimator) when $S^2=\bar{y}$ (where S^2 is the sample variance); also confidence intervals for a are continuous and usually more symmetric than those for k, which may be discontinuous."

The negative binomial is often used for empirical reasons. It has provided an adequate fit to count data. Martin and Katti (1965) fit various "contagious" distributions (distributions produced as mixtures of other distributions) to 35 data sets and found the negative binomial and Neyman Type A to have wide applicability. Bliss and Fisher (1953) used the NB to fit numerous biological data sets. Evans (1953) found the NB provided the only

satisfactory fit to counts of insect populations he examined. Christian and Pipes (1983) found the NB distribution compatible to the coliform frequencies in nine drinking water systems. In each case, the data could be considered as "overdispersed" relative to a Poisson model (variability larger than expected, or variance larger than the mean). In a Poisson model, the distribution of numbers of individuals per unit space or time is random, the presence of an individual not influencing the probability of the presence of another. Bliss (Bliss and Fisher, 1953) noted that "when the numbers of individuals per unit space or time cannot be assumed to have the same expected value, they may represent a mixture of several homogeneous Poisson distributions" in which the means are distributed as a positive continuous variate. This produces heterogeneity or density-dependence rather than the random spatial distribution associated with the Poisson. Animals that occur in herds display this tendency. So do microbes and insects which tend to aggregate or clump. Hubbard and Allen (1991) reasoned that insect distributions are heterogeneous because they "migrate to hospitable environments, or if the insects migrate little, (because) their eggs are laid in clumps." Taylor, Woiwod, and Perry (1978) reported on the rarity of randomness of spatial behavior in nature. Among the possible distributions that can fit heterogeneity, the negative binomial provides the simplest Poisson mixture. And because the gamma mixing variable can take on a wide variety of shapes, the resulting NB distributions fit many data sets.

Some analysts worked with data that fit only the form of the mean and variance of the NB, i.e., $E(Y)=\mu$ and $\mathrm{var}(Y)=\mu+a\mu^2$. Minkin (1991) arrives at this form using a random effects Poisson model. A multiplicative random effect was introduced to accommodate the extra-Poisson variability in data sets involving medical cancer clonogenic assays and the analysis of dose-response

curves. Chase and Hoel (1975) examine the mechanism of serial dilutions used to estimate particle concentration and show how small measurement errors of individual dilutions can produce a mean and variance with the same form as seen in the NB.

I became interested in plate counts of microbes while working in the Center for Biofilm Engineering, at Montana State University, on EPA-sponsored research concerning the efficacy of disinfectants. The counts appeared overdispersed relative to a Poisson model and the results of interest were tests of control versus treatment. My literature search lead me to believe that the NB distribution represented a reasonable model for this type of data for both empirical and methodological reasons. I wanted to compare the power of tests for data of NB form framed in a generalized linear model (GLM) setting with the specific purpose of testing control versus treatment. I was attracted to the characterization of the NB as a gamma mixture of Poissons. I followed the setting of the mixed Poisson (NB) regression model as presented in Lawless (1987). I was interested in the estimation of the variance parameter because it is an open, important problem. The two-parameter NB is not exponential family, the minimal sufficient statistic (order statistic) is not complete, and plots of the log-likelihood function illuminate the difficulty of the estimation problem for the variance parameter (Willson, Folks and Young (1986)). Many robust estimators have been proposed in recent literature. I wanted to compare the newest estimation techniques to the some of the most frequently used older techniques. My focus was two-fold: estimation of the variance parameter and testing for a difference between the means of two populations.

Notation and Terminology

Matrices shall be denoted by upper case letters (X), elements or scalars (in italics, x_{ij} and β_2) and vectors (not italic, bold for non-greek font, $\mathbf{x}_i$ and β) in lower case letters.

The extension of the NB(μ,a) to a log-linear regression model, or (generalized) linear model (GLM), denoted NB$(\mu(\mathbf{x}),a)$ is given in Lawless (1987). Given that the Y_i's are n independent negative binomial random variables with a common variance parameter, a, and the $\mathbf{x}_i$'s are px1 explanatory variables, the probability that Y_i equals y_i given $\mathbf{x}_i$ is

$$\Pr(Y_i = y_i|\mathbf{x}_i) = \frac{\Gamma(y_i + a^{-1})}{y_i!\,\Gamma(a^{-1})}\left(\frac{a\mu_i}{1+a\mu_i}\right)^{y_i}\left(\frac{1}{1+a\mu_i}\right)^{a^{-1}}, \; y_i = 0,1,\ldots \text{ where}$$

$E(Y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i) = e^{\mathbf{x}_i^t\beta}$, abbreviated μ_i, and $Var(Y_i) = \mu_i + a\mu_i^2$, abbreviated V_i. The link function, or link between the mean and the explanatory variables is $\log(\mu) = \mathbf{x}^t\beta$, hence the term log-linear model. I will refer to a as the *variance parameter*. Note that the variance parameter does not depend on i. Other authors refer to a (or to k=a^{-1}) as an index parameter, shape parameter or dispersion parameter. [I do not refer to a (or to k) as a dispersion parameter because it does not qualify as such in the way "dispersion parameter" is defined and used in some of the methods described below.]

The simplest way to adapt this generalized linear model setting to a negative binomial model for testing treatment versus control is to use a design matrix consisting of columns for an intercept and a control versus treatment contrast. Then each $\mathbf{x}_i^t$ is a row of a design matrix, X, and $\mathbf{x}_i$ is the transpose of the ith row. Just as in linear models for the one-way ANOVA setting, if we have n_1 observations from the control group (population one), and n_2

observations from the treatment group (population two), and $n=n_1+n_2$ then the design matrix consists of a column of 1's of length n and a column of 1's and -1's of lengths n_1 and n_2 , i.e.: $X=\begin{bmatrix}X_1\\X_2\end{bmatrix}\begin{matrix}n_1\times 2\\n_2\times 2\end{matrix}=\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}\begin{matrix}n_1\times 2\\n_2\times 2\end{matrix}$. The means for control and treatment are $\mu_1=e^{\beta_1+\beta_2}$ and $\mu_2=e^{\beta_1-\beta_2}$, or $\log(\mu_1)=\beta_1+\beta_2$ and $\log(\mu_2)=\beta_1-\beta_2$, respectively, where $\beta=\begin{bmatrix}\beta_1\\\beta_2\end{bmatrix}$ and β_1 is the intercept and β_2 is the slope of the contrast between treatment and control.

The objective is to test whether or not the treatment is effective. The interest is in a one-sided test. In the disinfectant setting, the disinfected object (treatment) should be at least as effective as the control (no treatment) in reducing the counts of microbes. The counts for the treatment group (Y_2) should be at least as small as those of control (Y_1). Thus, I would want to test the null hypothesis (H_0) that the treatment has no effect versus the alternative (H_1) that it reduces the counts of colonies. These hypotheses can be stated in either of two equivalent ways,

$$H_0: \beta_2 = 0 \text{ vs } H_1: \beta_2 > 0 \text{ , or}$$

$$H_0: \mu_1 = \mu_2 = \mu \text{ vs } H_1: \mu_1 > \mu_2.$$

For results reported to a regulatory agency, often test results and point estimates (rather than interval estimates or confidence intervals) are all that is required. If interval estimates are desired, they are obtainable by inverting the test statistics.

I studied a variety of test statistics and a variety of methods to estimate the variance parameter a. One test statistic was based on a generalization of Rao's score test which accommodates estimating equations from methods other than maximum likelihood. One advantage of (generalized) score tests is that parameters need only be estimated under the null hypothesis. The estimation

methods used in conjunction with generalized score tests were Extended Quasi-likelihood (EQL), Pseudolikelihood (PL), Conditional Maximum Likelihood (CML), Optimal Quadratic (OQ) and Maximum Likelihood (ML). The powers of these tests was compared to the power of the most commonly used test based on the complete distribution, the (generalized) Likelihood Ratio Test. I also studied a new type of estimation and testing, minimum Disparity estimation and a Disparity Difference Test (DDT), based on the Negative Exponential Disparity (NED) measure introduced by Lindsay (1994). Because it is easily applied and conventionally used by many researchers, a t-test (t) was performed on log-transformed NB data. I used a version of the t-test that allows for unequal variances, namely Welch's modified t-test.

CHAPTER 2

ESTIMATION METHODS

I shall explain the methods of estimation in general terms, then the resultant forms of the estimating equations applied to a NB GLM setting (p. 8) where possible. The adaptations of the equations and estimators to a one-way layout will follow.

In my reading, I noted that a number of authors examine the form or properties of the estimating equations themselves rather than those of the estimators [Godambe (1960), Godambe and Heyde (1987), Godambe and Thompson (1987, 1989), Anraku and Yanagimoto (1990), McCullaugh and Nelder (1989, see sections 9.4 and 9.5)]. They develop theory based on the distributions of the estimating functions rather than distributions of the estimators. Citing work by Boos (1980) and others, Godambe and Thompson (1989, p. 171) give examples where use of the "estimating function" improves accuracy of confidence intervals over those based on the corresponding "estimate". An *estimating equation* $g(\mathbf{y},\theta)$ is any function of the data and parameters having $E_\theta[g(\mathbf{y},\theta)]=0$ for all θ (Godambe and Thompson, 1984). Provided there are as many equations as parameters, the estimates $\tilde{\theta}_i$, are obtained by solving the vector equation $g(\mathbf{y},\theta)=0$ for θ.

Maximum Likelihood (ML) Estimation

Maximum likelihood is a long-used and familiar technique for estimation. The full distribution is assumed. When that distribution is correct, maximum likelihood estimators often provide the standard for efficiency against which robust estimators are compared. Maximum likelihood estimation for the NB is discussed by Fisher (1941). Its use in estimation for the variance parameter, $a = 1/k$ is discussed in Piegorsch (1990). Its application to the negative binomial in a generalized linear regression setting is summarized in Lawless (1987). The object of the estimation procedure is to maximize the likelihood, or equivalently the log-likelihood, for the given distribution.

For a sample from the general NB log-linear regression model, the likelihood is proportional to $L(\beta,a)=\prod_{i=1}^{n}\frac{\Gamma\left(y_i+a^{-1}\right)}{\Gamma\left(a^{-1}\right)}\left(\frac{a\mu_i}{1+a\mu_i}\right)^{y_i}\left(\frac{1}{1+a\mu_i}\right)^{a^{-1}}$, where $\mu_i=\mu(\mathbf{x}_i)=e^{\mathbf{x}_i^t\beta}$. Note that if y is an integer ≥ 1, then $\forall c>0$, $\frac{\Gamma(y+c)}{\Gamma(c)}=c(c+1)(c+2)\cdots(c+y-1)$; the gamma ratio equals one for $y=0$. It follows that $\frac{\Gamma\left(y_i+a^{-1}\right)}{\Gamma\left(a^{-1}\right)}=\prod_{j=0}^{y_i-1}a^{-1}(1+aj)=\left(\frac{1}{a^{y_i}}\prod_{j=0}^{y_i-1}(1+aj)\right)$, and thus $L(\beta,a)=\prod_{i=1}^{n}\left(\prod_{j=0}^{y_i-1}(1+aj)\right)\left(\frac{\mu_i}{1+a\mu_i}\right)^{y_i}\left(\frac{1}{1+a\mu_i}\right)^{a^{-1}}$. The log-likelihood is

$$l(\beta,a)=\log\left(L(\beta,a)\right)=\sum_{i=1}^{n}\left[\left(\sum_{j=0}^{y_i-1}\log(1+aj)\right)+y_i\log\mu_i-\left(y_i-a^{-1}\right)\log(1+a\mu_i)\right].$$

The ML estimating equations for the mean and variance parameters are:

$\frac{\partial l}{\partial\beta_s}=\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{(1+a\mu_i)}x_{is}=0$ for $s=1,\ldots,p$ and

$$\frac{\partial l}{\partial a}=\sum_{i=1}^{n}\left[\left(\sum_{j=0}^{y_i-1}\frac{j}{(1+aj)}\right)+a^{-2}\log(1+a\mu_i)-\frac{\left(y_i+a^{-1}\right)\mu_i}{(1+a\mu_i)}\right]=0.$$

Adapting the generalized linear model to a control versus treatment setting, we have $\mu(\mathbf{x}_i) = \mu_1 = e^{\beta_1+\beta_2}$ for $i = 1,\ldots,n_1$ and $\mu(\mathbf{x}_i) = \mu_2 = e^{\beta_1-\beta_2}$ for $i = n_1+1,\ldots,n$. The ML estimating equations for the means become $\frac{\partial l}{\partial \beta_1} = n_1\frac{(\bar{y}_1-\mu_1)}{(1+a\mu_1)} + n_2\frac{(\bar{y}_2-\mu_2)}{(1+a\mu_2)} = 0 = U_1(\beta_1,\beta_2;a)$, say, and $\frac{\partial l}{\partial \beta_2} = n_1\frac{(\bar{y}_1-\mu_1)}{(1+a\mu_1)} - n_2\frac{(\bar{y}_2-\mu_2)}{(1+a\mu_2)} = 0 = U_2(\beta_1,\beta_2;a)$. The ML estimates for the mean are $(\tilde{\beta}_1,\tilde{\beta}_2) = \left(\log\left[(\bar{y}_1\bar{y}_2)^{1/2}\right],\log\left[(\bar{y}_1/\bar{y}_2)^{1/2}\right]\right)$ or $(\tilde{\mu}_1,\tilde{\mu}_2) = (\bar{y}_1,\bar{y}_2)$ under H_1 and $\tilde{\mu} = \bar{y}$ under H_0.

The variance parameter estimate is obtained by inserting the mean estimates into $\frac{\partial l}{\partial a} = 0$ and solving for the root of the equation. A number of methods are applicable, including gradient methods, the scoring algorithm or Newton Raphson method (Walsh, 1975). For solving for the ML estimator of a, and *for all other methods and estimators requiring numerical techniques to obtain roots, I used the Newton Raphson* (NR) *method.* NR was used because it has good convergence properties when starting with good initial values (Walsh, 1975, chapter 4). Initial values were obtained using a grid search.

The above use of the GLM setting provides a model for obtaining an estimate for a common variance parameter in a one-way layout. An alternative method outside the GLM structure in given in Bliss and Owen (1958). In settings where the design matrix is not of the simplified ANOVA form, the mean estimates may not be obtainable in closed form and numerical techniques are required to calculate estimates of the $\tilde{\beta}$'s. Lawless (1987) suggests a method of profile likelihood to find $\tilde{a}$ in the general setting.

Extended Quasi-likelihood (EQL) estimation

Quasi-likelihood (QL) is a robust form of estimation used in generalized linear models (see McCullaugh and Nelder, 1989, chapter 9; first use of the term "quasi-likelihood" was by Wedderburn, 1974). The user specifies only the mean and variance structure rather than the complete form of the likelihood. The general assumptions are in agreement with the model as stated above. Namely, the components of the response vector $\mathbf{Y}$ are independent with mean vector μ and covariance matrix $\sigma^2\mathbf{V}(\mu)$. The mean vector is a known function of β and covariates $\mathbf{x}$. The covariance matrix is the product of σ^2, a scalar dispersion parameter which does not depend on β, and $\mathbf{V}(\mu)$, a diagonal matrix of known functions where the i^{th} diagonal element V_i depends on the i^{th} element of μ (rather than several components of μ). For the NB as defined above, σ^2=1 and $V_i = \mu_i + a\mu_i^2$.

The integral $Q(\mu;y) = \int_y^{\mu} \frac{y-t}{\sigma^2 V(t)} dt$, if it exists, is called the quasi-likelihood, or more correctly, the log quasi-likelihood for μ based on data y. When $Q(\mu;y)$ exists, the estimates for the mean parameters are the solutions to equations obtained by differentiating $Q(\mu;y)$ and are written in the general form $U(\tilde{\beta}) = 0$. The QL estimating equations, $U(\tilde{\beta}) = 0$, are also referred to as the quasi-score functions and have the same form whether or not $Q(\mu;y)$ exists. They can be written as $\underset{px1}{U}(\beta) = D^t V^{-1}(\mathbf{y} - \mu)/\sigma^2$, where D is the matrix of derivatives of the mean with respect to β, $D_{nxp} = \frac{\partial \mu}{\partial \beta^t} = \{D_{ij}\} = \left\{\frac{\partial \mu_i}{\partial \beta_j}\right\}$. Because the components of the response vector are independent, we can express the quasi-score function

for the complete data set as a sum of the individual contributions:

$$\underset{p\,x\,1}{U}(\beta)=\sum_{i=1}^{n}\mathbf{u}_i=\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{V_i}\frac{\partial\mu_i}{\partial\beta_{p\,x\,1}}\Big/\sigma^2.$$

For the NB, $D_{nxp}=\dfrac{\partial\mu_{nx1}}{\partial\beta^{t}_{1xp}}=\left\{\dfrac{\partial\left(e^{\mathbf{x}_i^t\beta}\right)}{\partial\beta_j}\right\}=\left\{x_{ij}e^{\mathbf{x}_i^t\beta}\right\}=\left\{x_{ij}\mu_i\right\}=\underset{nxn}{\left\{diag(\mu_i)\right\}}X_{nxp}$,

where $\underset{nxp}{X}=\begin{bmatrix}\mathbf{x}_1^t\\ \vdots\\ \mathbf{x}_n^t\end{bmatrix}$ and each $\underset{px1}{\mathbf{x}_i}$, the covariates for the i^{th} response, is the transpose of a row of the design matrix, X. The resultant QL estimating equation for the mean parameters is the same as that obtained by maximum likelihood. In the QL estimating equation for the mean, as in the ML estimating equation for the mean, the value of a is treated as fixed while solving for the mean parameters. An additional equation is needed to solve for a.

This method was "extended" to enable estimation of parameters in the variance function by adding a term to the estimating equation (introduced in Nelder and Pregibon, 1987). The Extended Quasi-likelihood equation can be written as: $Q^{+}=Q-\frac{1}{2}\log\left\{2\pi\sigma^2V(y)\right\}$, where $V(y)$ is the variance expressed as a function of y instead of μ, a sort of "empirical" variance. Thus, addition of this term does not change the estimates of the mean parameters. For the NB, $V(y)=y(1+ay)$ and σ^2=1.

For many distributions, the EQL equation is related to the log-likelihood equations. When using the NB mean and variance structure, Q^{+} is directly obtainable from the NB log-likelihood. The only difference in the equation is that the factorials $z!$ in the NB likelihood are replaced by Sterling's approximation, $z!\approx(2\pi z)^{\frac{1}{2}}z^ze^{-z}$. Because the Sterling approximation fails for z=0, an amended form of the approximation, $z!\approx\left\{2\pi(z+c)\right\}^{\frac{1}{2}}z^ze^{-z}$ where c>0, is used for discrete variables with zero in their support space. Nelder and

Pregibon (1987) suggest use of $c=1/6$. (See the Q^+ log-likelihood relationship for the NB in Appendix A).

An "adjustment for degrees of freedom" used to estimate variance or dispersion parameters is recommended by McCullaugh and Nelder (1989, p. 362; see also Nelder and Lee, 1992, p. 281). They suggest multiplying the term of Q^+ containing the empirical variance by $\frac{n-p}{n}$ to account for the fact that p parameters have been fitted to the means.

Adapting EQL to the NB null hypothesis model, the estimate for the mean using EQL is, again, $\tilde{\mu}=\bar{y}$, and p=1 parameter was fit to the mean. The estimate for the NB variance parameter is the solution to $\frac{\partial Q^+}{\partial a}=0=\tilde{U}(a)_{EQL}$, using $\tilde{\mu}=\bar{y}$, which reduces to solving for $\tilde{a}=\tilde{a}_{EQL}$ in $\sum_1^n\left[\frac{1}{\tilde{a}^2}\log\left(\frac{1+\tilde{a}\bar{y}}{1+\tilde{a}y_i}\right)-\frac{(n-1)}{n}\frac{y_i}{1+\tilde{a}y_i}+\frac{(n-1)}{n}\frac{(1+6y_i)}{2(\tilde{a}+6+6\tilde{a}y_i)}\right]=\frac{n-1}{2(\tilde{a}+6)}$. This is the equation listed in Clark and Perry (1989), adjusted for degrees of freedom. (See simplifications in Appendix A).

The above version of EQL was outlined in McCullaugh and Nelder (1983, pp. 212-214), but not present in the second edition (1989; pp. 373-374, 349-350, 360-362). In the later addition for the given model they suggest estimation of a via setting the mean deviance equal to unity, a type of "method of moments" approach. Their general suggestion in the later edition is to view any overdispersion as stemming from a dispersion parameter. Thus using $Var(Y)=\sigma^2V(\mu)$ or more generally $Var(Y_i)=\sigma_i^2V(\mu_i)$ where $\sigma^2\neq1$ and $V(\mu)$ does not contain any additional parameters beyond those found in the mean. This type of analysis could accommodate the NB modeled as a Poisson mixed by a gamma distribution in which the scale parameter (τ), rather than the shape parameter ($\nu=a^{-1}$), is held constant (McCullaugh and Nelder, 1989, p. 199 and

problem 9.1, p. 352; Nelder and Lee, 1992, p. 277). This results in an NB with the variance function as a linear function of the mean and a dispersion parameter, i.e., $Var(Y) = \sigma^2 V(\mu)$ where $V(\mu) = \mu$ and $\sigma^2 = (1+\tau)$, a function of the gamma scale parameter. Nelder and Lee (1992) point out that this model does not belong to the GLM family of distributions, i.e., one that can be written in exponential family form for fixed τ, and show that the Quasi-likelihood estimating equations for the mean will be different than those based on ML for this model. The NB parameterized with a is GLM family for fixed (but unknown)a, with canonical link of $\mathbf{x}^t\beta = \eta = \log\left(\frac{a\mu}{1+a\mu}\right)$ (McCullaugh and Nelder, 1989, p. 373).

I chose to follow the more common approach and assume that the shape parameter remains constant and that the mean varies with the scale of the gamma. In this setting, the mean estimates for the NB model coincide with those obtained by using a Poisson model. This choice is advantageous if one wants to consider the Poisson as a limiting case of the NB. There are differences in the both the Gammas and resultant NBs when one varies either the shape or scale parameters of the Gammas (Figure 1, p. 126). The difference between these choices is more evident in the gammas. Holding the shape fixed while increasing the scale parameter (Figure 1C) is similar to grabbing the right hand tail of the gamma distribution and stretching it out. The relative probability of values near zero is more closely preserved. Holding the scale fixed while increasing the shape parameter (Figure 1A) results in a change of symmetry and density values near zero. In this figure, the NBs produced by the gammas are plotted directly below the gammas. The NB shapes mimic the gamma shapes to a lesser degree. In both cases the NBs have the same means as the gammas that mixed them. The difference is in the

variance of the NBs. The NBs produced by the gammas with shape held constant (Figure 1D) exhibit a wider range of variance for the same change in mean with scale held constant (Figure 1B). The NB samples in this study were generated by holding the shape fixed while varying the mean by changing the scale.

Pseudolikelihood (PL) Estimation

Pseudolikelihood, the term used by Gong and Samaniego (1981), is a robust method for parameter estimation similar to EQL estimation in that one specifies only the mean and variance structure rather than the complete form of the distribution. It is used in those settings in which the variance can be expressed as a function of the mean and additional parameter(s), θ, where θ is single (or vector) valued. Assuming that the mean (regression) parameters are known and equal to the current estimate $\tilde{\beta}$, the general PL equation $l_{PL}\left(\theta,\tilde{\beta}\right)$, is obtained by incorporating the assumed mean and variance function ($v(\theta,\beta)$) into a normal log-likelihood equation, which has the form $l_{PL}\left(\theta,\tilde{\beta}\right)=-\sum_{i=1}^{n}\log\left(2\pi v_i\left(\theta,\tilde{\beta}\right)^{1/2}\right)-\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\mu_i\left(\tilde{\beta}\right)\right)^2\Big/v_i\left(\theta,\tilde{\beta}\right)$ (see Carroll and Ruppert, 1982, for examples of general variance functions.). The PL estimators are the maximizers of the "pseudo-normal" likelihood equations. The PL estimating equation for a variance function parameter is $0=\frac{\partial l_{PL}}{\partial\theta}=\sum_{i=1}^{n}\left[\left(\frac{r_i^2}{v_i}-1\right)\frac{\partial v_i}{\partial\theta}\frac{1}{v_i}\right]$, where $r_i=y_i-\mu_i\left(\tilde{\beta}\right)$ is the i^{th} residual. A modification of the PL estimating equation that incorporates leverage was introduced by Davidian and Carroll (1987). The modification to $0=\frac{\partial l_{PL}}{\partial\theta}$

replaces $\left(\frac{r_i^2}{v_i}-1\right)$ with $\left[\frac{r_i^2}{v_i}-(1-h_i)\right]$ where the h_i are diagonal elements of the projection or "hat" matrix produced from estimation of the mean parameters.

PL estimation was used in the NB setting by Breslow (1989, 1990), who estimated the mean parameters using QL and used PL for the single equation to estimate the variance parameter. The equations use $\theta=a$ and $v_i=\mu_i+a\mu_i^2=V_i$ (previous NB notation for variance function). The PL estimating equation for a is $\tilde{U}(a)_{PL}=\sum_{i=1}^{n}\left[\frac{r_i^2}{V_i}-(1-h_i)\right]\frac{\partial V_i}{\partial a}.\frac{1}{V_i}=0$, where the h_i are the diagonal elements of the projection matrix that arises at convergence of the (QL) iterated weighted least squares solution for the mean parameters for a given value of the variance parameter. The projection matrix, H, is written: $H=Q\left(Q^tQ\right)^{-1}Q^t$, where $Q_{nxp}=\left(diag\left(\frac{\mu_i}{V_i^{1/2}}\right)\right)_{nxn}X_{nxp}$.

PL estimation for the variance function is usually alternated with use of generalized least squares for estimation of β. Use of GLS for estimating β. would amount to minimizing only the second term in $l_{PL}\left(\theta,\tilde{\beta}\right)$ above and arriving at the same solutions for β as QL. Note that one could use the same PL "pseudo-normal" method to obtain estimates for the mean parameters given the current estimate of the variance parameter(s). In this case, only for the Gaussian model with constant variance, would the QL and PL estimators for mean parameters be the same. The PL estimates differ because β occurs in both terms of the PL equation, whereas the added term in EQL is an empirical variance expression which does not contain the mean parameters (Nelder, 1992). For use in variance function estimation, Davidian and Carroll (1988) note that the PL method is asymptotically equivalent to weighted regression on

squared residuals with estimated weights, both being based on the method of moments. Thus the estimating equation for variance parameter(s) is unbiased and the estimates are consistent under general conditions. They prefer PL to EQL for this reason and others based on asymptotic results.

I used PL estimation for the variance parameter only, as implemented for NB by Breslow (1989). For the NB single population model the mean estimate (using QL) and variance estimator (using PL) are $\tilde{\mu} = \bar{y}$ and $\tilde{a}_{PL} = \frac{S^2 - \bar{y}}{\bar{y}^2}$, where $S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the usual sample variance. So, the pseudolikelihood estimator for a in this setting is just the Method of Moments (MOM) estimator. (See calculations and simplifications in Appendix B).

Optimal Quadratic (OQ) Estimation

If, in addition to knowledge about the functional relationship between the mean and variance, one had knowledge about the skewness and kurtosis of the distribution, one could follow the outline of Godambe and Thompson (1989) and use their optimal quadratic estimating equations. In earlier work, Godambe and Thompson (1987) had shown that the QL equation is optimal among linear estimating equations, i.e., for equations of the form $g = \sum_{1}^{n} (y_i - \mu)\alpha_i$ QL provides the optimal constants α_i. The optimality property is defined as equations with minimum variance or maximal information or, equivalently, the highest correlation with the score statistic among all estimating equations of the form g. The emphasis on highest correlation with the score function $[\partial \log f(y,\theta)/\partial\theta]$ stems from the fact that under regularity conditions the score statistic provides the minimal sufficient partitioning of the sample space even when the ML

estimator is not minimal sufficient. In this sense the score function contains all the information in the sample.

Godambe and Thompson extended Godambe's work on the optimal combination of "orthogonal" estimating equations (1985, 1987) to include higher moments. They introduced an "extended quasi-score function" by adding a term to the QL equation and an additional equation to estimate a dispersion parameter. Godambe and Thompson demonstrate the connection between the quasi-score function and the extended quasi-score function and claim that the former is the natural substitute for the latter when the likelihood is undefined. Requirement of knowledge about third and fourth moments raises questions about efficiency and robustness relative to maximum likelihood and to methods requiring knowledge of only the first two moments.

The extended quasi-score function, or OQ estimating equations are written (Godambe and Thompson, 1989) in terms of a parameter vector θ_{px1} and a dispersion parameter σ^2. The means (μ's) and variances (V's) can be any specified function of the parameters in θ_{px1} and the variance is not assumed to depend on θ_{px1} only through the means. The OQ equations are of the general form $\sum_{i=1}^{n} h_{1i} w_{1i} + \sum_{i=1}^{n} h_{2i} w_{2i} = 0$. Here the orthogonal estimating functions are $h_{1i} = y_i - \mu_i$ and $h_{2i} = (y_i - \mu_i)^2 - (\sigma^2 V_i) - \gamma_1 (\sigma^2 V_i)^{1/2} (y_i - \mu_i)$. The orthogonality implies that $E(h_{1i} h_{2j}) = 0 \; \forall i, j = 1 \ldots n$ and that the estimating equations are additive or provide "additive information". The optimal weights (optimality as defined in previous paragraph) are $w_{1i} = \dfrac{(\partial \mu_i / \partial \theta_r)}{\sigma^2 V_i}$ and $w_{2i} = \dfrac{\left[\gamma_1 \left(\dfrac{\partial \mu_i}{\partial \theta_r}\right) - \sigma \left(\dfrac{\partial V_i}{\partial \theta_r}\right) \Big/ V_i^{1/2}\right]}{(\sigma^2 V_i)^{3/2} (\gamma_2 + 2 - \gamma_1^2)}$ (proof in Godambe and Thompson (1989)). The first

summation ($\sum_{i=1}^{n} h_{1i} w_{1i} = 0$) is just the quasi-score function from QL. The second summation ($\sum_{i=1}^{n} h_{2i} w_{2i} = 0$) incorporates the higher moments, skewness (γ_1) and kurtosis (γ_2).

For the general GLM NB model (p. 8) with means expressed in terms of px1 explanatory variables **x** and parameter vector β and a common variance parameter (a), the OQ estimating equations reduce to QL for the mean parameters and solving the additional OQ estimating equation for the variance parameter: $\tilde{U}(a)_{OQ} = \sum_{i=1}^{n} \left\{ \frac{\left[(y_i - \tilde{\mu}_i)^2 - \tilde{\mu}_i(1+a\tilde{\mu}_i) - (1+2a\tilde{\mu}_i)(y_i - \tilde{\mu}_i) \right]}{(1+a\tilde{\mu}_i)^2} \right\} = 0$. Under H_0 for the treatment versus control NB model, the QL estimators are $\tilde{\mu} = \bar{y}$ and $\tilde{a}_{OQ} = \frac{\left(\frac{n-1}{n}\right) S^2 - \bar{y}}{\bar{y}^2}$ (See calculations in Appendix C).

Conditional Maximum Likelihood (CML) Estimation

A technique used to obtain an estimating equation for a parameter of interest in a multiple parameter distribution is conditional likelihood. The full likelihood is factored into a conditional likelihood and a residual likelihood. The conditional likelihood depends on only the parameter of interest. The residual likelihood depends on the other parameters. The estimating equation for the parameter of interest is obtained by maximizing the conditional likelihood rather than the full likelihood. This is the same rationale that underlies the popular Restricted Maximum Likelihood Estimate (REML) for variance components in analysis of variance settings (Searle, et. al., 1992, Chapter 3).

The following factorization was presented in Anraku and Yanagimoto (1990). The negative binomial likelihood for a single population can be factored into a conditional likelihood, L_C, which depends on the variance parameter but not on the mean, and a residual likelihood, L_R, as follows. Conditioning on $t=\sum_{i=1}^{n} y_i$, factor the likelihood as

$L(y_i;\beta,a) \quad = \quad L_C(y_i;a|t) \quad \times \quad L_R(t;\beta,a)$, which is

$$\prod_{i=1}^{n}\frac{\Gamma(y_i+a^{-1})}{y_i!\Gamma(a^{-1})}\left(\frac{a\mu}{1+a\mu}\right)^{y_i}\left(\frac{1}{1+a\mu}\right)^{a^{-1}} = \frac{\prod_{i=1}^{n}\frac{\Gamma(y_i+a^{-1})}{y_i!\Gamma(a^{-1})}}{\binom{n/a+t-1}{t}}\times\binom{n/a+t-1}{t}\left(\frac{a\mu}{1+a\mu}\right)^{t}\left(\frac{1}{1+a\mu}\right)^{n/a}.$$

The conditional likelihood, $L_C(y_i;a|t)$, is proportional to

$$LC=\frac{\prod_{i=1}^{n}\frac{\Gamma(y_i+a^{-1})}{\Gamma(a^{-1})}}{\frac{\Gamma(n/a+t)}{\Gamma(t)}}=\frac{\prod_{i=1}^{n}\{1/a(1/a+1)(1/a+2)\cdots(1/a+y_i-1)\}}{n/a(n/a+1)(n/a+2)\cdots(n/a+t-1)}$$

$$= \frac{\prod_{i=1}^{n}\prod_{j=0}^{y_i-1} 1/a(1+aj)}{\prod_{j=0}^{t-1} 1/a(n+aj)} = \frac{\prod_{i=1}^{n}\prod_{j=0}^{y_i-1}(1+aj)}{\prod_{j=0}^{t-1}(n+aj)}, \text{ and,}$$

$$\log(LC) = \left[\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\log(1+aj)\right] - \left[\sum_{j=0}^{t-1}\log(n+aj)\right].$$

The maximizer of the conditional likelihood is the solution $\tilde{a}_{CML}$ to $\frac{\partial \log(LC)}{\partial a} = 0 = \left[\sum_{i=1}^{n}\sum_{j=0}^{y_i-1}\frac{j}{1+aj}\right] - \left[\sum_{j=0}^{t-1}\frac{j}{n+aj}\right]$. It is obtainable using root finding techniques.

Anraku and Yanagimoto state that "$\bar{y}$ is a reasonable estimator of μ irrespective of the estimator of a." They suggest $\tilde{a}_{CML}$ as a robust alternative to the maximum likelihood estimate for the variance parameter when $\frac{n-1}{n}S^2 > \bar{y}$. Anraku and Yanagimoto claim that the uniqueness of $\tilde{a}_{CML}$ when $S^2 > \bar{y}$ follows from work by Levin and Reeds (1977) in terms of the compound multinomial distribution. When $S^2 < \bar{y}$, they define $\tilde{a}_{CML}$ as zero. I used $\bar{y}$ and $\tilde{a}_{CML}$ as estimates under the null hypothesis setting.

Although the conditioning method as described above is not applicable to the general GLM setting, it is applicable to the one-way layout structure with a common variance parameter using conditioning on the group totals. The one-way layout was studied by Anraku and Yanagimoto (1990). Results from their simulation showed that the CML estimators for a, $\frac{a}{(a+1)}$ and $\frac{1}{a}$ performed better than those based on ML and MOM in terms of bias, mean square error and Kullback-Leibler risks for multiple populations.

Negative Exponential Disparity (NED) Estimation

The Negative Exponential Disparity estimator belongs to a class of estimators known as minimum disparity estimators (Lindsay, 1994). These estimators correspond to the minimization of a "disparity" or a measure of the distance between a pair of densities, namely the data (or empirical) density and the model (or probability) density. In a simplified sense, one can think of looking for an estimator which yields the best match between the heights corresponding to the histogram of the data and the heights corresponding to a family of (discrete) probability densities. The general theory for minimum disparity estimators is easily applied to discrete distributions. Though it can be applied with modifications to continuous distributions (Basu and Lindsay, 1994; Basu and Sarkar, 1994), I will present the general theory in the context of discrete distributions only.

Let the sample space be Y={0,1,2,...,K} with K possibly infinite, and assume m_θ is a family of probability densities on Y indexed by θ, a vector of parameters. Assume, unless otherwise noted, that summations (denoted by $\sum$) are taken over the entire sample space. Assume that the data are n iid observations made from m_θ. Let $d(y)$ be the proportion of the n sample observations which has value y. Lindsay defines the Pearson residual function $\delta(y)$ as $\delta = \delta(y) = \frac{[d(y) - m_\theta(y)]}{m_\theta(y)}$ (this name was used because the model-weighted sum of the squared residuals $\sum m_\theta(y)\delta(y)^2$ is Pearson's chi-squared distance). The disparity measure, ρ, between the data proportions, $d(y)$, and the model density values, $m_\theta(y)$, is a function of the Pearson residual function and defined as $\rho_\theta(d, m_\theta) = \sum G(\delta(y))m_\theta(y)$ where G is a strictly convex thrice-differentiable function. Assuming the differentiability of the model density,

minimization of the disparity measure involves the solution of an estimating equation for θ_j of the form $-\frac{\partial}{\partial\theta_j}\rho(d,m_\theta)=\sum A(\delta(y))\frac{\partial m_\theta(y)}{\partial\theta_j}=0$,where $A(\delta)=(1+\delta)G'(\delta)-G(\delta)$ is termed the residual adjustment function (RAF). The form and properties of the estimation procedure depend on the choice of G. Lindsay presents disparity measure methods as an unifying concept and lists the choices of G (and associated RAFs, A) for many estimation techniques, including maximum likelihood (ML) and the common distance-type measures of minimum Pearson's chi-squared (PCS), minimum Neyman's chi-squared (NCS), minimum Kullback-Leibler divergence (KL) and minimum Hellinger distances (HD) (see Lindsay, 1994, pp. 1086-1089, 1101, 1103). The shape of the RAF determines the tradeoff between the estimator's robustness and efficiency. RAF construction or choice is listed as a starting point for building new disparity measures. The RAFs chosen with strict convexity and differentiability insure some desirable properties of the estimators.

Lindsay (1994, p. 1103) introduces the Negative Exponential disparity (NED) as a new disparity measure, determined by $G(\delta)=e^{-\delta}-1$. NED estimators are shown to be second-order efficient, and robust in the sense that they reduce the effects of outliers and "inliers" in the data (proofs in Lindsay). Inliers (term first used by Lindsay), or smaller than expected sample proportions, are defined as values of δ near -1. They can be due to empty cells, i.e., $d(y)$=0. The RAFs in Figure 2 (p. 127) present a combination of Lindsay's Figures 3 and 5. Curvature towards the x-axis indicates the degree of robustness to outliers (for δ>0) and inliers (for δ<0). The NED has the only RAF in the plot that is robust for both (is convex for δ>0 and concave for δ<0).

I chose to apply the NED method for estimation and testing. To my knowledge, it has not been used for the NB(μ,a) distribution. Construction for the estimators and a test follow.

NED applied to NB single population

I provide here some notation and abbreviations of disparity functions in terms of NED and the NB(μ,a) family.

Let $\theta=(\mu(\beta),a)$ be the vector of parameters which indexes the NB family of pmfs, and $m_\theta=m_\theta(y)=\frac{\Gamma\left(y+a^{-1}\right)}{y!\Gamma\left(a^{-1}\right)}\left(\frac{a\mu}{1+a\mu}\right)^y\left(\frac{1}{1+a\mu}\right)^{a^{-1}}$, $y=0,1,2,\ldots$, denote the model density or likelihood. I will also use the NB likelihood with the parameterization of $k=a^{-1}$, yielding $m_\theta=\frac{\Gamma(y+k)}{y!\Gamma(k)}\left(\frac{\mu}{\mu+k}\right)^y\left(\frac{k}{k+\mu}\right)^k$. Let $l_\theta=l_\theta(y)=\log[m_\theta(y)]$ denote the log-likelihood. Thus, $\frac{\partial m_\theta}{\partial\theta_j}=m_\theta\frac{\partial l_\theta}{\partial\theta_j}=m_\theta u_j$, where u_j is the score function or derivative of the log-likelihood with respect to θ_j (similarly, u_μ will denote derivative of the log-likelihood with respect to μ). Recall, the Pearson residual, $\delta=\delta(y)=\frac{[d(y)-m_\theta(y)]}{m_\theta(y)}$, takes on the value δ=-1 whenever $d(y)=0$. A sample proportion of zero, $d(y)=0$, results when the value y does not occur in the realized data. Use of this fact results in some simplifications in numerical calculations of several equations below.

The disparity function is $G(\delta)=e^{-\delta}-1$ and has derivatives $G'(\delta)=-e^{-\delta}$ and $G''(\delta)=e^{-\delta}$. Thus, $A^*(\delta)=(1+\delta)G'(\delta)-G(\delta)=1-(2+\delta)e^{-\delta}$ and $A(\delta)=A^*(\delta)-A^*(0)=2-(2+\delta)e^{-\delta}$ is the RAF, adjusted so that $A(0)=0$.

The disparity measure is defined as

$$\begin{aligned}\rho_\theta(d,m_\theta) &= \sum G(\delta(y))m_\theta(y) \\ &= G(-1)+\sum\left[G(\delta(y))-G(-1)\right]m_\theta(y) \\ &= (e-1)+\sum\left[e^{-\delta(y)}-e\right]m_\theta(y), \text{ for NED.}\end{aligned}$$

The equations in the second and third lines are simplifications used in numerical computations and are equivalent to the first equation because $G(-1)$ is constant and $\sum m_\theta(y)=1$. The equations are a useful representation since $\left[G(\delta)-G(-1)\right]=0$ whenever $\delta=\delta(y)=-1$, so the disparity measure need be summed only over observed data values rather than the entire support space. I shall refer to equations summed over observed data points as "working" equations. They are the forms I used in computation.

The general form for the NED estimating equations is $-\frac{\partial}{\partial\theta_j}\rho(d,m_\theta)=0=\sum A(\delta)\frac{\partial m_\theta(y)}{\partial\theta_j}=\sum\left[A(\delta)-A(-1)\right]m_\theta(y)u_j$. The last equation is the working equation and is equivalent because $A(-1)$ is a constant and $\sum m_\theta(y)u_j=E(u_j)=0$. So, the estimators are the solutions to $0=\sum\left[e-(2+\delta)e^{-\delta}\right]m_\theta(y)u_j$, where u_j is the score function for the estimated parameter.

To estimate the mean of NB(μ,a), include the mean score function, $u=u_\mu=\frac{y-\mu}{\mu(1+a\mu)}$. The NED estimating equation for the mean becomes

$$-\frac{\partial}{\partial\mu}\rho(d,m_\theta)=0=\sum\left[A(\delta)-A(-1)\right]m_\theta(y)u_\mu=\sum\left\{\frac{\left[e-(2+\delta)e^{-\delta}\right]m_\theta}{\mu(1+a\mu)}\right\}(y-\mu).$$

The estimator for the NB mean, $\tilde{\mu}$, can be iteratively reweighted. The updated estimate $\tilde{\mu}_{r+1}=\frac{\sum w_r y}{\sum w_r}$, where the weights from the rth iteration, w_r, are

$w_r = \dfrac{\left[e-(2+\delta_r)e^{-\delta_r}\right]m_{\theta r}}{\tilde{\mu}_r(1+a\tilde{\mu}_r)}$, and where a is the current estimate of the variance parameter.

The estimator for the NB variance parameter, a, must be obtained by a root finding routine in conjunction with the NED equations. This is because one cannot factor or otherwise separate out the variance parameter in the score function. The working form of the estimating equation is, again, $-\dfrac{\partial\rho(d,m_\theta)}{\partial a} = 0 = \sum\left[A(\delta)-A(-1)\right]m_\theta(y)u_a$, where u_a is the variance parameter score function, i.e., $u = u_a = \sum_{j=0}^{y-1}\left(\dfrac{j}{1+aj}\right)+a^{-2}\log(1+a\mu)-\dfrac{\left(y+a^{-1}\right)\mu}{(1+a\mu)}$. I solve for the root of this NED estimating equation using the Newton-Raphson method. In order to apply this numerical technique, I need an additional partial derivative.

For the disparity measures in general,

$$\frac{-\partial^2\rho}{\partial\theta_j^{\,2}} = \frac{\partial}{\partial\theta_j}\left(\frac{-\partial\rho}{\partial\theta_j}\right) = \frac{\partial}{\partial\theta_j}\left[\sum A(\delta)\frac{\partial m_\theta}{\partial\theta_j}\right]$$

$$= \sum\left[\frac{\partial A(\delta)}{\partial\theta_j}\frac{\partial m_\theta}{\partial\theta_j} + A(\delta)\frac{\partial^2 m_\theta}{\partial\theta_j^{\,2}}\right] = \sum\left\{\left(\frac{\partial A(\delta)}{\partial\delta}\frac{\partial\delta}{\partial m_\theta}\frac{\partial m_\theta}{\partial\theta_j}\right)\frac{\partial m_\theta}{\partial\theta_j} + A(\delta)m_\theta\left[\left(\frac{\partial l}{\partial\theta_j}\right)^2 + \frac{\partial^2 l}{\partial\theta_j^{\,2}}\right]\right\}$$

$$= \sum m_\theta\left\{-A'(\delta)(\delta+1)\left(\frac{\partial l}{\partial\theta_j}\right)^2 + A(\delta)\left[\left(\frac{\partial l}{\partial\theta_j}\right)^2 + \frac{\partial^2 l}{\partial\theta_j^{\,2}}\right]\right\}.$$

For NED and the NB variance parameter, note that $A'(\delta) = (1+\delta)e^{-\delta}$, which is zero whenever $\delta(y) = -1$. Use of $\left[A(\delta)-A(-1)\right]$ in place of $A(\delta)$ does not change the equation, since $\sum m_\theta\left[\left(\dfrac{\partial l}{\partial\theta_j}\right)^2 + \dfrac{\partial^2 l}{\partial\theta_j^{\,2}}\right] = E\left[\left(\dfrac{\partial l}{\partial\theta_j}\right)^2 + \dfrac{\partial^2 l}{\partial\theta_j^{\,2}}\right] = 0.$

The NB working equation becomes

$$\frac{-\partial^2\rho}{\partial a^2}=\frac{\partial}{\partial a}\left(\frac{-\partial\rho}{\partial a}\right)=\sum m_\theta\left\{\left[(1+\delta)^2e^{-\delta}\right]\left(\frac{\partial l}{\partial a}\right)^2+\left[e-(2+\delta)e^{-\delta}\right]\left[\left(\frac{\partial l}{\partial a}\right)^2+\frac{\partial^2 l}{\partial a^2}\right]\right\},$$

where $\frac{\partial l}{\partial a}=u_a$ is listed above and

$$-\frac{\partial^2 l}{\partial a^2}=-\frac{\partial u_a}{\partial a}=\sum_{j=0}^{y-1}\left(\frac{j}{1+aj}\right)^2+2a^{-3}\log(1+a\mu)-\frac{2a^{-2}\mu}{1+a\mu}-\frac{\left(y+a^{-1}\right)\mu^2}{(1+a\mu)^2}.$$

Thus, the estimator for a is obtained by iterative solutions to the Newton Raphson equation, updating estimates using $\tilde{a}_{r+1}=\tilde{a}_r-\frac{(-\partial\rho/\partial a)}{\left(-\partial^2\rho/\partial a^2\right)}\Big|a=\tilde{a}_r,\mu=\tilde{\mu}.$

Estimates for μ and a are obtained alternately until convergence of the overall disparity measure ρ. The convergence criteria for both the individual estimates and the overall system or set of estimates is at each stage based on the convergence in the relative size ofρ, i.e., $(\text{abs}(\rho_{new}-\rho_{old}/\rho_{new})<\text{tol}.$

Treatment versus Control setting

The NED methodology may be extended to a treatment versus control setting. Just as the estimates in a single population setting can be thought of as minimizing an expectation, $\rho=E(G(\delta))$, one can think of the two population setting as minimizing a ρ defined as the expected value of a conditional expectation. Here I will calculate ρ for the sample from each population and use the sampling technique of weights proportional to size. Thus, I will use weights $w_i=n_i/n$. and minimize $\rho=E_{\mathbf{x}}\left[E\left(G|\mathbf{x}_i\right)\right]=\sum_{i=1}^{2}\frac{n_i}{n}\sum G\left(\delta\left(y|\mathbf{x}_i\right)\right)m_\theta\left(y|\mathbf{x}_i\right),\quad (*)$

where $\mathbf{x}_i^t$ for $i=1,2$ are the unique rows of X. Here, $\mathbf{x}_1^t=[1\ \ 1]$ and $\mathbf{x}_2^t=[1\ \ -1]$ indicate the groups of control or treatment.

In the 2 population setting with parameters $\theta=(\mu_1,\mu_2,a)$, use $\delta_i=\delta\left(y|\mathbf{x}_i\right)=\frac{\left[d\left(y|\mathbf{x}_i\right)-m_\theta\left(y|\mathbf{x}_i\right)\right]}{m_\theta\left(y|\mathbf{x}_i\right)}$ as the Pearson residuals for the sample proportions and model density per group or population. Similarly,

$G_i = G\big(\delta(y|\mathbf{x}_i)\big)$ and $A_i = A\big(\delta(y|\mathbf{x}_i)\big)$ are functions are defined in terms the two populations via δ_i. Again, let $G_i' = \partial G_i/\partial\delta_i$ and $A_i' = \partial A_i/\partial\delta_i$.

Solving for the parameters in the mean one obtains

$$-\frac{\partial\rho}{\partial\beta_r} = \sum_{i=1}^{2}\frac{n_i}{n}\sum A\big(\delta(y|\mathbf{x}_i)\big)\frac{\partial m_\theta(y|\mathbf{x}_i)}{\partial\beta_r} = \sum_{i=1}^{2}\sum\frac{n_i}{n}A\big(\delta(y,\mathbf{x}_i)\big)m_\theta(y|\mathbf{x}_i)\frac{\partial l_\theta(y|\mathbf{x}_i)}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_r}$$

$$= \sum_{i=1}^{2}\sum w_i(y|\mathbf{x}_i)(y-\mu_i)\mathbf{x}_i \text{ for } r=1,2 \text{ where } w_i = w(y|\mathbf{x}_i) = \frac{n_i}{n}A\big(\delta(y|\mathbf{x}_i)\big)\frac{m_\theta(y|\mathbf{x}_i)}{(1+a\mu_i)}.$$

So, $\dfrac{-\partial\rho}{\partial\beta_1} = \sum w_1(y-\mu_1)+\sum w_2(y-\mu_2) = 0$,

and $\dfrac{-\partial\rho}{\partial\beta_2} = \sum w_1(y-\mu_1)-\sum w_2(y-\mu_2) = 0$.

Solving these simultaneously leads to $\dfrac{-\partial\rho}{\partial\beta_1}+\dfrac{-\partial\rho}{\partial\beta_2} = 0 = 2\sum w_1(y-\mu_1)$ and $\dfrac{-\partial\rho}{\partial\beta_1}-\dfrac{-\partial\rho}{\partial\beta_2} = 0 = 2\sum w_2(y-\mu_2)$. Thus one can iteratively solve for the estimate of the mean of the population (treatment or control) by using the sample from that population and the same technique as for a single population, i.e.,

$$\tilde{\mu}_{i,r+1} = \frac{\sum w_{i,r}y}{\sum w_{i,r}}.$$

One could solve for the β parameters, because $\mu_1 = e^{\beta_1+\beta_2}$ and $\mu_2 = e^{\beta_1-\beta_2}$ imply $\mu_1\mu_2 = e^{2\beta_1}$ and $\mu_1/\mu_2 = e^{2\beta_2}$. As estimates of the β's we can use $\tilde{\beta}_1 = \log\Big[(\tilde{\mu}_1\tilde{\mu}_2)^{\frac{1}{2}}\Big]$ and $\tilde{\beta}_2 = \log\Big[(\tilde{\mu}_1/\tilde{\mu}_2)^{\frac{1}{2}}\Big]$. This choice is in the spirit of Lehmann (TPE, p. 112) using the invariance property of MLEs, $\hat{g}(\theta) = g\big(\hat{\theta}\big)$. If the invariance property holds for NED estimators, the β's would be estimated as shown. These estimates seem a reasonable starting point in any case. Note, however, that only the $\tilde{\mu}_i$'s are needed in order to construct a test.

To solve for the variance parameter, histogram information from both populations is needed. Again go back to the definition of ρ and minimize the disparity measure. Using $\rho = E_{\mathbf{x}}\Big[E\big(G|\mathbf{x}_i\big)\Big] = \sum_{i=1}^{2}\frac{n_i}{n}\sum G\big(\delta(y|\mathbf{x}_i)\big)m_\theta(y|\mathbf{x}_i)$ the NED

estimating equation for the variance parameter is $-\frac{\partial \rho}{\partial a} = \sum_{i=1}^{2} \frac{n_i}{n} \sum A\left(\delta\left(y|\mathbf{x}_i\right)\right) m_\theta\left(y|\mathbf{x}_i\right) \frac{\partial l_\theta}{\partial a} = 0$. The NED variance estimator, $\tilde{a}_{NED}$, provides the root or solution to this equation.

To apply Newton-Raphson use the derivative $\frac{\partial}{\partial a}\left(-\frac{\partial \rho}{\partial a}\right) =$ $\sum_{i=1}^{2} \frac{n_i}{n} \sum m_\theta\left(y|\mathbf{x}_i\right) \left\{ -A'\left(\delta\left(y|\mathbf{x}_i\right)\right)\left(1+\delta\left(y|\mathbf{x}_i\right)\right)\left(\frac{\partial l_\theta}{\partial a}\right)^2 + A\left(\delta\left(y|\mathbf{x}_i\right)\right)\left[\left(\frac{\partial l_\theta}{\partial a}\right)^2 + \frac{\partial {l_\theta}^2}{\partial a^2}\right] \right\}$ and iterate until convergence using updates $\tilde{a}_{r+1} = \tilde{a}_r - \frac{\left(-\frac{\partial \rho}{\partial a}\right)}{\frac{\partial}{\partial a}\left(-\frac{\partial \rho}{\partial a}\right)}$. The working equations are obtainable using the same steps as for the single population estimates. Convergence criterion are based on convergence of ρ as in the single population setting (above).

CHAPTER 3

TESTING METHODS

Likelihood Ratio Test (LRT)

The statistic for the generalized Likelihood Ratio test is the logarithm of a ratio of likelihoods, $LRT = 2n(l_{H0} - l_{H1})$. McCullagh and Nelder refer to the LRT for GLMs as a "deviance" and, analogous to (Normal error) Analysis of Variance, to partitioned LRTs as Analysis of Deviance. A discussion of the difficulties that arise is covered in McCullaugh and Nelder (1989, pp. 35-36). The LRT has an asymptotic chi-square distribution with degrees of freedom equal to the difference in the numbers of parameters estimated under the null and alternative hypotheses. (McCullaugh and Nelder,1989; see Appendix C).

Disparity Difference Test (DDT)

Disparity Difference Tests (DDT) are presented by Lindsay (1994) in the simple setting of testing the parameters equal to a given value. Using the estimators obtained in the treatment versus control setting (above) I apply DDT to the desired two-sample test. DDTs constructed from NED estimators have an asymptotic Chi-square distribution if the family of model densities satisfies the conditions in Lehmann (1991, pp 409 and 429) and additional bounded expectations (Assumption 31, p. 1109 of Lindsay,1994). The DDT has an analogous form to the Likelihood Ratio Test (LRT) and is expressed as $DDT = 2n(\rho_{H0} - \rho_{H1})$. The NB family does satisfy the required conditions (proofs in Appendix D).

The disparity measures differenced in the DDT are based on conditional ρ, a weighted average of the measures calculated for the two samples. The test amounts to determining the degree of improvement or reduction in disparity which occurs when more parameters are estimated under H_1 than estimated under H_0. The asymptotic distribution of the DDT is chi-square with degrees of freedom (df) equal to the difference in the numbers of parameters calculated in the two models (Lindsay, 1994, Theorem 6 and Appendix A).

Figure 3 (p. 128) illustrates the difference in the fit to a data set due to disparity measures minimized for estimation under H_0 (p. 27) and H_1 (p. 30). This sample was generated from two NBs with $n_1 = n_2$=15 (overall n=30), $\mu_1 > \mu_2$ and a common value of a. The top left histogram is that of the total sample or the combination of the 2 samples into a single sample. The X's show the best "fit" under H_0 (the pmf based on estimation under H_0 multiplied by the total sample size). The bottom 2 histograms are those of the 2 samples (control and treatment) considered separately. The diamonds represent the best fit under H_1 to the sample means and a common variance parameter. The X's show the fit to each sample based on H_0 estimation. The difference in the disparity under H_0 and H_1 is a function of the distances between the X's or diamonds and the tops of the histograms, respectively (as pictured in the bottom row). [The top right histogram is the sum of the 2 lower histograms and the diamonds are the sum of the diamond heights. This combined sample with combined frequency estimates (diamonds) is not used in analysis but presented to show the similarity of fit to combined samples (top row) using the estimates obtained under H_0 versus H_1 (X's versus diamonds). The best fit under H_0 (top left) appears to give us the same fit as the sum of the H_1 fits on the separate samples applied to the combined sample (top right).] Again, differences in disparity are

not based on disparity calculations applied to the combined sample. The disparity measures are calculated from the fits in the bottom row where one can see a difference between the X's and diamonds. Each disparity measure is the NED-based function ((*) p. 30) of the differences between the histogram heights and the heights based on estimation (X's and diamonds). The H_0 disparity is ρ_{H0}=0.2987 and the H_1 disparity is ρ_{H1}=0.1451, resulting in a DDT statistic of 9.216 and a p-value of less than 0.01. Thus the conclusion is that there is a difference between the means of the 2 populations.

Welch Modified Two-Sample t-test (t)

This version of the t-test is conventionally recommended when the variances are unequal (Miller,1986, section 2.3; Snedecor and Cochran, 1980). The statistic is $t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_d}$, where $s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Its distribution under the null hypothesis can be approximated by a t-distribution where the degrees of freedom (df) are calculated by a formula based on the variances of the two samples, i.e., $\mathrm{df} = \left[\frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}\right]^{-1}$, where $c = \frac{s_1^2}{n_1 s_d^2}$. In the simulation study, the t is applied to log(y+0.5), where y is the sample data, as this is the standard transformation used on microbial count data (Niemelä, 1983). The intent is not to provide the true variance stabilizing (or normalizing) transformation but to mimic what is commonly used in practice. The Welch t-test is used instead of the standard t-test because it is available in software packages and should handle problems remaining if the transformation is less than optimal.

Generalized Score Tests (S)

Boos (1992) presented generalizations of Rao's score test (Rao 1948; see also Cox and Hinkley, 1974) which make use of general estimating equations (rather than just derivatives of the (Normal) log-likelihood) and empirical variance estimates. Rao's (original) score statistics (notation from Boos (1992)) have the general form $S(\tilde{\theta})^t \tilde{I}_f^{-1} S(\tilde{\theta})$, where $S(\theta)$ is the vector of partial derivatives of the log-likelihood function, $\tilde{\theta}$ is the vector of restricted maximum likelihood estimates under H_0 and $\tilde{I}_f$ is the Fisher information of the sample evaluated at $\tilde{\theta}$. One advantage of the score statistics is that they only require computation of the parameter estimates under the null hypothesis. They are asymptotically equivalent to the likelihood ratio statistics under the null hypothesis, H_0. Both the likelihood ratio statistic and the score statistic are invariant under reparameterization or non-linear transformations.

Recall, a (general) estimating equation $g(\mathbf{y},\theta)$ is defined as any function of the data and parameters having zero mean for all parameter values (Godambe and Thompson, 1984). Provided there are as many equations as parameters, the estimates $\tilde{\theta}_i$ are obtained by solving the vector equation $g(\mathbf{y},\theta)=0$ for θ. Boos (1992) refers to the estimating equations as "score functions".

The generalized score test is constructed in the following manner, outlined in Boos (1992, p. 328).

> "Find the asymptotic covariance matrix of the score function $g(\mathbf{y},\theta_0)$ under H_0, say Σ_g. Then define the generalized score statistic to be $T_{GS} = g(\mathbf{y},\theta)^t \tilde{\Sigma}_g^- g(\mathbf{y},\theta)$, where $\tilde{\Sigma}_g^-$ is a generalized inverse of a consistent estimate $\tilde{\Sigma}_g$ of Σ_g. Usually a version of $\tilde{\Sigma}_g^-$ is available which is easily computed."

The score statistics have an asymptotic null chi-squared distribution. Boos shows how the various forms of generalized score tests arise from Taylor expansion of the estimating equations.

General information on Score tests in a GLM setting are found in Pregibon (1982). Breslow (1989, 1990) presents the theory and form of the score tests in the NB setting for testing a hypothesis about a subset of the mean vector. Estimation of the mean parameters is via QL and the variance parameter via an additional equation. Breslow uses PL for the variance parameter but the following discussion would hold for any estimation procedure described above (see Appendix E). The QL estimating equations for the mean are $\underset{p\times 1}{U}(\beta,a)=\sum_{i=1}^{n}\mathbf{u}_i=\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{V_i}\frac{\partial\mu_i}{\partial\beta_{p\times 1}}=\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{(1+a\mu_i)}\mathbf{x}_i=X^t\left[diag\left(\frac{1}{1+a\mu_i}\right)\right](\mathbf{y}-\mu)$,

where $\mathbf{x}_i$, the covariates for the i^{th} observation, is the transpose of a row of X (refer to section on EQL Estimation). Breslow assumes that the mean structure is correctly specified but allows for possible misspecification of the variance. The empirical covariance matrix for score statistics of the vector of mean parameters, β, is $G=\sum_{i=1}^{n}\mathbf{u}_i\mathbf{u}_i^t=\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2\mu_i^2}{V_i^2}\mathbf{x}_i\mathbf{x}_i^t=X^t\left[diag\left(\frac{(y_i-\mu_i)^2\mu_i^2}{V_i^2}\right)\right]X$. The negative expectation of the partial derivatives of the mean score statistic is $A=-\sum_{i=1}^{n}E\frac{\partial\mathbf{u}_i}{\partial\beta^t}=\sum_{i=1}^{n}\frac{\mu_i^2}{V_i}\mathbf{x}_i\mathbf{x}_i^t=X^t\left[diag\left(\frac{\mu_i^2}{V_i}\right)\right]X$. The asymptotic variance of $\tilde{\beta}$ is estimable by $A^{-1}GA^{-1}$, even if the variance is misspecified. This is called the "empirical covariance matrix". If the variance is correctly specified then $E(G)=A$ and one may estimate Var(β) by A^{-1}, which Breslow refers to as "model based".

In the general setting, β is partitioned into sub vectors of lengths p_1 and p_2, i.e., $\beta_{px1} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}_{p2\,x\,1}^{p1\,x\,1}$. The set of covariables and other matrices are conformably partitioned. Generalized score tests for testing $H_0 : \beta_2 = \beta_2^0$ are based on $U_2 = X_2^t \left[diag\left(\frac{\mu_i}{V_i} \right) \right] (\mathbf{y} - \mu) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mu_i}{V_i} \mathbf{x}_{2i}$ where $\tilde{\beta}$ is the quasi-likelihood estimate under H_0, i.e., $\tilde{\beta}_2 = \beta_2^0$ and $\tilde{\beta}_1^0$ is the solution to $U_1 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mu_i}{V_i} \mathbf{x}_{1i} = 0$. The test statistic is $T_{GS} = U_2{}^T \tilde{\Sigma}_g^{-1} U_2$, where $\tilde{\Sigma}_g = I_{2\cdot 1} = A_{22} - A_{21} A_{11}^{-1} A_{12}$ for a model based score test and $\tilde{\Sigma}_g = I_{2\cdot 1}^e = G_{22} - A_{21} A_{11}^{-1} G_{12} - G_{21} A_{11}^{-1} A_{12} + A_{21} A_{11}^{-1} G_{11} A_{11}^{-1} A_{12}$ for an empirical score test. The matrices A and G have been partitioned into appropriate sub matrices of the dimensions $p_1 \times p_1$, $p_1 \times p_2$, etc.

For the treatment versus control setting testing $H_0 : \beta_2 = 0$ with $n_1 = n_2 = n/2$, the model based score test is $T_{GS}^m = \frac{n(\bar{y}_c - \bar{y}_t)^2}{4\bar{y}(1 + \tilde{a}\bar{y})}$ where $\tilde{a}$ is the null hypothesis estimator and $\bar{y}_t$, $\bar{y}_c$ are the sample means for treatment and control. A model based score test statistic is calculated for each of the estimation methods described above. The empirical score test statistic for the same setting is $T_{GS}^e = \frac{n^2(\bar{y}_c - \bar{y}_t)^2}{4 \sum_{total} r_i^2}$ for all methods where $r_i = y_i - \bar{y}$ are the residuals. The model score test statistic with $\tilde{a}$ estimated via OQ Estimation gives the same test statistic as the empirical score test. (For more details on Score Tests, see Appendix E).

CHAPTER 4

SIMULATION STUDY

A simulation study was conducted with two objectives: (i) compare estimators of the NB variance parameter and (ii) compare two-sample test methods.

Study Methods

Data generated under the null hypothesis (H_0) are generated from a single population setting. Data generated under the alternative hypothesis (H_1) are from two populations with different means but a common variance parameter. Each sample is comprised of two subsamples with $n_1=n_2=n/2$.

For simulation under the null hypothesis, a 3x3x2 factorial design was used. The first factor was the NB mean (3 levels: μ=1,2,5), the second factor was the NB variance parameter (3 levels: a=.2,.5,1) and the third factor was sample size (2 levels: $n_1=n_2$=15, $n_1=n_2$=25). One reason for this choice of levels is that they overlap with those from previous simulation studies that compared estimation methods for the NB variance parameter (Willson et. al. (1984), van de Ven (1993), Clark and Perry(1989), Piegorsch (1990)). A second reason for the choice of levels is due to applicability to analyses of drinking water and disinfection studies in environmental microbiology. Sample estimates of means and NB variance parameters reported in the literature often fall within the ranges of the levels chosen. The range of $\tilde{\mu} \in [1,5]$ was reported in water quality studies [El-Shaarawi, Esterby and Dutka (1981); Pipes and Christian (1982); Pipes, Ward and Ahn (1977)]; and $\tilde{a} \in [.2,1]$ (or $\tilde{k} \in [1,5]$) was estimated from

microbiological count data (Pipes, Ward and Ahn (1977); Maul, El-Shaarawi and Block (1985); Maul and El-Shaarawi (1991)); $n \in$30-50 is about the maximum sample sizes in environmental microbiology. I noted that the chosen parameter ranges were more representative of coliform counts or other microbes occurring at low density. I found estimates of negative binomial parameters reported in the literature that were outside this range (eg: means in Maul, El-Shaarawi and Block (1985); variance parameter in Pipes and Christian (1982), Christian and Pipes (1983)). The estimate values outside the ranges were due to microbes that commonly occur at moderate to high concentrations and due to the greater spatial and temporal variability that often occurs in environmental sampling as opposed to a controlled laboratory setting.

For simulation under the alternative hypothesis, the mean used to generate one of the subsamples was kept at the null hypothesis setting and the mean of the other subsample was varied for the desired power comparisons. I shall refer to both the choice of parameters and (sometimes) the set of samples generated for that choice using NB($\mu_1=\mu_2=\mu,a$) or NB($\mu_1>\mu_2,a$) as the "H_0 setting" or "power setting" respectively.

The computing environment used was S-PLUS version 3.2. Random NB counts were generated as random Poisson variates with means generated as random gamma variates with the desired shape and scale. Thus, the NB counts are generated as a gamma mixture of Poissons. The random NB generator available in S-PLUS was based on the Pascal form that uses only integer k ($=a^{-1}$) and I wrote a routine for general k. The estimation and testing routines were built using S-PLUS commands. I noted that Venables (1994) has written S-PLUS code to fit and analyze NB GLMs for a nested sequence of models using analysis of deviance (LRT).

The variance parameter was estimated using the single population assumption (under H_0) for all methods. Because use of the generalized score tests does not require estimation under the alternative hypothesis, the 2-sample variance parameter estimate was not calculated for several methods and comparisons were made on the single population (H_0) estimates.

The estimation methods are listed below along with their acronymns and the page numbers on which they are described.

Estimation Method	Acronym	Description
Maximum Likelihood	ML	pg. 12
Extended Quasi-Likelihood	EQL	pg. 14
Pseudolikelihood	PL	pg. 18
Optimal Quadratic	OQ	pg. 20
Conditional Maximum Likelihood	CML	pg. 23
Negative Exponential Disparity	NED	pg. 25

The estimation methods are rated according to the size of the errors of the estimates ($\tilde{a}-a$) calculated under the null hypothesis. The criteria for comparison of the estimation methods are average bias, mean square error (MSE) and means of root squared errors (MRSE) as a robust alternative to MSE. The MRSE is discussed in the next to last section of this chapter. The differences in MRSE are tested via simultaneous confidence intervals using the "all possible contrasts" form of Hotelling T^2 (Seber 1984, section 3.4; Miller 1966, p. 197; Scheffé 1959, section 3.5).