DETECTING OVERDISPERSION IN DATA WITH CENSORING

by

Hsing-Yi Chang
Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2162T

July 1996

DETECTING OVERDISPERSION IN DATA WITH CENSORING

by

Hsing-Yi Chang

A dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics

Chapel Hill

1996

ABSTRACT

HSING-YI CHANG. Testing Overdispersion in Data With Censoring. (Under the direction of Chirayath M. Suchindran.)

The term overdispersion refers to the situation that the variance of the outcome exceeds the nominal variance. Overdispersion in general has two effects. The first effect is that summary statistics have a larger variance than anticipated under the simple model. The second is a possible loss of efficiency in using statistics appropriate for the assumed distribution. Censoring is common in medical experiments, and hence estimation methods must allow for it if they are to be generally useful. The review of the statistical literature shows that many models and test statistics for overdispersion do not deal with censored data.

We extend two methods to include censoring to detect the existence of overdispersion. The first method is mixing an exponential family with some unknown distribution whose first two moments are specified. The second is the double exponential family developed by Efron (1986), which introduces an additional parameter into the one parameter exponential family to control for the variance independently of the mean. We use the score statistics to test the null hypothesis of no overdispersion.

Simulations are used to study the behavior of the test statistics in data with different degrees of censoring with different sample sizes. The results indicate that (1) the test statistics have enough power in detecting the existence of overdispersion when the degree of censoring is mild ($< 40\%$) and sample sizes

are sufficiently large for the overdispersed geometric distribution; (2) the test statistics have enough power in detecting the existence of overdispersion in all the situations simulated for overdispersed exponential distributions; and (3) test statistics based on the mixture of exponential family are normally distributed when the amount of overdispersion is large, and normality does not hold when the amount of overdispersion is small; while the test statistics based on the double exponential family do not distribute normally in either cases.

Both test statistics are applied to three sets of data which represent different types of survival time, different degree of censoring and different sample sizes. These data pertain to (1) the waiting time to conception, (2) survival time after myocardial infarction and (3) duration of breast feeding. The test statistics can detect the existence of overdispersion in all cases.

ACKNOWLEDGEMENT

First, I would like express my sincere gratitude to my advisor Dr. C. M. Suchindran, for his guidance, patience and encouragement. I appreciate the editing and support from Dr. R. Bilsborrow; the editing and constructive comments from Dr. D. Quade; the suggestions for improvement from Dr. J. Cai and the review and comments from Dr. B. Margolin.

I would also like to thank my extended family in Chapel Hill, the Chapel Hill Chinese Christian Fellowship, for their love, encouragement and prayers.

Finally, I deeply appreciate my parents, sisters and brother for their support throughout all my study in the U.S.

CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION AND REVIEW OF LITERATURE

This study reviews the models and statistics derived for the overdispersed exponential family and applies them to censored data. The goals are (1) to examine the effects of censoring on tests of overdispersion, (2) to extend the existing overdispersion models to accommodate data with censored observations, and (3) to derive the statistics which can detect overdispersion in data with censoring based on the extended models. The details of the study plan will be presented after a brief review of related works.

1.1 Definition of Overdispersion and Its Effects

The term overdispersion refers to the situation in which the variance of the outcome exceeds the nominal variance (McCullagh and Nelder, 1990). For example, if we assume $Y_1,\ldots Y_n$ are independently and identically distributed in a Poisson distribution, the mean θ of the underlying distribution is optimally estimated by $\overline{Y} = \sum Y_j/n$. Overdispersion is most simply represented by supposing that Y_j has a Poisson distribution of mean Θ_j, where $\Theta_1,\ldots\Theta_n$ are independently and identically distributed with a gamma distribution of mean μ and index γ, then the variance of Y_j has the following expression

$$n \text{ var } (\overline{Y}) = \mu\ (1 + \mu/\gamma),$$

which is the original variance μ inflated by a factor (Cox, 1983).

Overdispersion is not uncommon in practice. In fact, some believe that overdispersion is the norm in practice and nominal dispersion the exception. The incidence and degree of overdispersion encountered greatly depend upon the field of application. For example, in large-scale epidemiological studies concerning the geographical variation in the incidence of disease, the binomial variance is often an almost negligible component of the total variance. Unless there are good external reasons for relying on the binomial assumption, it is wise to be cautious and to assume that overdispersion is present unless and until it is shown to be absent (McCullagh and Nelder, 1990).

Overdispersion in general has two effects. One is that summary statistics have a larger variance than anticipated under the simple model. This has long been recognized and is commonly allowed for by an empirical inflation factor, either assumed from prior experience or estimated. The second effect is a possible loss of efficiency in using statistics appropriate for the assumed distribution (Cox, 1983).

1.2 Review of Studies on Overdispersion

Studies of overdispersion in data have focused on two regimes. One is a detailed representation of overdispersion by a specific model. The main models include a sampling density which is a mixture of the exponential families. Most of the time the log likelihood ratio is used to determine the goodness-of-fit of the model. The other is to introduce procedures which add a dispersion parameter to the exponential family (Albert and Pepple, 1989), and develop test statistics to detect the existence of the dispersion parameter. The test statistic is derived under the null hypothesis that the overdispersion parameter is zero.

Commonly, the beta-binomial model (Williams, 1975; Liang and McCullagh, 1993) and the correlated binomial (Kupper and Haseman, 1978) are used to model the overdispersed binomial distribution; the negative binomial model is used to model the overdispersed Poisson distribution (Collings and Margolin, 1985; Margolin et al., 1981), and the Dirichlet multinomial is used to model the overdispersed multinomial (Paul et al., 1989; Kim and Margolin, 1992). Since the quasi-likelihood function requires only the first two moments of the distribution, it has been applied to test overdispersion models (Breslow, 1984, 1989, 1990; O'Hara Hines and Lawless, 1993; Liang and McCullagh, 1993).

Test statistics have been obtained for the exponential family (Dean, 1992), the extra-binomial (Prentice, 1986) and the extra-Poisson (Collings and Margolin, 1985; Breslow, 1989) distributions.

1.2.1 Representation of Overdispersion by a Mixture of Exponential Family

In one-parameter exponential families such as the binomial and Poisson, the variance is a function of the mean. The existence of overdispersion, which is a common practical complication, leads to a failure of the variance-mean relation (Cox, 1983; Efron, 1986). Therefore, many efforts have been made to develop models to represent overdispersed exponential families.

A. The Overdispersed Binomial Distribution

In certain toxicological experiments with laboratory animals, the outcome is the occurrence of dead or malformed fetuses in a litter. Many researchers have indicated that the simple one-parameter binomial model generally provides a poor fit to this type of binary data. A commonly used model to deal with the situation is the beta-binomial model.

Williams (1975) proposed a beta-binomial model to handle the data from toxicological experiments designed to investigate the teratogenic or fetotoxic effect of a chemical on laboratory animals, where the experimental units are pregnant females. He assumed that within each litter the binary responses form a set of a Bernoulli trials whose probability parameter varies between litters in the same treatment group according to a two-parameter beta distribution. The parameters of the beta distribution were estimated for each treatment by maximum likelihood methods, and treatments were then compared with respect to these parameter values by using asymptotic likelihood ratio tests.

Prentice (1986) extended the beta-binomial distribution to include a range of negative correlations, and proposed approaches to carry out regression analysis for the binary response rate and for the pairwise correlation between the binary variation. One particular problem Prentice (1986) discussed was measurement error in the regression variables representing the source of overdispersion.

Kupper and Haseman (1978) presented an alternative to the beta-binomial model. They used a type of "correlated binomial" model, which allowed for possible within-litter dependence. They included second-order correlation in the model. The advantage of the model is that the correlation can be negative.

Pack (1986) investigated the properties of the likelihood ratio tests and simpler t-tests by simulation under an assumed beta-binomial model for parameter values typically found in toxicological studies. He found that likelihood ratio methods are at least as powerful as simpler approaches and in certain situations can be significantly more powerful. He also found the beta-binomial to be superior to alternative parametric models, such as the correlated binomial

models of Kupper and Haseman (1978), contrary to Kupper and Haseman's claim. His conclusion was that if the beta binomial assumption is justified, then likelihood ratio tests are the most powerful means of testing the appropriateness of the model.

Although there are many procedures for testing the binomial distribution against overdispersion alternatives, there has been very little consideration of how to test the null hypothesis of a beta-binomial distribution against all other distributions. Garren et al. (1994) proposed a test statistic based on combining Pearson statistics from individual litter sizes and evaluated its null distribution by both exact calculation and simulation. The method begins by estimating the two unknown parameters of the beta-binomial distribution via maximum likelihood (MLE), then computes a Pearson type goodness-of-fit statistic for each litter size observed in the sample. For each statistic corresponding to an observed litter size, its separate significance level is computed by simulating beta-binomial pseudo-statistics. The beta-binomial goodness-of-fit statistic is then determined by the smallest level of significance among all litter sizes.

The test seems to be powerful when the data are generated by certain mixtures of binomial random variables, or when small numbers of rare values exist in the data. It can also be extended to the logit link when multiple doses are available. Its weakness is in using of estimated values for parameters at the beginning. It is possible to re-evaluate the estimates after the simulations, but this is likely to be computationally intensive.

B. The Overdispersed Poisson Distribution

The random sample case of detecting non-specific hypervariation departures from the Poisson has received considerable attention in the literature. The standard test for this situation is the so-called "variance" test or "index of dispersion" test. The Poisson index of dispersion is $s^2/\bar{X}$. It can be shown that under the null hypothesis of a Poisson distribution, the test statistic $(n-1)s^2/\bar{X}$ is approximately $\chi^2(n-1)$, or $\sqrt{\frac{s^2}{\bar{X}}}$ is approximately normal with mean 1 and standard deviation $\sqrt{\frac{2}{n}}$.

A classical approach to the problem of the overdispersed Poisson is to treat the Poisson means associated with each observed count as latent variables that are sampled from a specific parametric distribution (Breslow, 1990; Dean, 1992). Most authors have considered a gamma mixing distribution, which leads to a negative binomial distribution for the observed data (Collings and Margolin, 1981, 1985). Margolin et al. (1981) used this model to analyze the Ames Salmonella/microsome test data.

Paul and Plackett (1978) examined the effect of the Poisson mixtures, especially as represented by the negative binomial distribution, on statistical inference. They found that probabilities of rejecting the null hypothesis of no overdispersion are increased, sometimes considerably. The findings were confirmed by Collings and Margolin (1985).

Collings and Margolin (1985) proposed methods for detecting negative binomial departures from a Poisson model. They dealt with three kinds of problems: (A) where the mean response is constant, (B) where the mean response depends on a single covariate and the regression line is through the origin, or (C)

where it takes on a fixed number of values according to a one-way layout. The negative binomial form they used is

$$\Pr\{Y = y\} = \frac{\Gamma(y + c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{cm}{1+cm}\right)^{y}\left(\frac{1}{1+cm}\right)^{-c},$$

where $y = 0, 1, 2,\ldots, 0 < m < \infty$, and $0 \leq C < \infty$. The variance of Y is $m(1 + cm)$, and by convention, values for c = 0 are understood to be evaluated in the limit as c $\rightarrow$ 0. The specific detection of negative binomial departure from the Poisson distribution may be reformulated in terms of hypothesis testing for the cases mentioned above. In each case, the null hypothesis is $H_o: c = 0$ and the alternative hypothesis is $H_1: c > 0$. The tests are based on the likelihood ratio method. The tests for the situations mentioned previously are listed below.

Table 1.1 Tests For Non-Specific Hypervariation and For Negative Binomial Departures From Poisson.

Cases	Non-specific hypervariation	Negative binomial
A	$S_A = \sum_{i=1}^{n} (Y_i - \bar{Y}_+)^2/\bar{Y}_+$	
B	$S_B = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2/\hat{Y}_i$	$T_B = \sum_{i=1}^{n} (Y_i - \beta\hat{m})^2/\bar{Y}_+$
C	$S_c = \sum_{i=1}^{r}\sum_{j=1}^{n_j} (Y_{ij} - \bar{Y}_{i+})^2/\bar{Y}_{i+}$	$T_C = \sum_{i=1}^{r}\sum_{j=1}^{n_j} (Y_{ij} - \bar{Y}_{i+})^2/\bar{Y}_{++}$

All of these tests have distributions which are approximately χ^2 under H_0. For example, the distribution of T_C conditional on $U(Y) = (\bar{Y}_{1+}, \bar{Y}_{2+}, \ldots, \bar{Y}_{n+})$ is

approximately weighted X^2

$$\sum_{i=1}^{r} \omega_i \, \chi^2(n_i - 1),$$

where $\omega_i = \overline{Y}_{i+}/\overline{Y}_{++}$. Similarly, the distribution of S_C is approximately $\chi^2(n-r)$. They found that the statistics based on the negative binomial distribution have higher power than the non-specific hypervariation in testing departures from Poisson.

Engel (1984) proposed a generalization of the log-linear modeling technique for the negative binomial model as an extension of the Poisson model. He concentrated on a two-way cross-classification. The models he considered are negative binomial with (1) only the shape parameter varying with factorial effects, i.e., X_{ij} is negative binomial (α_{ij}, $p = \frac{\theta}{1+\theta}$), with

$$E(X_{ij}) = \alpha_{ij}\theta = m_{ij}$$
$$\mathrm{Var}(X_{ij}) = \alpha_{ij}\,\theta(1+\theta) = m_{ij}(1+\theta)\ ;$$

and (2) only the scale parameter varying with factorial effects, i.e., X_{ij} is negative binomial (α, $p_{ij} = \frac{\theta_{ij}}{1+\theta_{ij}}$), with

$$E(X_{ij}) = \alpha\ \theta_{ij} = m_{ij}$$
$$\mathrm{Var}(X_{ij}) = \alpha\ \theta_{ij}(1+\theta_{ij}) = m_{ij}\ (1 + \frac{m_{ij}}{\alpha}\ ).$$

He suggests starting the analysis by testing the hypothesis whether the distribution is Poisson, and then fitting the full factorial model based on the

Poisson distribution for X_{ij}. If the hypothesis is rejected, a choice between model (1) and model (2) can be based on a plot of $(X_{ijk} - \widehat{m}_{ij})/\sqrt{\widehat{m}_{ij}}$ against the estimates $\widehat{m}_{ij} = \overline{X}_{ij+}$ of m_{ij} under the full Poisson model. If the variance of these quantities is more or less constant, model (1) is preferred; if it increases with $\widehat{m}_{ij}$, model (2) is more suitable. Of course, it is possible that neither model (1) nor model (2) is satisfactory in the case of extra-Poisson variation.

Consul and Famoye (1992) proposed another way to represent the overdispersed Poisson distribution. The generalized Poisson distribution fits over- or under- dispersed count data. Consul and Famoye (1992) defined the random variable Y to have a generalized (Lagrangian) Poisson distribution (GPD) if its probability distribution is given by

$$P(Y = y) = \begin{cases} \theta(\theta + \lambda y)^{y-1} \exp\{-(\theta + \lambda y)\}/y!, \text{ for } y = 0, 1, 2, \ldots \\ \qquad 0; \text{ for } y > m \text{ when } \lambda < 0 \end{cases}$$

and zero otherwise, where $\theta > 0$, $\max(-1, -\theta/4) \leq \lambda \leq 1$ and m is the largest positive integer for which $\theta + m\lambda > 0$ when λ is negative. The GPD reduces to the Poisson model when $\lambda = 0$ and possesses the property of overdispersion for all values of $\lambda > 0$ and the property of underdispersion for all values of $\lambda < 0$. The mean μ for GPD is

$$\mu = \theta\,(1-\lambda)^{-1} = \theta\,\rho.$$

One can write the corresponding generalized Poisson regression (GPR) in the form

$$P(Y = y \mid \underline{x}\,) = \begin{cases} \mu\,[\mu + (\rho-1)y]^{y-1}\rho^{y-1} \exp\{-[\mu + (\rho-1)y]\}/y!,\ y = 0,1,2,\ldots \\ \qquad 0; \text{ for } y > m \text{ when } \rho < 1. \end{cases}$$

The mean is $E(Y \mid \underline{x}) = \mu(\underline{x})$ and the variance is $Var(Y \mid \underline{x}) = \rho^2 \ \mu(\underline{x}) > 0$. When $\rho = 1$, P reduces to the Poisson distribution; when $\rho > 1$, P is the overdispersed Poisson; and for $1/2 \leq \rho < 1$, P is underdispersed when $\mu > 2$.

They used the method of maximum likelihood and moments for the estimation of parameters. They also derived asymptotic tests for the adequacy of the model and for the significance of regression parameters.

C. The Overdispersed Multinomial Distribution

The Dirichlet-multinomial model has been used to model the overdispersed multinomial distribution. Paul et al. (1989) compared test statistics for the multinomial (binomial) assumption against Dirichlet-multinomial (beta-binomial) alternatives.

The tests they considered were the likelihood ratio and $C(\alpha)$ test statistics for goodness-of-fit. Since a boundary problem existed with the asymptotic null distribution of the likelihood ratio statistic, they used a 50:50 mixture of zero and chi-square with 1 degree of freedom of the null distribution of likelihood ratio statistic. They also derived the $C(\alpha)$ test for goodness-of-fit of the multinomial distribution against the Dirichlet-multinomial alternatives and showed that the statistic is the generalization of the $C(\alpha)$ goodness-of-fit of the binomial distribution against beta-binomial alternatives. The distribution of the $C(\alpha)$ test approximates the statistic Z to the standard normal distribution.

Having done Monte-Carlo simulations, they found that (1) the empirical significance levels for the likelihood ratio test were significantly lower than the

normal level when the sample sizes was less than 40, even if the correct critical value was used. On the other hand, the $C(\alpha)$ test performed pretty well except when both the size and the underlying probability were small; and (2) the two tests had very similar powers over the range of conditions used in the simulations.

The $C(\alpha)$ statistic is similar to the T_k statistic derived by Kim and Margolin (1992), which is a Pearson chi-square statistic, but with altered weights. Kim and Margolin (1992) presented an improved approximate sampling distribution for the corresponding $C(\alpha)$ test statistic. The elements of the improvement included better control of test size and increased power against overdispersion. Kim and Margolin (1992) also considered the Pearson chi-squared test for the detection of overdispersion, obtained the asymptotic efficiency relative to T_k, and showed that it is always less than or equal to 1. In their Monte Carlo study, they observed that (1) under the null hypothesis, the $C(\alpha)$ is too conservatively approximated by a standard normal, whereas the test based on T_k and Pearson chi-squares exhibit satisfactory tail area behavior; and (2) the gain of power of T_k relative to the Pearson chi-square test is small unless there is substantial inequality among the Dirichlet-multinominal sample sizes.

D. The Mixture of Exponential Families

Dean and Lawless (1989) developed tests for detecting extra-Poisson variation in Poisson regression models. They obtained the tests as score tests against arbitrary mixed Poisson alternatives, and provided approximations for computing the significance level and power of the tests against the negative binomial alternatives. The mixed model they used was an extension of that of

Collings and Margolin's (1985) to general regression situations. The model assumed that the random effects had finite first and second moments. If the random effects follow a gamma distribution, then the random variable has a negative binomial distribution.

Dean and Lawless's (1989) findings were that for detecting overdispersion, tests analogous to those based on their models were superior to tests based on the Pearson statistic for more general regression situations, at least for moderate amounts of overdispersion. But, as they pointed out, the test was designed to be particularly effective against one type of extra-Poisson variation.

Later on, Dean (1992) generalized the idea to the overdispersed exponential family. Dean (1992) derived a method for obtaining tests for overdispersion with respect to a natural exponential family. He claimed that the tests were powerful against arbitrary alternative mixture models where only the first and second moments of the mixed distribution were specified. He used the natural exponential family with a probability density function

$$f(Y;\theta) = \exp\{a(\theta)Y - g(\theta) + C(Y)\}, \tag{1.1}$$

where Y represents the response variable and θ is an unknown parameter on which the distribution of Y depends. Let $Y_1, \ldots, Y_n$ be a sample of independent observations, where the Y_i's are then from a natural exponential family, with corresponding θ_i a function of a $p \times 1$ vector of covariates $\underline{x}_i$ and regression parameters $\underline{\beta}$; that is, $\theta_i = \theta_i(\underline{x}_i;\underline{\beta})$, $i = 1, \ldots, n$. The mean and variance of Y_i are

$$E(Y_i) = \mu_i(\theta_i) = \{a_i'(\theta_i)\}^{-1} g_i'(\theta_i)$$

$$Var(Y_i) = \sigma_i^2(\theta_i) = \{a_i'(\theta_i)\}^{-2}\{g_i''(\theta_i) - a_i''(\theta_i)E(Y_i)\},$$

where the (′) denotes differentiation with respect to θ. Then, Dean (1992) constructed a large 'overdispersed' family for which $Var(Y_i) > \sigma_i^2$, with equality when the distribution is equation (1.1). From this large family he constructed a score test for the hypothesis that equation (1.1) is adequate. To construct the extended family, he let the density of Y_i given θ_i^* be $f(Y_i \mid \theta_i^*)$, as given by the right-hand side of (1.1), where the θ_i^*'s are continuous independent random variates with finite mean and variance.

$$E(\theta_i^*) = \theta_i(\underset{\sim}{x}_i \,;\, \underset{\sim}{\beta}), \;\; Var(\theta_i^*) = \tau b_i(\theta_i) \;> 0. \tag{1.2}$$

He assumed that $\mathrm{E}\{(\theta_i^* - \theta_i)^\gamma\} = \alpha_\gamma$, $\alpha_\gamma = o(\tau)$, $\gamma \geq 3$. When $\tau \to 0$, this model reduces to $f(Y_i; \theta_i)$. Similarly to Cox's (1983) method, he expanded $f(Y_i; \theta_i^*)$ in a Taylor series about θ_i and took expectations of θ_i^* to obtain

$$f_M(Y_i) = f(Y_i; \theta_i) \left\{ 1 + \sum_{r=2}^{\infty} \frac{\alpha_r}{r!} D_r(Y_i \,;\, \theta_i) \right\},$$

where

$$D_r(Y_i \,;\, \theta_i) = \left\{ \frac{\partial^{(r)}}{\partial \theta_i^{*(r)}} f(Y_i \,;\, \theta_i^*) \mid_{\theta_i^* = \theta_i} \right\} \{f(Y_i \,;\, \theta_i)\}^{-1}.$$

Hence, the contribution to the log-likelihood function from the ith observation is

$$l_i = \log f(Y_i \,;\, \theta_i) + \log\left\{ 1 + \sum_{r=2}^{\infty} \frac{\alpha_r}{r!} D_r(Y_i \,;\, \theta_i) \right\}.$$

He derived the score test statistic to test the hypothesis $\tau = 0$ along the line of Breslow's (1989) suggestion to replace the second derivatives by their expectation under the null hypothesis before evaluation.

Dean (1992) applied the method to the following three types of overdispersed Poisson models:

a) $\theta_i = \log \mu_i = \underline{x}_i^T \underline{\beta} + z_i$, and $b(\theta_i) = 1$, where z_i are iid random variables with $E(z_i) = 0$ and $Var(z_i) = \tau < \infty$, $i = 1, \ldots, n$.

b) the multiplicative random effects model:

$\theta_i = \mu_i$, $\theta_i^* = \nu_i \mu_i$, $E(\nu_i) = 1$, $Var(\nu_i) = \tau < \infty$, where the ν_i are iid positive variates, given that the fixed effects and the random effects, the ν_i's, are added when $\mu_i = \exp(\underline{x}_i^T \underline{\beta})$; that is in the log linear model.

c) an alternative model with a simple variance inflation:

$\theta_i = \mu_i$, $\theta_i^* = \nu_i \mu_i$, $E(\nu_i) = 1$, $Var(\nu_i) = \tau/\mu_i$, $\tau < \infty$, where the ν_i's are independent positive random variates.

The tests for these three types of extra-Poisson models are listed in Table 1.2.

Similarly, he applied the technique to extra-binomial variation and obtained the score statistics for various situations. He adjusted the test statistics using the leverage matrix for regression, so that they converge to normality quickly. The $\widehat{h}_{ii}$'s in Table 1.2 are used for this purpose. Here $\widehat{h}_{ii} = h_{ii}(\widehat{\beta})$, where $h_{ii}(\beta)$ is the ith diagonal element of the matrix $\boldsymbol{H} = W^{1/2}(X^T W X)^{-1} X^T W^{1/2}$, with $W = \text{diag } (\mu_1, \ldots \mu_n)$ and $X = X(\beta)$ being an $n \times p$ matrix with ijth entry $\mu^{-1}(\partial \mu_i / \partial \beta_j)$, $i = 1, \ldots, n$; $j = 1, \ldots, p$. Since Dean (1992) mixed the natural exponential family with the distribution whose first two moments are defined, some information on the distribution is lost.

Table 1.2. Tests for Different Types of Extra-Poisson Variation

	The Over-dispersion model		
	(1)	(2)	(3)
θ	$\log\mu$	μ	μ
$a(\theta)$	θ	$\log\theta$	$\log\theta$
$g(\theta)$	e^{θ}	θ	θ
$b(\theta)$	1	θ^2	θ

Score test statistic — Adjusted statistic

Model(1) $P_A = \dfrac{\sum\{(Y_i-\hat{\mu}_i)^2-\hat{\mu}_i\}}{\{2\sum\hat{\mu}_i^2\}^{1/2}}$; $P'_A = \dfrac{\sum\{Y_i-\hat{\mu}_i)^2-(1-\hat{h}_{ii})\hat{\mu}_i\}}{\{2\sum\hat{\mu}_i^2\}^{1/2}}$

Model(2) $P_B = \dfrac{\sum\{(Y_i-\hat{\mu}_i)^2-Y_i\}}{\{2\sum\hat{\mu}_i^2\}^{1/2}}$; $P'_B = \dfrac{\sum\{(Y_i-\hat{\mu}_i)^2-Y_i+\hat{h}_{ii}\hat{\mu}_{ii}\}}{\{2\sum\hat{\mu}_i^2\}^{1/2}}$

Model(3) $P_C = \dfrac{1}{\sqrt{2n}}\sum\left\{\dfrac{(Y_i-\hat{\mu}_i)^2-Y_i}{\hat{\mu}_i}\right\}$; $P'_C = \dfrac{1}{\sqrt{2n}}\sum\left\{\dfrac{(Y_i-\hat{\mu}_i)^2-Y_i+\hat{h}_{ii}\hat{\mu}_i}{\hat{\mu}_i}\right\}$

Type of over-dispersion

Model(1): $E(Y_i)\simeq\mu_i$, $Var(Y_i)\simeq\mu_i(1+\tau\mu_i)$ for τ small.

(2): $E(Y_i)=\mu_i$, $Var(Y_i)=\mu_i(1+\tau\mu_i)$.

(3): $E(Y_i)=\mu_i$, $Var(Y_i)=\mu_i(1+\tau)$.

E. The Double Exponential Families

Efron (1986) developed double exponential families to generalize any exponential family regression models, especially in binomial and Poisson regressions. The double exponential families allow the introduction of a second parameter that controls variance independently of the mean. The ordinary one-parameter exponential family of density functions can be written as

$$g_{\mu,n}(y) = e^{n[\eta y - \psi(\mu)]}\,[dG_n(y)], \tag{1.3}$$

where μ is the expectation parameter, $\mu = \int_{-\infty}^{\infty} y\ g_{\mu,n}(y)\ dG_n(y)$; y is the natural statistic; η is the natural, or canonical parameter, a monotone function of μ; $\psi(\mu)$ is a normalizing function, chosen to make the density integrate to 1; and n is the sample size.

Given an exponential family (1.1), the family of density functions

$$\widetilde{f}_{\mu,\phi,n}(y) = c(\mu,\phi,n)\ \phi^{1/2}\ \{g_{\mu,n}(y)\}^{\phi}\ \{g_{y,n}(y)\}^{1-\phi}[dG_n(y)] \tag{1.4}$$

is called the double exponential family, with parameters μ, ϕ and n. The constant $c(\mu,\phi,n)$ is defined to make $\int_{-\infty}^{\infty} \widetilde{f}_{\mu,\phi,n}\ (y)\ dG_n(y) = 1$. Using the density (1.2) as a constituent of a regression analysis, we can estimate the unknown parameters μ and ϕ from the data (Efron 1986). Since Efron (1986) proved that the constant $c(\mu,\phi,n)$ in (1.4) is nearly equal to 1, and the Kullback-Leibler distance for $g_{\mu,n}$ is

$$\mathrm{I}_n(\mu_1,\ \mu_2) = E_{\mu,n} \log\,[g_{\mu_1,n}\ (y)\ /\ g_{\mu_2,n}(y)] = n\ \mathrm{I}(\mu_1,\ \mu_2),$$

(1.4) can be rewritten as

$$f_{\mu,\phi,n}(y) = \phi^{1/2}\ g_{y,n}(y)\ \exp\,[\,-2n\phi\ \mathrm{I}(y,\ \mu)]\ .$$

The log-likelihood function for the double exponential family with an approximation to the normalizing constant (or, equivalently, to the extended quasi-likelihood function) is $l(\mu,\ \phi;\ y) = \sum (1/2)\ [\ln(\phi_i) - 2\ \phi_i I(y_i,\ \mu_i)]$. For a Poisson count, $I(y_i,\ \mu_i) = 2\ \{y_i \log(y_i/\mu_i) - (y_i - \mu_i)\}$. The estimates $\widehat{\mu}$ and $\widehat{\phi}$ that

maximize the log likelihood function may be obtained iteratively: At each iteration the parameters in the mean are estimated by maximum quasi-likelihood with weight vector $\widehat{\phi}$, and then the parameter estimates in ϕ are updated by maximization of $l(\widehat{\mu}, \widehat{\phi}; y)$. The likelihood ratio test statistic based on the double exponential family distribution is DLR $= 2[\, l(\widehat{\mu}, \widehat{\phi}\, ; y) - l(\widehat{\mu}_0, \widehat{\phi}_0\, ; y)]$, where $\widehat{\mu}$ and $\widehat{\phi}$ are the maximum likelihood estimates under the full model and $\widehat{\mu}_0$ and $\widehat{\phi}_0$ are the estimates under the reduced model with $\alpha = 0$.

The double exponential family score test statistic is DS $= (2\overline{D}^2)^{-1} \sum \widehat{D}_i\, z_i$ $(z_i z_i')^{-1} \sum \widehat{D}_i\, z_i'$, where $\sum z_i = 0$ (defined by Ganio and Schafer (1992)), and $\widehat{D}_i = D\ (y_i,\ \widehat{\mu}_i)$ is the ith component of the deviance statistic for the fit to the reduced model and $\overline{D}$ is the average of the $n\widehat{D}_i$'s. The null asymptotic distribution of DS is chi-squared with q degrees of freedom if the double exponential family model is correct.

Ganio and Schafer (1992) point out that in some cases the double exponential model may not be realistic. For example, if Y/m is a double binomial proportion with mean P, then $Var\ (Y/m)\ = P(1-P)/(m\phi)$ must be small when m is large. Although there is good reason to suspect that extra-binomial variation is most consequential when m is large, the double-binomial model for variance may not be realistic. The other drawback to the likelihood methods is that convenient probability distributions incorporating the mean-variance relationship, such as those from the double exponential family, may be unrealistic (Ganio and Schafer, 1992). In addition, it is likely that inference regarding parameters in the variance is not robust--at least for variance tests from the normal distribution (Ganio and Shafer, 1992).

Following the idea of the double exponential model (Efron, 1986), Ganio and Schafer (1992) proposed diagnostic tools for assessing the dependence of extra-Poisson variation on explanatory variables and for comparing several common models for overdispersion. These tools were based on likelihood ratio and score tests for regression terms in the dispersion parameter of the generalized linear model, using double exponential family and 'pseudolikelihood' formulations.

Ganio and Schafer (1992) took some function of the dispersion parameter in a generalized linear model to depend on explanatory variables. Suppose that Y_1, $Y_2, \cdots, Y_n$ are independent response variables, with

$E(Y_i) = \mu_i = h(\eta_i), \qquad \eta_i = x_i'\beta,$

$Var(Y_i) = V(\mu_i)/(a_i\phi_i), \quad \phi_i = g(\gamma_i), \quad \gamma_i = \lambda + z_i'\alpha$, and

where x_i and β are $p \times 1$ vectors of known explanatory variables and unknown parameters, $h(\ )$ is a known monotonic differential function, $V(\ )$ is a known positive function, the a_i's are known constants, λ is a scale parameter, and z_i's are taken to be centered such that $\sum z_i = 0$. If $\alpha = 0$, then ϕ_i is a constant dispersion parameter and $E(Y_i)$ and $Var(Y_i)$ define a generalized linear model.

As an alternative, the method of moments may be used to estimate the parameters in the variance after the mean parameters have been estimated by maximum quasi-likelihood (Breslow, 1984). Ganio and Schafer (1992) called it the pseudolikelihood function for the model of interest. The function is

$$l_p(\mu, \phi; y) = \sum (1/2)[\ln(\phi_i) - \phi_i R(y_i, \mu_i)], \text{ where } R(y_i, \mu_i) = (y_i - \widehat{\mu}_i)^2 / [V(\widehat{\mu}_i)/a_i].$$

Then, they derived likelihood ratio and score tests to test the hypothesis $\alpha = 0$ in the model specified above. The pseudolikelihood ratio test statistic is PLR $= 2$ [$l_p(\widehat{\mu},\ \widehat{\phi}\ ;\ y) - l_p(\widehat{\mu}_0,\ \widehat{\phi}_0;\ y)$]. The score test statistic for the pseudolikelihood is PS $= (2\overline{R}^2)^{-1} \sum \widehat{R}_i z_i\ (\sum z_i z_i')^{-1} \sum \widehat{R}_i z_i'$, where $\widehat{R}_i = R(\ y_i\ ,\ \widehat{\mu}\ )$, the ith component of the Pearson goodness of fit statistic for the reduced model with $\alpha = 0$, and $\overline{R}$ is the average of these. This is of the same form as DS, but with the deviance components replaced by Pearson components.

Ganio and Schafer (1992) used the F test for overall significance in the regression of unweighted least squares fit of squared residuals (deviance (DSF) or Pearson (PSF)) on the suspected explanatory variable to explore the necessity of regression terms in the dispersion parameter. DSF is defined as $(\mathrm{SSR}_z/q)/\mathrm{MSE}_z$, where q is the dimension of z_i and SSR_z and MSE_z are the regression sum of squares and residual mean squares from the least squares fit to the regression of the squared deviance residual on $\underset{\sim}{z}_i$. The denominator is an empirical rather than model-based estimate of the variance of squared residuals. PSF is the F statistic from using Pearson residuals rather than deviance residuals (Ganio and Schafer, 1992). Applying the methods to eight different types of simulated data, they found that DSF and PSF are more robust than DS and PS, but slightly conservative.

Despite the problem in the double exponential family, Nelder and Lee (1992) found the extended quasi-likelihood estimator to be generally superior to the maximum pseudolikelihood estimator in several simulation studies. As a possible explanation, Nelder and Lee (1992) suggested that deviance residuals generally are more normally distributed than Pearson residuals, therefore, the log

likelihood ratio based on the double exponential family should be more appropriate than the log likelihood ratio based on pseudolikelihood.

1.2.2 Representation of Overdispersion by an Extra Dispersion Parameter

Quasi-likelihood has provided a useful tool for analyzing overdispersed data, since it needs only specification the mean and the relationship between the mean and variance of the observations (Wedderburn, 1974). For a one-parameter exponential family, the log likelihood is the same as quasi-likelihood. For example, to model the extra-Poisson variation, the connection between the mean and variance of a frequency N can be expressed by

$$V(N) = c_1 E(N) + c_2\{E(N)\}^2, \text{ or}$$

$$V(N) = c\{E(N)\}^b, \qquad 1 < b < 2.$$

Then, we can extend generalized linear models (GLM) to fit these models by allowing a variable variance, for example, $V(y) = \phi\mu$, or $V(\mu) = \mu^\alpha$ (Paul and Plackett, 1978).

A. Breslow's Method

Considering two types of variance structure, Breslow (1984) modified an iterative reweighted least squares scheme to accommodate the extra-Poisson variation when fitting log-linear models to tables of frequencies or rates.

Suppose one observes counts d_i, given p explanatory variables arranged in row vectors x_i, and possibly also fixed denominators n_i for $i = 1, \ldots, N$. Let λ_i denote the unknown rates or frequencies and assume that, conditional on λ_i and n_i, the d_i have independent Poisson distributions with mean $E(d_i|\lambda_i) = \lambda_i n_i$. Further suppose that $\log(\lambda_i) = x_i\beta + \epsilon_i$, where β is a column vector of p unknown

regression parameters and the ϵ_i are random error terms having means 0 and a constant unknown variance σ^2.

Consider first a weighted least squares fit to the transformed observation $y_i = \log(d_i/n_i)$. Provided that the d_i are reasonably large, the y_i may be regarded as having an approximately normal distributions with means $x_i\beta$ and variance $\sigma^2 + \tau^2$, where $\tau^2 = 1/E(d_i)$ is estimated by $1/d_i$. This suggests estimating σ^2 and carrying out the least square analysis using the empirical weights $w_i = (\sigma^2 + \widehat{\tau}^2)^{-1}$. If the weights were correctly chosen to be the true inverse variances, one would have in expectation

$$\sum_{i=1}^{N} \frac{(y_i - x_i\widehat{\beta})^2}{(\sigma^2 + \tau_i^2)} = N - p. \tag{1.5}$$

Equation (1.5) may be solved recursively by rewriting it as

$$\sigma^2 = \frac{1}{N-p} \sum_{i=1}^{N} \frac{(y_i - x_i\widehat{\beta})^2}{\{1 + (\sigma^2 d_i)^{-1}\}}.$$

If the generalized linear model is defined as having $E(d_i) = \mu_i \cong \exp(\log\, n_i + x_i\beta)$ and $Var(d_i) = E\{Var(d_i|\lambda_i)\} + Var\{E(d_i|\lambda_i)\} \cong \mu_i + \sigma^2\mu_i^2$, the relationship in (1.5) can be rewritten as

$$\sum_{i=1}^{N} \frac{(d_i - \widehat{\mu}_i)^2}{\{\widehat{\mu}_i(\widehat{\mu}_i\sigma^2 + 1)\}} = N - p. \tag{1.6}$$

As before, the new value of σ^2 can be estimated by

$$\sigma^2 = \frac{1}{N-p} \sum_{i=1}^{N} \frac{(d_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i(\widehat{\mu}_i + \sigma^{-2})} .$$

The moment estimator based on (1.6) corresponds to intuition since at convergence the residual sum of squares equals its degrees of freedom. Procedures based on equation (1.6) seem most appropriate when the expected frequencies are small. An alternate approach for small frequencies would be to specify a parametric family of error distributions, for example, if one takes λ_i to have the gamma distribution, then the observed frequencies d_i are negative binomial. Such parametric procedures have an advantage over the moment estimators proposed by Breslow (1984) for σ^2 in that they facilitate the assessment of the statistical significance of the extra-Poisson variation. On the other hand, they require explicit assumptions about the nature of the error distribution, and thus lose generality.

Breslow (1989) subsequently developed the score statistics for testing the mean and the variance using the methods of quasi-likelihood, with variance parameters estimated either by moments or by pseudo-likelihood. That provides a simple approach to overdispersion in the Poisson regression and other generalized linear models (Breslow, 1989). Although less efficient than full maximum likelihood when the overdispersed model holds, quasi-likelihood provides consistent estimates of the mean even if the variance function is misspecified (Breslow, 1989).

Since the score test on the variances involves estimation of third and fourth moments, this statistic and the variance estimate are highly unstable in small samples (Breslow, 1989). Therefore, Breslow (1989) suggested replacing estimates

of the elements of the information matrix by their expectations under the null hypothesis before evaluating the information matrix.

Later, Breslow (1990) developed test statistics for evaluating the significance of added variables in a regression equation for the mixed Poisson models, where the structural parameter ϕ that determines the mean-variance relationship $V(\mu;\ \phi) = \mu + \phi\mu^2$ is estimated by the method of moments and regression coefficients are estimated by quasi-likelihood. He investigated two versions of the Wald and score tests-- one calculated from the usual model-based covariance, and the other using an "empirical" covariance matrix. His findings were that (1) the simpler Poisson analysis produces approximately unbiased regression coefficients, even though the overdispersion is not accounted for; (2) although tests and standard errors based on Poisson theory are seriously in error in the presence of overdispersion, the empirical standard error and especially the empirical score test obtained in conjunction with the Poisson analysis perform reasonably well provided the sample size is sufficiently large; and (3) the empirical standard error and score test perform less well in small samples than do the model-based quasi-likelihood moment procedures, probably because of the lack of precision in the empirical variance.

1.2.3 Other Methods

A. Residual Plots as a Diagnostic Tool

Using residual diagnostics is another way to check overdispersion. Liang and McCullagh (1993) examined a series of examples with overdispersion drawn from the literature via the residual analysis and a formal comparison. They fitted

the standard models without overdispersion, then plotted the standardized residuals against litter size, since they believed that some of the overdispersion was related to litter size. They also developed a formal test to check which of the two commonly used dispersion models, the beta-binomial variance or the constant dispersion factor relative to the binomial, is better for describing overdispersion. The statistic used is the ratio of the squared residual of the 'well-fitting' model to the nominal variance. The approximated distribution is a $\sigma^2\chi^2$, where σ_i^2 satisfies the linear model

$\sigma_i^2 = \alpha + \beta\,(m_i - 1)$, where m_i is the litter size.

The methods presented in their paper are useful when the number of responses is small or modest. They recommended that both methods be adopted to strengthen the conclusion when examining the adequacy of the variance expression for overdispersion. However, we need to be cautious when applying the methods to small litter sizes, which could have skewed residuals.

B. Errors-in-x Regression

Adding covariates may reduce the overdispersion. Zeger and Edelstein (1989) assumed that overdispersion comes from errors in the x variable. They fitted a Poisson regression model with a surrogate x variable to help assess the efficacy of vitamin A in reducing child mortality in Indonesia. They developed a parametric error-in-x regression model for this purpose, then applied it to a community study where the mortality rate in villages receiving supplemental vitamin A was 35% less than in control villages. However, at the baseline time, the control villages were found to have slightly higher rates of xerophthalmia, an ocular disease caused by vitamin A deficiency. Therefore, they wanted to adjust

the mortality comparison for differences in baseline vitamin A levels as indicated by xerophthalmin prevalence. They assumed the number of deaths among preschoolers in the villages, given a random effect, had a Poisson distribution. Then, they used the second outcome, which was the number of cases of xerophthalmia in the villages at the baseline, as a surrogate for the random effect, which was the true health status at the baseline. One limitation of the model is that it can adjust for only one random covariate at a time. Extensions of this are conceptually feasible, but computationally difficult.

1.2.4 Overdispersed Survival Data

Overdispersion has also been reported in survival data. Models dealing with overdispersed failure time must include censoring, a unique characteristic of survival data, to be useful. It is common that at the close of life-testing experiments on industrial reliability, not all components may have failed (Cox and Oakes, 1988). In clinical trials, some patients (many, it is to be hoped) will survive to the end of the clinical trials or may be lost to be follow-up. Efforts have been made in modeling the heterogeneity in survival time as well as in hazard function.

A. Modeling Mortality as a Function of Time

Overdispersion is also found in mortality data measured at different time points. O'Hara Hines and Lawless (1993) applied survival analysis and generalized linear models to analyze the results of toxicity experiments in which different treatments were applied to groups of animals, and the resulting mortality in each group was measured at a number of discrete time points over the course of the

experiments. They incorporated several models for overdispersion into the generalized linear model framework for multinomial data to model the cumulative mortality as a function of time and the covariates representing experimental conditions. They did not allow censoring in the model.

Following Breslow's (1984, 1990) idea, O'Hara Hines and Lawless (1993) estimated the regression coefficients by quasi-likelihood and estimated the overdispersion parameter by the method of moments. They reached a conclusion similar to that of Breslow (1990) that a generally satisfactory regression analysis can be obtained by using the multinomial model estimating equations for regression coefficients combined with robust variance estimators. The concern they shared with Breslow (1990) is that hypothesis testing employing robust variance estimates has a lower power than the tests using appropriate model-based variance estimates.

Keiding et al. (1990) presented a case of modeling the excess mortality of Danes unemployed at the time of the census of 1970, who were followed over the next 10 years. In testing the hypothesis of non-differential excess mortality across mortality groups, Keiding et al. (1990) used two approaches: (1) including of further covariates (this was partly successful, but led to a fairly complicated model); and (2) modeling the remaining heterogeneity as random. They modeled the mortality rate as a function of the survival time and other covariates, then fitted a power transformation model which had a similar variance structure to that used by Breslow (1984).

One interesting problem Keiding et al. (1990) dealt with was not having information about individual times at risk. They chose two approximations: (1)

that the individual times at risk are all identical within the stratum, and (2) the times at risk within the stratum are generated from independent, identically distributed exponential random variables. Keiding et al. (1990) fitted the models for both collective and individual frailty, which is an unobservable random variable acting multiplicatively on the hazard. They found that the individual frailty model seemed to be more satisfactory because it modeled the heterogeneity explicitly among individuals. They did not consider censoring in their models.

B. Modeling Overdispersed Failure Time

Modeling survival time is one of the major interests in survival analysis. The fact that individuals differ substantially in their endowment for longevity is well known. There are two ways of looking at the issue. One is to consider the existence of heterogeneity in failure times. Another is to focus on the heterogeneity in the hazard function describing factors influencing the force of mortality.

i. Models for Overdispersed Failure Times

Exponential distributions have been widely used to model survival time in previous studies. That the distribution has only one adjustable parameter often means that methods based on it are rather sensitive to even modest departures of the variance (Cox and Oakes, 1988). Recently, lots of discussions on the distributions of failure time have focused on their relations to the exponential distribution, in particular whether they are over- or under- dispersed relative to the exponential family (Cox and Oakes, 1988). For example, if each individual survival time is exponentially distributed but the rate varies randomly between individuals, and if the rate has a gamma distribution, the unconditional

distribution becomes the Pareto distribution, which is overdispersed relative to the exponential distribution.

Mendenhall and Hader (1958) considered a failure population which can be divided into subpopulations, each representing a different type or cause of failure. They obtained estimates of population parameters for the case where the subpopulation failure times were exponentially distributed and sampling was censored at a predetermined test termination time. Their primary attention was directed towards the case of two populations of failure, each exponentially distributed. Then they extended the estimation to the case of any number of failure subpopulations, each distributed according to a Weibull distribution. They concluded that the maximum likelihood estimation procedure appeared to give satisfactory results when the sample size was sufficiently large and the test termination time was large relative to the average lifetime of each subpopulation. When the sample size and the termination time were small, the estimates were badly biased and had large variances.

Jewell (1982) considered arbitrary nonparametric mixtures of the exponential and the Weibull (fixed shape) distributions for a lifetime, which was characterized as a Laplace transformation. He investigated the maximum likelihood estimates of the mixing distribution and found it was supported on a finite number of points. However, the convergence of the algorithm described by Jewell (1982) could be very slow. Typically, more than 100 iterations were needed for convergence to 3 decimal places.

ii. Models for Heterogeneous Hazards (Frailty Models)

Studies have been done on the effects on human survival of either differences in individual susceptibility to specific causes of death or differences in general susceptibility to all causes of death. Vaupel et al. (1979) modified standard life table methods to construct cohort, period and individual life tables, and to explore the impact of heterogeneity in the force of mortality on the dynamics of total mortality. They introduced a multiplicative frailty z to the force of mortality so that the modified force of mortality μ of an individual in population group i at exact age x, at some instant in time y, and with a frailty z is then

$$\mu_i(x, y, z) = z \times \mu_i(x, y, 1).$$

An individual with a frailty of 1 might be called a "standard" individual. Vaupel et al. (1979) assumed that the frailty of an individual at birth had a gamma distribution, then constructed cohort, period, and individual life tables. In doing this, they corrected the over-estimation of life expectancy and the potential gain in life expectancy from health and safety interventions by standard life table methods.

Hougaard (1984) generalized the frailty distributions by means of a Laplace transformation and then applied them into nonnegative exponential families having the frailty as a canonical statistic. He examined the consequences of different distributions, such as the gamma, uniform, Weibull and lognormal, with a special interest in the inverse Gaussian distribution. He found that the latter made the population homogeneous over time, whereas with the gamma distribution the relative heterogeneity was constant, independent of age. He also

discussed the model for competing risks by using cause-specific frailty.

Later, Hougaard (1986) proposed a three-parameter family of distributions on the positive numbers for heterogeneous populations. Similarly, the family was characterized by a Laplace transformation. This time Hougaard (1986) focused on so-called stable distributions. This family has a natural exponential distribution in one of the parameters. He derived the moments, convolution, infinite divisibility, unimodality and other properties of the family. Although we cannot tell which of the distributions is the correct explanation of heterogeneity, the methods demonstrate that it is possible in practice to analyze models formulated after an explicit consideration of heterogeneity.

To include both observable and unobservable factors, Heckman and Singer (1982) proposed a nonparametric maximum likelihood estimator for the distribution of an unobservable frailty which simultaneously estimates parameters in a wide class of conditional duration distributions, given observed covariates and unobservable frailty. Although they did not specify the distribution of the duration, they expressed it as a member of the exponential family. They provided an iterative procedure to find a local maximum for the log likelihood function. They advised running the EM algorithm from a variety of starting values to guard against failure in locating a global optimum. They demonstrated that the estimator was consistent for both the distribution of unobservable frailty and the coefficient vectors associated with the covariates. However, they also stated that the distribution function of the unobservable frailty was difficult to estimate. Though it is feasible to consider censoring, the computations are very intensive. Hence, they did not impose censoring or truncation on the model.

Heckman et al. more recently (1990a) developed a nonparametric method for testing the hypothesis that duration can be represented by a mixture of the exponential distribution and unobservable frailty. They developed (1) both Bayesian and classical finite sample tests, (2) a nonparametric test for the presence of a mover-stayer model for the class of discrete mixture models; and (3) a consistent estimator for the number of points of support of a discrete mixture of exponential models.

Heckman and Walker (1990b) applied the methods to estimating fecundability from data on waiting times to first conception. They reanalyzed the data on Hutterite women used by Sheps (1973) and reached the following conclusion: (1) the data are consistent with the hypothesis that the fecundability of each woman is constant; (2) the data cannot distinguish between a mixture of the exponential models with other models consistent with population homogeneity and a decline in fecundability for each woman with time at risk; and (3) the data are consistent with a "mover-stayer" model, where the stayer proportion is interpreted as a population sterility proportion.

Andersen et al. (1993) reviewed frailty models with an unobservable random variable acting multiplicatively on the intensity of the hazard. They pointed out the limitations in these models: (1) the frailty variable did not depend on time; (2) only one frailty variable was allowed even though an individual may experience several types of frailty; and (3) the frailty was modeled parametrically everywhere. In fact they assumed the frailty to have a gamma distribution with an unknown (inverse) scale parameter and an unknown shape parameter.

iii. The Beta-Geometric Distribution

The beta-geometric distribution has been applied in studies of human fecundability, the time required to achieve pregnancy (Suchindran, 1972; Weinberg and Gladen, 1986). If all noncontracepting, sexually active couples had the same per-cycle conception probability, then the number of cycles required to achieve pregnancy would have the geometric distribution. In fact, there is ample evidence that couples vary in their fecundability. Weinberg and Gladen (1986) therefore considered the per-cycle conception probability as fixed over time for a given couple but varying across couples according to some underlying distribution. If the beta distribution is chosen, then the duration of cycles to pregnancy has a beta-geometric distribution.

Heckman and Walker (1990b) developed test criteria to determine whether or not data on waiting time to conception can be represented by a mixture of the exponential models or, in discrete time, by a mixture of geometric models. They could not reject the hypothesis that a mixture of geometric model explained the Hutterite data. On the other hand, they found a variety of models fit the data in the sense of passing the conventional goodness-of-fit tests. Nonetheless, they concluded that the true model was a mixture of exponential model or in discrete time, a mixture of geometric model because of the simplicity of the mixture of exponential models and the biological plausibility of a constant (monthly) hazard for time to first conception (given fecundity).

1.3 Goals of This Research

Survival data are distinguished from most other types of data by the widespread occurrence of censoring. Censoring occurs when the outcome of a particular unit (patient or component) is unknown at the end of the study. Thus we may know only that a particular patient was still alive six months into the study, but the exact failure time is unknown either because the patient withdrew from the study or because the study ended while the patient was still alive. Censoring is common in medical experiments, so estimation methods must allow for it if they are to be generally useful.

The review of research above shows that many models and test statistics for overdispersion do not deal with censored survival time. In the following chapters, this research will modify the model scheme and test statistics for the overdispersed exponential family to account for data with censoring, and illustrate the use of these tests in analytical and numerical applications to real-world data.

In Chapter 2, we derive the test statistics to detect the existence of overdispersion based on two models. The first one is the mixture of exponential families (Dean, 1992), which mixes a natural exponential family with an unknown distribution whose first two moments are specified. The second one is the double exponential families developed by Efron (1986). The survival functions based on both methods are derived. Score statistics are then obtained to test the null hypothesis of no overdispersion in data with censoring.

Chapter 3 applies the results from Chapter 2 to some specific distributions such as the overdispersed geometric and exponential distributions. The results are then used to explore the properties of the test statistics via simulation. The

distributions of the tests when the dispersion parameter is at the boundary as well as inside the parameter space are examined in Chapter 4. The power of the test statistics for different degree of overdispersion, censoring and various sample sizes is also obtained through simulation.

Having learned the behavior of the test statistics, we apply them to real data. First is on waiting time to first conception, given by Weinberg and Gladen (1986). This data set has been analyzed by various authors testing overdispersion in waiting time. The second set of data is on duration of breast-feeding from the 1988 National Maternal and Infant Health Survey in the U.S. We examine the existence of overdispersion after including covariates. The results are presented in Chapter 5.

Chapter 6 summarizes the results of the research and provides suggestions for further research.

CHAPTER II

PROPOSED MODELS

In this chapter we propose two models to accommodate overdispersion when there is censoring. One is the mixture of exponential families (Dean, 1992), where the first two moments of the mixing distribution are specified. We derive the survival function based on the mixture, and then establish the likelihood of data, taking right censoring into account. We use the score statistics to test the existence of overdispersion, i.e., test the null hypothesis that the overdispersion parameter is equal to zero.

The second model considered is the double exponential family, which introduces an additional parameter into the one-parameter exponential family to control the variance independently of the mean. On the basis of the properties of double exponential families, we derive the survival function. In this way, we can model overdispersion in censored data while carrying out the usual regression analyses for the mean as a function of the predictors. Again, we use the score statistic to test the null hypothesis that the dispersion parameter is equal to zero.

2.1 The Mixture of Exponential Family Distributions

As mentioned in the previous chapter, we assume that each component of $\boldsymbol{Y}$ has a distribution in the exponential family, taking the form

$$f(y;\theta) = \exp\{\ a(\theta)\ y - g(\theta) + c(y)\ \}, \tag{2.1}$$

for some specific functions $a(\ .\ )$, $g(\ .\ )$ and $c(\ .\ )$. Let the density of Y_i given θ_i^* be $f(y_i \mid \theta_i^*)$, as given by the right-hand side of (2.1), where the θ_i^*'s are continuous independent random variates with finite mean and variance, with

$$E(\theta_i^*) = \theta_i(\underline{x}_i;\beta), \text{ and } Var(\theta_i^*) = \tau b_i\ (\theta_i) > 0. \tag{2.2}$$

Dean (1992) assumes that $E\{(\theta_i^* - \theta_i)^r\} = \alpha_i;\ \ \alpha_i = o(\tau), \quad r \geq 3.$
In the limit as $\tau \rightarrow 0$, this model reduces to $f(y_i\ ;\ \theta_i)$, as given by (2.1). The probability model specified by (2.2) is $f_M(y_i) = E_*\{\ f(y_i;\ \theta_i^*)\ \}$, where E_* denotes the expectation over the distribution of θ_i^*.

In the following discussion, the dependence on θ_i of the functions $\mu_i(\theta_i)$, σ_i^2 (θ_i), $a_i(\theta_i)$, $b_i(\theta_i)$, and $g_i(\theta_i)$ will be suppressed for simplicity of notation. The mixture of exponential family distributions can be expressed as

$$f_M(y_i) = f\{1 + \frac{\tau b_i}{2}\ f''f^{-1} + o(\tau)f^r f^{-1}\},$$

$$= f + \frac{\tau b_i}{2}\ f'' + o(\tau)f^r, \tag{2.3}$$

where

$$f' = \frac{\partial f}{\partial \theta_i} = (a_i'y - g_i')\ f, \quad \text{and } f'' = \frac{\partial^2 f}{\partial \theta_i^2} = (a_i''y - g_i'')\ f + (a_i'Y - g_i')^2\ f.$$

2.2.1 The Survival Function

The information contributed by the censored observations is their survival probability up to the censoring time. The survival function can be derived as $1 - F_M$, where F_M is the cumulative probability function of the mixture of exponential family distributions. The expression of the survival function is as

follows:

$$\mathcal{G}_M(T_i) = 1 - [\int f\ dt + \frac{\tau b_i}{2} \int f'' dt + o(\tau) \int f^r dt] = 1 - [F + \frac{\tau b_i}{2} \int f'' dt]$$

$$= \mathcal{G} - \frac{\tau b_i}{2} \frac{\partial^2}{\partial \theta_i^2} \int_0^T f\ dt$$

$$= \mathcal{G} - \frac{\tau b_i}{2} \frac{\partial^2}{\partial \theta_i^2} (1 - \mathcal{G})$$

$$= \mathcal{G} + \frac{\tau b_i}{2} \frac{\partial^2}{\partial \theta_i^2} \mathcal{G},$$

where $F = \int_0^t f(z_i) dz$, the cumulative probability under the null hypothesis, and $\mathcal{G} = 1 - F$, the survival probability under the null hypothesis.

2.1.2 The Score Statistic

Testing the existence of overdispersion is equivalent to testing the null hypothesis for overdispersion parameter, or $\tau = 0$. We choose to use the score statistic for this task, because it has some computational advantage in that only maximization at $\tau = \tau_0$ is required. Therefore the adequacy of the basic model specified by θ can be tested by augmentation in various directions without re-maximization.

To construct the score statistic, the likelihood function is needed. The likelihood function taking into account of censoring is

$$\mathcal{L}_M = \prod_{i=1}^{n} [f + \frac{\tau b_i}{2} f'']^{\delta_i} [\mathcal{G} + \frac{\tau b_i}{2} \frac{\partial^2}{\partial \theta_i^2} \mathcal{G}\ ]^{1-\delta_i},$$

where $\delta_i = 1$ if the event occurs, and $\delta_i = 0$ otherwise.

The score statistic is

$$U_i\,(\widehat{\theta}_i,\tau=0)=\frac{\partial\, l_i}{\partial\tau}\;|_{\,\tau=0}\;=\frac{1}{2}\Big\{\,\delta_i\;b_i\;(a_i'\,)^2\;[(Y_i-\mu_i)^2-(a_i')^{-2}(g_i''-a_i''Y_i)]$$

$$+(1-\delta_i)\;b_i\;[\,\frac{1}{\vartheta}\,\frac{\partial^2}{\partial\theta_i^2}\,\vartheta]\,\Big\}\,,$$

where l_i is the log of likelihood. Notice that for those who failed $U_i(\widehat{\theta}_i)$ compares two estimates of σ_i^2, one based on the sample variance and the other derived from the form of the variance in the base model (2.1), i.e., based on there being no overdispersion.

We now have a situation for the application of asymptotic methods with censored data: the variance of the score statistic, or the Fisher information matrix $\mathfrak{I}(\theta)$, involves the potential censoring times for individuals under study, whereas these would otherwise often not be known for individuals that fail. Further, it is not clear that potential, but unobserved, censoring times should affect the inference about overdispersion even if they are available. For these reasons, the observed information matrix,

$$\mathbf{I}(\theta)=\;-\frac{\partial^2\mathrm{logL}(\theta)}{\partial\theta^2},$$

or $\mathbf{I}(\widehat{\theta})$ is commonly substituted for $\mathfrak{I}(\theta)$ (Kalbfleisch and Prentice, 1980).

The elements of the observed information matrix $\mathbf{I}(\beta,\,\tau)$ are

$$\mathbf{I}\,(\beta,\,\tau)=\begin{bmatrix} I_\beta & I_{\beta\tau} \\ I_{\beta\tau}^T & I_\tau \end{bmatrix},$$

where I_β and I_τ are $p \times p$ and 1×1 matrices, respectively. Let X be an $n \times p$ matrix with ir-element $\frac{\partial \theta_i}{\partial \beta_r}$, $\underline{1}$ be an $n \times 1$ unit vector, and W_1 and W_2 be diagonal matrices with ith diagonal elements,

$$W_{1i} = \left\{ -\frac{\partial^2 l}{\partial \theta_i^2} \right\} |_{\tau = 0} = \delta_i \, (g_i'' - a_i'' \, y_i) \; +$$

$$(1 - \delta_i) \, [-\tfrac{1}{\vartheta} \, \tfrac{\partial^2}{\partial \theta_i^2} \, \vartheta + \tfrac{1}{\vartheta^2} \, (\tfrac{\partial}{\partial \theta_i} \, \vartheta)^2];$$

$$W_{2i} = \left\{ -\frac{\partial^2 l}{\partial \theta_i \, \partial \tau} \right\} |_{\tau = 0} = \tfrac{1}{2} \left\{ \delta_i \, [\, b_i \, (\, (a_i'' y_i - g_i'') + \, (a_i' \, y_i - g_i')^2 \right.$$

$$- (a_i''' \, y_i - g_i''') - 3 \, (a_i'' y_i - g_i'') \, (a_i' \, y_i - g_i') - \, (a_i' \, y_i - g_i')^3 \,)$$

$$- b_i' \, [\, (a_i'' \, y_i - g_i'') + (a_i' \, y_i - g_i')^2 \,]$$

$$\left. + (1 - \delta_i) \, [\, \tfrac{b_i}{\vartheta^2} \, (\, \tfrac{\partial \vartheta}{\partial \theta_i} \,) \, (\, \tfrac{\partial^2}{\partial \theta_i^2} \, \vartheta) - \tfrac{1}{\vartheta} (\, (\tfrac{\partial b_i}{\partial \theta_i}) \, (\tfrac{\partial^2}{\partial \theta_i^2} \, \vartheta) + b_i \, \tfrac{\partial^3}{\partial \theta_i^3} \, \vartheta) \,] \right\};$$

and

$$I_\tau = \left\{ -\sum_{i=1}^{n} \frac{\partial^2 l}{\partial \tau_i^2} \right\} |_{\tau = 0} = \sum_{i=1}^{n} \frac{b_i^2}{4} \left\{ \delta_i \; (a_i' \, y_i - g_i')^2 + (\, 1 - \delta_i \,) \; (\, \tfrac{1}{\vartheta} \; \tfrac{\partial^2}{\partial \theta_i^2} \, \vartheta \,)^2 \right\}.$$

Then $I_\beta = X^T W_1 X$ and $I_{\beta\tau} = X^T W_2 \, \underline{1}$. The variance of the test statistic is $V^2 = I_\tau - \underline{1}^T W_2 X (X^T W_1 X)^{-1} X^T W_2 \; \underline{1}$, and a standardized test statistic for testing $\tau = 0$ is

$$S_{cM} = \frac{\sum_{i=1}^{n} U_i(\widehat{\theta}_i)}{\widehat{V}} ,$$

where I_β and I_τ are $p \times p$ and 1×1 matrices, respectively. Let X be an $n \times p$ matrix with ir-element $\frac{\partial \theta_i}{\partial \beta_r}$, $\underline{1}$ be an $n \times 1$ unit vector, and W_1 and W_2 be diagonal matrices with ith diagonal elements,

$$W_{1i} = \left\{ -\frac{\partial^2 l}{\partial \theta_i^2} \right\} |_{\tau = 0} = \delta_i \, (g_i'' - a_i'' \, y_i) \; +$$

$$(1 - \delta_i) \, [\, -\frac{1}{\vartheta} \frac{\partial^2}{\partial \theta_i^2} \vartheta + \frac{1}{\vartheta^2} (\frac{\partial}{\partial \theta_i} \vartheta)^2];$$

$$W_{2i} = \left\{ -\frac{\partial^2 l}{\partial \theta_i \, \partial \tau} \right\} |_{\tau = 0} = \frac{1}{2} \Big\{ \, \delta_i \, [\, b_i \, (\, (a_i'' y_i - g_i'') + \, (a_i' \, y_i - g_i')^2$$

$$- (a_i''' \, y_i - g_i''') - 3 \, (a_i'' y_i - g_i'') \, (a_i' \, y_i - g_i') - \, (a_i' \, y_i - g_i')^3 \,)$$

$$- b_i' \, [\, (a_i'' \, y_i - g_i'') + (a_i' \, y_i - g_i')^2 \,]$$

$$+ (1 - \delta_i) \, [\, \frac{b_i}{\vartheta^2} \, (\, \frac{\partial \vartheta}{\partial \theta_i} \,) \, (\, \frac{\partial^2}{\partial \theta_i^2} \vartheta) - \frac{1}{\vartheta} (\, (\frac{\partial b_i}{\partial \theta_i}) \, (\frac{\partial^2}{\partial \theta_i^2} \vartheta) + b_i \, \frac{\partial^3}{\partial \theta_i^3} \vartheta) \,] \Big\};$$

and

$$I_\tau = \left\{ -\sum_{i=1}^{n} \frac{\partial^2 l}{\partial \tau_i^2} \right\} |_{\tau = 0} = \sum_{i=1}^{n} \frac{b_i^2}{4} \Big\{ \, \delta_i \, \, (a_i' \, y_i - g_i')^2 + (\, 1 - \delta_i \,) \, \, (\, \frac{1}{\vartheta} \, \frac{\partial^2}{\partial \theta_i^2} \vartheta \,)^2 \, \Big\}.$$

Then $I_\beta = X^T W_1 X$ and $I_{\beta\tau} = X^T W_2 \, \underline{1}$. The variance of the test statistic is $V^2 = I_\tau - \underline{1}^T W_2 X (X^T W_1 X)^{-1} X^T W_2 \, \underline{1}$, and a standardized test statistic for testing $\tau = 0$ is

$$S_{cM} = \frac{\sum_{i=1}^{n} U_i(\widehat{\theta}_i)}{\widehat{V}},$$

with $\widehat{V} = V(\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$. This test statistic is asymptotically standard normal as $n \to \infty$ for complete data. Large values of S_{cM} provide evidence against the null hypothesis (Cox and Hinkley, 1974).

To summarize this section, we began with the general exponential family, then mixed it with an unknown distribution whose first and second moments are specified. The mixture allows us to introduce an additional parameter τ, which varies the dispersion of the family without changing the mean. We derive the survival function based on the extended family. Similar to the mixing density, the survival function involves only the original survival function and the overdispersion parameter. Thus we test the existence of overdispersion in censored data. We use the score statistic for this task.

The discussion in this section suggests that the functions a_i and g_i depend on the index i only through covariates $\underline{x}_i$. Dean (1992) suggested that the results can be applied to more general situations in which some of Y's have a different underlying probability distribution. However, such situations are rare.

2.2 The Double Exponential Families

The second model is based on the double exponential families. The terminology follows Efron (1986), who defines an ordinary one-parameter exponential family as

$$g_\mu(y) = \exp\, n\, [\, \eta\, Y - \psi(\mu)\,]\, [dG(y)], \tag{2.4}$$

where $G(y)$ is the carrier measure for the exponential family, so that

$Pr\{A\} = \int_A g_\mu(y)\ dG(y)$ for measurable sets A; and n is the sample size, in accordance with the familiar situation in which y is actually the average of n independent quantities z_i, each of form (2.5), except with $n = 1$, as follows:

$$y = \sum_{i=1}^{n} \frac{z_i}{n} \qquad \text{where } z_i \overset{ind}{\sim} g_{\mu,1}.$$

Note that the definitions used above, and most of the results that follow can be stated in ways that do not require the exponential family to be expressed in form (2.4). Efron (1986) defined the double exponential family with parameter μ, ϕ and n, given an exponential family (2.4), is

$$\tilde{f}_{\mu,\phi,n}(y) = c(\mu, \phi, n)\ \phi^{1/2}\ \{g_{\mu,n}(y)\}^{\phi}\ \{g_{y,n}(y)\}^{1-\phi}\ [dG_n(y)]. \tag{2.5}$$

The constant $c(\mu, \phi, n)$ is defined to make $\int_{-\infty}^{\infty} \tilde{f}_{\mu,\phi,n}(y)\ dG(y) = 1$. The possible values of μ in (2.4), those for which $g_{\mu,n}(y)$ is a genuine density, lie in an interval of the real line. We assume y's lie in this interval too. Then $\mu = y$ is the maximum likelihood estimate (MLE) of μ. That is, $g_y(y)$ maximizes $g_{\mu,n}(y)$ over the allowable choices of μ (Efron, 1986).

In addition to the definitions following (2.5), we need to introduce notations for the variance function and Kullback-Leibler distance of an ordinary exponential family $g_{\mu,n}(y)$ as follows:

Variance function:

$$V(\mu) = \text{var}_{\mu,1}(z) = \int_{-\infty}^{\infty} (z-\mu)^2\ g_{\mu,1}(z)\ dG(z) \tag{2.6}$$

Kullback-Leibler distance:

$$\mathbb{I}\,(\mu_1,\,\mu_2) = E_{\mu_1,1}\log[\; g_{\mu_1,1}(y) \;/\; g_{\mu_2,1}(y)\;]. \tag{2.7}$$

These definitions apply to the case $n = 1$. For the general case $g_{\mu,n}$ considered in (2.4), the variance function is

$$\mathrm{var}_{\mu,n}\,(y) = \frac{V(\mu)}{n},$$

and the Kullback-Leibler distance is

$$\mathbb{I}_n\,(\mu_1,\,\mu_2) = E_{\mu_1,n}\log[\; g_{\mu_1,n}(y) \;/\; g_{\mu_2,n}(y)\;] = n\;\mathbb{I}\,(\mu_1,\,\mu_2). \tag{2.8}$$

Twice the Kullback-Leibler distance $\mathbb{I}\,(y,\,\mu)$ is the ***Deviance*** $\mathbf{D}\,(y,\,\mu)$.

There are several useful facts about the double exponential family which make the inference about the dispersion parameter ϕ straight forward. They are:

Fact 1. The constant $c(\mu,\,\phi,n)$ nearly equals 1;

Fact 2. The density $\widetilde{f}_{\mu,\phi,n}(y)$ has mean value approximately equal to μ and variance approximately equal to $V(\mu)/(n\phi)$;

Fact 3. With ϕ and n fixed, (2.5) is an exponential family indexed by μ:

$$\widetilde{f}_{\mu,\phi,n}(y) \;= a_{\phi,n}\,(\mu)\; b_{\phi,n}(y)\,\exp\{n\phi\;[\eta y - \psi(\mu)]\;\}\;dG_n(y),$$

with natural statistic y, natural parameter η, and expectation parameter approximately equal to μ.

Fact 4. With μ and n fixed, (2.5) is an exponential family with density indexed by ϕ:

$$\tilde{f}_{\mu,\phi,n}(y) = c(\mu,\, \phi,\, n)\; \phi^{1/2}\; e^{-n\phi I(y,\mu)}\; g_{y,n}\,(y)\; [dG_n(y)],$$

with natural statistic $-n\; I(y,\mu)$ and natural parameter ϕ, where $I(y,\mu)$ is the Kullback-Leibler distance of an ordinary exponential family.

Fact 5. The density $\tilde{f}_{\mu,\phi,n}(y)$ represents approximately the same probability distributionas the ordinary one-parameter exponential family with n changed to $n\phi$:

$$\int_A \tilde{f}_{\mu,\phi,n}(y)\; dG_n(y) \doteq \int_A g_{\mu,\, n\phi}(y)\; dG_{n\phi}(y), \text{ for any interval } A.$$

Fact 4 is an obvious result from the definition of the double exponential family. Fact 5 holds exactly, if fact 1 is true. In some cases, the constant c is not close to 1 and involves μ and ϕ, in which case we have to re-evaluate fact 5. We discuss this in the sections on applications below.

2.2.1 The Survival Function

We derive the survival function utilizing the first and fifth facts listed above for the double exponential family. The cumulative probability from time 0 to t is

$$F = \int_0^t \tilde{f}_{\mu,\phi,n}(y)\; dG(y) \doteq \int_0^t \exp[n\phi(\eta\; y - \psi(\mu)]\; dG_\phi(y).$$

By definition, we know the survival probability $S_{n\phi} = 1 - F$, where ϕ is the

dispersion parameter. To guarantee that ϕ is positive, we re-define it as e^{τ}. When $\tau \to 0$, the distribution reduces to the ordinary exponential family.

2.2.2 The Score Statistic

Similar to the procedure for the mixture of the exponential family, we construct the likelihood function taking censoring into account before deriving the score statistic. The likelihood function is

$$\mathcal{L}_{\phi} = \prod_{i=1}^{n} [\ e^{\frac{1}{2}\{\tau - e^{\tau} D_i(y_i\ ,\ \mu_i)\}}\ f_y\ (y)\ ]^{\delta_i}\ [\ S_{\phi}]^{1-\delta_i}\ ;$$

and the log likelihood is,

$$l_{\phi} = \sum_{i=1}^{n} \left\{ \frac{1}{2}\ \delta_i\ \{\ [\tau - e^{\tau} D_i(y_i\ ,\ \mu_i)\ ] + \log\ (f_y(y)\ )\ \} + (\ 1-\delta_i)\ \log\ (S_{\phi})\ \right\},$$

where $\delta_i = 1$ if the event is observed, and $\delta_i = 0$ otherwise. Notice that the term f_y does not involve parameters, so it will be dropped in deriving the test statistic. Then the inference about τ for those who failed involves only the deviance $D_i(y_i\ ,\ \mu_i)$.

The score statistic for testing the null hypothesis that $\tau = 0$ is

$$\sum_{i=1}^{n} U_i\ (\widehat{\mu}_i\ ),$$

where

$$U_i\ (\widehat{\mu}_i\ ,\ \tau = 0) = \frac{\partial\ l_i}{\partial\ \tau}\ \Big|_{\tau = 0}\ = \frac{1}{2}\ \delta_i\ \ [1 - D_i(y_i\ ,\ \mu_i)\ ] + (1-\delta_i)\ \frac{1}{S_{\phi}}\ \frac{\partial}{\partial\ \tau}\ S_{\phi}.$$

As we can see from this, the part involving the censored individual is somewhat complicated.

Similar to the derivation of the elements of the information matrix for the mixture of the exponential family, regression parameters are considered. Let μ_i be a function of a $p \times 1$ vector of covariates $\underline{x_i}$ and regression parameters $\underline{\beta}$; that is, $\mu_i = \mu_i\,(\underline{x_i};\,\beta)$, $i = 1, \ldots, n$. Let X be an $n \times p$ matrix with ir-element $\frac{\partial \mu_i}{\partial \beta_r}$, $\underline{1}$ be an $n \times 1$ unit vector, and W_1 and W_2 be diagonal matrices with ith diagonal elements,

$$W_{1i} = \Big\{ -\frac{\partial^2 l}{\partial \mu_i^2} \Big\} \Big|_{\tau = 0} = \delta_i \frac{1}{2\,V(\mu_i)} + (\,1 - \delta_i\,) \frac{\partial^2 \log\,(S_\phi)}{\partial\,\mu_i^2},$$

$$W_{2i} = \Big\{ -\frac{\partial^2 l}{\partial\,\mu_i\,\partial\,\tau} \Big\} \Big|_{\tau = 0} = \delta_i \frac{(y_i - \mu_i)}{2\,V(\mu_i)} + (\,1 - \delta_i\,) \frac{\partial^2 \log\,(S_\phi)}{\partial\,\mu_i\,\partial\,\tau}, \text{ and}$$

$$I_\tau = \Big\{ -\frac{\partial^2 l}{\partial\,\tau_i^2} \Big\} \Big|_{\tau = 0} = \frac{\delta_i}{2}\,D(y_i\,,\,\mu_i) + (\,1 - \delta_i\,) \frac{\partial^2 \log\,(S_\phi)}{\partial\,\tau^2}.$$

We know that $\frac{1}{2} \frac{\partial}{\partial\,\mu} D(y,\,\mu) = \frac{-(y - \mu)}{V(\mu)}$, the term for events in each element, are easy to derive. Again, we have $I_\beta = X^T W_1 X$, and $I_{\beta\tau} = X^T W_2\,\underline{1}$. The variance of the test statistic is $V^2 = I_\tau - \underline{1}^T W_2 X (X^T W_1 X)^{-1} X^T W_2\,\underline{1}$, and a standardized test statistic for testing $\tau = 0$ is

$$S_{cD} = \frac{\sum_{i=1}^{n} U_i(\widehat{\mu}_i)}{\widehat{V}},$$

with $\widehat{V} = V(\widehat{\mu}_1,\ \ldots, \widehat{\mu}_n)$. Large values of S_{cD} provide evidence against the hypothesis. The parameters in score statistics based on both models can be obtained through iteration based on the models under the null hypothesis.

In this section, we define the double exponential family based on the one-parameter exponential family. For most of the situations, the double exponential family behaves like $g_{\mu,n\phi}(y)$, that is, the ordinary exponential family with sample size changed from n to $n\phi$. The extended family is an exponential family in μ when ϕ and n are fixed and an exponential family in ϕ when μ and n are fixed. We derived the survival function utilizing a number of facts about the extended family. Those facts now allow us to derive the test statistic to detect the existence of overdispersion in censored data.

2.3 Discussion

We have derived the tests for testing the existence of overdispersion with censoring based on the mixture of the exponential families (Dean, 1992) and the double exponential families (Efron, 1986). The mixture of the exponential family mixes a distribution from the exponential family with an unknown distribution whose first two moments are specified, whereas the double exponential family introduces an additional parameter to accommodate overdispersion.

Both of the methods require the complete representation of the survival function. Fortunately, only the survival function under the null hypothesis is needed for the mixture of the exponential family. The survival function of the double exponential function can be derived directly if the constant is approximately 1. However, the facts that the constant may depend on the distribution of the random variable y, and when the constant is different from one, it complicates the derivation of the survival function.

In this chapter, we have developed test statistics to detect overdispersion in data with censoring in general forms. Since survival functions are required in the statistics derived from either the mixture of exponential families or the double exponential family, detailed derivations for some specific distributions are presented in the next chapter.

CHAPTER III

APPLICATIONS

Models with explanatory variables based on exponential family distributions are commonly used in modeling accelerated lifetime. If the model is to be used with censored data, it is helpful if both the density and the survival functions can be expressed in reasonably explicit form. For this reason, we apply the methods described in the previous chapter to some special distributions, such as the overdispersed geometric distribution and the exponential distribution.

In each case, we derive three score statistics for testing the hypothesis of no overdispersion. These tests are based on the models specified in Chapter 2 and the test based on the full representation of the overdispersed model. Test statistics based on the overdispersed geometric distribution are presented first, followed by test statistics based on the overdispersed exponential distribution.

3.1 Test For An Overdispersed Geometric Distribution

The geometric probability model is often used for the discrete waiting time to an event. If the probability of an event varies among individuals, the observed variance will be larger than the variance calculated under the simple geometric model. One common practice is to let the event probability follow a beta distribution (Sheps and Menken, 1973; Suchindran, 1972; Weinberg and Gladen, 1986), and then do the statistical analyses using the beta-geometric model.

3.1.1 Mixture of Exponential Family Distributions

The first model we apply is the mixture of exponential family distributions. Suppose the time to the first event given probability $\theta^* = \theta$ follows a geometric distribution. Then we have

$$f(t_i \mid \theta_i^*) = (1-\theta_i)^{t_i - 1}\theta_i \ ,$$

which can be expressed in the form of a natural exponential family:

$$\exp\{\ t_i \log\ (1-\theta_i) - \log\ (1-\theta_i) + \log\ \theta_i\}.$$

Under Dean's formulation in Chapter 2, the conditional expected waiting time is $E(T_i \mid \theta_i^*) = \frac{1}{\theta_i^*}$ and the variance is $Var\ (T_i \mid \theta_i^*) = \frac{1-\theta_i^*}{(\theta_i^*)^2}$. Let the parameter θ_i^* follow a beta distribution, $\sim$ beta[$\frac{\theta}{\gamma}$, $\frac{1-\theta}{\gamma}$], then the expected value is $E(\ \theta_i^*) = \theta_i\ (\ \underline{x}\ ,\ \underline{\beta}\ )$, and the variance $E(\theta_i^* - \theta_i)^2 = \tau\ \theta_i\ (1-\theta_i)$. The mixed distribution is $f_M\ (\ t_i\ ) = f\ (\ t_i\ ) + \frac{\tau\ b_i}{2}\ f'' + o\ (\ r\ )\ f^r$. Table 3.1 lists the a_i, g_i, b_i, τ and the survival function $\mathcal{S}$ under the null hypothesis, following Dean's notation.

We, then, substitute a_i, g_i, b_i, τ to the formulation of the score statistics based on the mixture of the exponential family in Section 2.1 to get the test statistics. The score statistic is

$$U(\widehat{\theta}_i) = \frac{\partial l}{\partial \tau}\Big|_{\tau=0} = \frac{1}{2\ (1-\widehat{\theta}_i)}\ \ \delta_i\ \ (\widehat{\theta}_i\ t_i^2 - \widehat{\theta}_i\ t_i - 2\ t_i + 2)$$

$$+\ \frac{\widehat{\theta}_i}{2(1-\widehat{\theta}_i)}\ (1-\delta_i)\ (t_i^2 - t_i)\ .$$

Elements of the observed information matrix under the null hypothesis are then derived. They are:

$$W_{1i} = \delta_i\,(g_i'' - a_i''\,t_i)\; + (1-\delta_i)\,\frac{t_i}{(1-\theta_i)^2}$$

$$= \delta_i\,\frac{1}{(1-\theta_i)^2}\,(\frac{1-2\theta_i}{\theta_i^2} + t_i) + (1-\delta_i)\,\frac{t_i}{(1-\theta_i)^2}\,,$$

$$W_{2i} = \frac{\delta_i}{2}\,\{b_i\,[(g_i''' - a_i'''\,t_i) - 2\,(a_i''\,t_i - g_i'')(a_i'\,t_i - g_i')]$$

$$-\,b_i'\,[(a_i''\,t_i - g_i'') + (a_i'\,t_i - g_i')^2]\} +$$

$$\frac{(1-\delta_i)}{2}\,[\,\frac{-\,b_i\,t_i^2\,(t_i-1)}{(1-\theta_i)^3} - \frac{b_i'\,t_i\,(t_i-1)(1-\theta_i) - b_i t_i(t_i-1)(t_i-2)}{(1-\theta_i)^3}]$$

$$= -\,\delta_i\,\frac{2-3t_i+t^2}{(1-\theta_i)^2} - (1-\delta_i)\,\frac{t_i\,(t_i-1+\theta_i-t_i^2\,\theta_i+t_i\,\theta_i^2-\theta_i^2)}{(1-\theta_i)^2},\ \text{and}$$

$$I_\tau = \frac{b_i^2}{4}\,\{\delta_i\,[(a_i''\,t_i - g_i'') + (a_i'\,t_i - g_i')^2]^2\; + (1-\delta_i)\,\frac{t_i^2\,(t_i-1)^2}{(1-\theta_i)^4}\}$$

$$= \frac{b_i^2}{4}\,\{\delta_i\frac{(t_i^2\,\theta_i - 2\,\theta_i - t_i\theta_i + 2)^2}{\theta^2(1-\theta_i)^4} + (1-\delta_i)\,\frac{t_i^2\,(t_i-1)^2}{(1-\theta_i)^4}\}.$$

We substitute the elements above into $I_\beta = X^T W_1 X$, and $I_{\beta\tau} = X^T W_2\,\underline{1}$ to get the variance of the test statistic $V^2 = I_\tau - \underline{1}^T W_2 X (X^T W_1 X)^{-1} X^T W_2\,\underline{1}$ and the standardized score statistic

$$S_c = \frac{\sum_{i=1}^{n} U(\theta_i)}{V}\,.$$

Table 3.1 Elements of the Mixture of the Geometric Distribution.

	Overdispersed Geometric Model
θ	$\frac{1}{\mu}$
$a(\theta)$	$\log(1-\theta)$
$g(\theta)$	$\log(\frac{1-\theta}{\theta})$
$b(\theta)$	$\theta\,(1-\theta)$
τ	$\frac{\gamma}{1+\gamma}$
$\mathcal{G}$	$(1-\theta)^t$
$E(T_i)$	$\frac{1}{\theta}$

3.1.2 The Double Exponential Formulation for a Geometric Distribution

Double exponential families are defined in the previous chapter,

$$\widetilde{f}_{\mu,\phi,n}(y) = c\,(\mu,\phi,n)\,\phi^{1/2}\,\{g_{\mu,n}(y)\}^{\phi}\,\{g_{y,n}(y)\}^{1-\phi}[dG_n(y)]$$

Before we derive its score statistic, we need to evaluate if the constant c is approximately 1. If c is a function of μ and ϕ and is not close to 1, then it cannot be ignored. Our derivations depend on Efron's (1986) expression for an exponential family density,

$$g_{\mu,n}(y) = g_{y,n}(y)\,\exp[-n\mathbb{I}(y,\,\mu)], \tag{3.1}$$

where n is the sample size and y is the average of n independent quantities of z_i each of form (2.6). In order to evaluate c, we need to evaluate

$$\frac{1}{c(\mu,\ \phi,n)} = \int_{-\infty}^{\infty} f_{\mu,\phi,n}\ (y)\ d\ G_n(y).$$

We use a summation instead of an integral since we are dealing with discrete time:

$$\frac{1}{c(\mu,\ \phi,n)} = \phi^{1/2} \sum_{y=1}^{\infty} \exp\ [-n\ \phi\ \mathbb{I}(y,\ \mu)]\ g_{y,n\phi} R_{n,\phi} dG_n(y), \tag{3.2}$$

where $R_{n,\phi} = [g_{y,n}(y)/g_{y,n\phi}(y)](dG_n/dG_{n\phi})(y).$ (3.3)

Using expression (3.1), (3.2) can be expressed as

$$\frac{1}{c(\mu,\ \phi,n)} = \phi^{1/2} \sum_{y=1}^{\infty} g_{\mu,n\phi} R_{n,\phi} dG_{n\phi}(y). \tag{3.4}$$

We then apply expressions (3.3) and (3.4) to the geometric distribution, which can be expressed as

$$g_{\mu,1}(t) = (\frac{1}{\mu})\ (1-\frac{1}{\mu})^{y-1},$$

where $g_{\mu,n}(y)$ has a negative binomial distribution with parameter $n\mu$ and n. Here we have not written the density form in canonical form (2.5), since it is not necessary for evaluating the constant c. Definition (3.3) gives

$$R_{n,\phi} = \frac{g_{y,n}}{g_{y,n\phi}} = \frac{C_{n-1}^{y-1}\ (\frac{1}{ny})^n\ (1-\frac{1}{ny})^{y-n}}{C_{n\phi-1}^{y-1}\ (\frac{1}{n\phi y})^{n\phi}(1-\frac{1}{n\phi y})^{y-n\phi}},$$

which depends on y. To evaluate c, we substitute $R_{n,\phi}$ into expression (3.4) and obtain

$$\frac{1}{c(\mu,\ \phi,n)} = \phi^{1/2} \sum_{y=1}^{\infty} \frac{C_{n-1}^{y-1}\ (\frac{1}{ny})^n\ (1-\frac{1}{ny})^{y-n}}{C_{n\phi-1}^{y-1}\ (\frac{1}{n\phi y})^{n\phi}(1-\frac{1}{n\phi y})^{y-n\phi}}\ C_{n\phi-1}^{y-1}\ (\frac{1}{n\phi\mu})^{n\phi}(1-\frac{1}{n\phi\mu})^{y-n\phi}.$$

When $n = 1$, the expression on the right-hand side becomes

$$\frac{1}{c(\mu,\ \phi,n)} = \phi^{1/2} \sum_{y=1}^{\infty} \frac{(\frac{1}{y})\ (1-\frac{1}{y})^{y-1}}{(\frac{1}{y})^{\phi}(1-\frac{1}{\phi y})^{y-\phi}}\ (\frac{1}{\mu})^{\phi}(1-\frac{1}{\phi\mu})^{y-\phi}. \tag{3.5}$$

Equation 3.5 can be arranged as

$$\frac{1}{c(\mu,\ \phi,n)} = \phi^{1/2+\phi} \sum_{y=1}^{\infty} \frac{(\frac{1}{y})\ (1-\frac{1}{y})^{y-1}}{(\frac{1}{y})^{\phi}(1-\frac{1}{\phi y})^{y-\phi}}\ \frac{C_{\phi-1}^{y-1}}{C_{\phi-1}^{y-1}}\ (\frac{1}{\phi\mu})^{\phi}(1-\frac{1}{\phi\mu})^{y-\phi}.$$

$$= \phi^{1/2+\phi} \sum_{y=1}^{\infty} \gamma(y)\ g_{\mu,\phi},$$

where $\gamma(y) = \dfrac{(\frac{1}{y})\ (1-\frac{1}{y})^{y-1}}{(\frac{1}{y})^{\phi}(1-\frac{1}{\phi y})^{y-\phi}}\ \dfrac{1}{C_{\phi-1}^{y-1}}$, a function of y; and

$g_{\mu,\phi}$ is a negative binomial distribution with $p = \dfrac{1}{\phi\mu}$. Then

$\sum_{y=1}^{\infty} \gamma(y)\ g_{\mu,\phi}$ is the expectation of $\gamma(y)$ over the negative binomial

distribution $g_{\mu,\phi}$. A Taylor expansion is used to evaluate the expectation of $\gamma(y)$:

$$E[\gamma(y)] = E[\gamma(y_0)] + \sum_{j=1}^{k} E[\frac{(y-y_0)^j}{j!}\ \gamma^{(j)}(y_0) + \Delta_k(y,\ y_0)\ ],$$

where

$$\Delta_k(y, y_0) = \frac{(y-y_0)^{k+1}}{(k+1)!}\ \gamma^{(k+1)}(h\ y_0 + (1-h)\ y) \text{ for some } 0 < h < 1;$$

and $\gamma^{(j)}$ is the jth derivative of γ. Due to the complex expression for the constant $c(\mu, \phi,1)$, we take only the first two terms in the expansion

$$E[\gamma(y)] = \gamma(\mu) + \frac{1}{2!} E[\ (y-\mu)^2]\ \gamma''(\mu), \tag{3.7}$$

and substitute them to evaluate the constant $c(\mu,\phi,1)$ (see Appendix A). We have defined $\phi = e^{\tau}$ in Chapter 2. We evaluate the fact that the constant $c \to 1$ as $\tau \to 0$. The constant c then is plugged back to the double exponential family formulation for the geometric distribution.

The double exponential family based on a geometric distribution is

$$\widetilde{f}_{\mu,\phi,n}(y) = c(\mu,\phi,1)\ \ \phi^{1/2} C_{\phi-1}^{y-1}\ (1-\frac{1}{\phi\ \mu})^{y-\phi}(\frac{1}{\phi\ \mu})^{\phi};$$

and the survival function is

$$\widetilde{S}_{\mu,\phi,n}(y) = 1 - c(\mu,\phi,1)\ \ \phi^{1/2} \sum_{j=1}^{y} C_{\phi-1}^{j-1}\ (1-\frac{1}{\phi\ \mu})^{j-\phi}(\frac{1}{\phi\ \mu})^{\phi}.$$

The score statistic based on the double exponential family taking censoring into account, is given by:

$$U_i(\widehat{\mu_i}) = \frac{\partial\ l_{\phi}}{\partial \tau}\Big|_{\tau=0}$$

$$= \delta_i\ [\frac{1}{2} + \frac{\partial\ c(\mu,\phi,1)}{\partial \tau}\ \Big|_{\tau=0}\ (1-\frac{1}{\mu})^{y-1}\ \frac{1}{\mu}$$

$$+ \frac{\partial}{\partial \tau} C_{\phi-1}^{j-1}\ (1-\frac{1}{\phi\ \mu})^{j-\phi}(\frac{1}{\phi\ \mu})^{\phi}\ \Big|_{\tau=0}]$$

$$-(1-\delta_i)\frac{1}{S_0}\left[\frac{\partial\, c(\mu,\phi,1)}{\partial\tau}\mid_{\tau=0}\sum_{j=1}^{y}(1-\tfrac{1}{\mu})^{j-1}\frac{1}{\mu}\right.$$

$$\left.+\frac{\partial}{\partial\tau}\sum_{j=1}^{y}C_{\phi-1}^{j-1}(1-\frac{1}{\phi\,\mu})^{j-\phi}(\frac{1}{\phi\,\mu})^{\phi}\mid_{\tau=0}\right],$$

where

$$\frac{\partial c(\mu,\phi,1)}{\partial\tau}\mid_{\tau=0}=\frac{1}{2\,(\mu-1)}[1-2\,\mu^2\,\Psi(2,\mu)-2\,G-2\,\mu\,\Psi(\mu)+2\log(\mu-1)$$

$$+\mu\,\Psi(2,\mu)+\mu^3\,\Psi(2,\mu)+2\,\mu\,G-2\log(\mu-1)];$$

$$\frac{\partial}{\partial\tau}C_{\phi-1}^{j-1}(1-\frac{1}{\phi\,\mu})^{j-\phi}(\frac{1}{\phi\,\mu})^{\phi}\mid_{\tau=0}=\log(y-1)-\log(\mu-1)-\frac{\mu}{\mu-1};$$

$$\frac{\partial}{\partial\tau}\sum_{j=1}^{y}C_{\phi-1}^{j-1}(1-\frac{1}{\phi\,\mu})^{j-\phi}(\frac{1}{\phi\,\mu})^{\phi}\mid_{\tau=0}$$

$$=\frac{1}{\mu\,(1-\frac{1}{\mu})}\sum_{j=1}^{y}(1-\tfrac{1}{\mu})^{j}\,[\,\log(\mathrm{j})+\log(1-\tfrac{1}{\mathrm{j}})\,]+[\,S_0-(1-\tfrac{1}{\mu})\,]\log(\mu-1)$$

$$+\frac{S_0}{(\mu-1)}(y+\mu+1)+1;$$

$S_0=(1-\frac{1}{\mu})^{y}$, the survival function under the null hypothesis;

$\Psi(\mu)=\frac{\partial}{\partial\,\mu}[\,(\mu-1)!\,]$; $\Psi(n,\mu)=\frac{\partial}{\partial\,\mu^{(n)}}[\,(\mu-1)!]^{(n)}$, the nth derivative of the

gamma function, and G is the Euler constant,

$\underset{n\to\infty}{\text{limit}}\,[\sum_{i=1}^{n}\frac{1}{i}-\log(n)]$, with the numerical value 0.5772156649....

Since the derivation of the information matrix is complicated, we use the MAPLE program to do the derivation, especially the part for censoring. The MAPLE program is attached in Appendix B. When the expression is too messy, we substitute the censored time and the average waiting time under the null hypothesis into the calculation to get the numerical value directly.

3.1.3 The Beta-Geometric Distribution

For comparison, we also use the exact distribution to derive the test statistic. Weinberg and Gladen (1986) have worked out the detailed expression for the beta-geometric distribution and applied it to fecundability studies.

Let T denote the number of cycles required for conception. For a given couple with fecundability p, T follows a geometric distribution:

$$\Pr\,(T = t \mid p) = (1-p)^{t-1}\, p.$$

Now we assume that fecundability follows a beta distribution with mean parameter θ and shape parameter γ. The variance of the fecundability is $\frac{\theta\,(1-\theta)\,\gamma}{1+\gamma}$. Removing the conditioning on p by integrating over the beta yields the probability that conception occurs at t for a randomly selected couple:

$$\Pr\,(T = t) = \frac{\theta \prod_{j=1}^{t-1} [1-\theta+(j-1)\gamma]}{\prod_{j=1}^{t} [1+(j-1)\gamma]},$$

and the probability that the couple has experienced $t-1$ unsuccessful cycles is:

$$\Pr\,(T > t-1) = \prod_{j=1}^{t} \frac{[(1-\theta)+(j-1)\gamma]}{[1+(j-1)\gamma]}, \text{ when } t \geq 1.$$

This distribution has mean $\frac{(1-\gamma)}{(\theta-\gamma)}$ and variance $\frac{\theta\,(1-\theta)}{(\theta-\gamma)^2\,(\theta-2\gamma)}$.

The contribution of n couples to the likelihood of conception is then

$$\prod_{i=1}^{n} [\ \Pr\ (T=t)\ ]^{\delta_i} [\ \Pr\ (T>t)\ ]^{1-\delta_i},\ \delta_i = 1 \text{ if conception occurs, 0 otherwise.}$$

The log of the likelihood is

$$l = \sum_{i=1}^{n} \delta_i \{ \log\theta + \sum_{j=1}^{t_i-1} \log[(1-\theta)+(j-1)\gamma] - \sum_{j=1}^{t_i} \log[1+(j-1)\gamma] \}$$

$$+ \sum_{i=1}^{n} (1-\delta_i) \sum_{j=1}^{t_j} \{ \log[(1-\theta)+(j-1)\gamma] - \log[1+(j-1)\gamma] \}.$$

Testing $\tau = 0$ is equivalent to testing $\gamma = 0$. Again, we use the score statistic for this task. The score statistic is

$$U_e(\hat{\theta}) = \frac{\partial l}{\partial \gamma}\Big|_{\gamma=0} = \frac{1}{2} \left\{ \sum_{i=1}^{n} \delta_i \left[\frac{(t_i-2)\,(t_i-1)}{1-\hat{\theta}} - (t_i-1)\,t_i \right] \right.$$

$$\left. + \sum_{i=1}^{n} (1-\delta_i) \left[\frac{(t_i-1)\,t_i}{1-\hat{\theta}} - (t_i-1)\,t_i \right] \right\}.$$

The elements of the observed information matrix, which are required for evaluating the variance of the score statistic, are

$$W_{1i} = \delta_i \left(\frac{1}{\theta_i^2} + \frac{t-1}{(1-\theta_i)^2} \right) + (1-\delta_i) \frac{t}{(1-\theta_i)^2},$$

$$W_{2i} = \delta_i \frac{(t_i-1)(t_i-2)}{2\,(1-\theta_i)^2} - (1-\delta_i) \frac{t_i\,(t_i-1)}{2(1-\theta_i)^2}, \text{ and}$$

$$I_\gamma = \frac{\delta_i}{6}\,[\,\frac{(t_i-1)(t_i-2)(2t_i-3)}{(1-\theta_i)^2} - t_i\,(t_i-1)(2t_i-1)] +$$

$$\frac{1-\delta_i}{6}\,[\,\frac{t_i\,(t_i-1)(2t_i-1)}{(1-\theta_i)^2} - t_i\,(t_i-1)(2t_i-1)].$$

3.2 The Test for Overdispersion In the Exponential Distribution

We apply methods based on the mixture of the exponential family and the double exponential family to model the overdispersed exponential distribution and derive the test statistic for testing the existence of overdispersion. For the full representation of the overdispersed model, we use the Pareto distribution, which is a compound distribution of the exponential and the gamma distributions.

3.2.1 The Mixture of the Exponential Family Distributions

When the survival time T follows the exponential distribution with a parameter θ, the probability density function is

$$f(t) = \theta\, e^{-\theta\, t} \text{ for } t \geq 0,\ \theta > 0;$$
$$= 0 \text{ for } t < 0.$$

Let θ^* be a random variable with density $r_{\theta^*}(\,.\,)$ and suppose that the conditional density of T given $\theta^* = \theta$ is

$$f_{T\,|\,\theta^*}(\,t\,|\,\theta^*\,) = \theta^*\, e^{-\theta^*\, t} = \exp\,(\,\log\,(\theta^*) - \theta^*\, t).$$

Let γ (θ^*) follow a gamma distribution with parameters [$\frac{1}{\gamma}$, $\gamma\theta$], so E (θ_i^*) $= \theta_i$ ($\underline{x}$, $\underline{\beta}$), and E ($\theta_i^* - \theta_i$) $= \tau\ \theta_i^2$, where $\tau = \gamma$. The mixture is

$$f_M\ (\ t_i\ ) \simeq f\ (\ t_i\ ) + \frac{\tau\ b_i}{2}\ f'' = \ \theta_i\ e^{-\theta_i\ t_i}\ + \frac{\tau\ \theta_i^2}{2}\ (-2\ t_i\ + \theta_i\ t_i^2\ )\ e^{-\theta_i\ t_i}\ ,$$

and the survival function based on the mixture is

$$\mathcal{S}_M\ (\ t_i\ ) = e^{-\theta_i\ t_i}\ + \frac{\tau\ \theta_i^2}{2}\ \ t_i^2\ e^{-\theta_i\ t_i}\ .$$

The a_i, g_i, b_i, τ and the survival function $\mathcal{S}$ under the null hypothesis are listed in Table 3.2.

The score statistic under the the mixture of exponential distributions, taking censoring into account, is

$$U(\ \widehat{\theta}_i\ ) = \ \frac{\partial\ l_i}{\partial\ \tau}\ |_{\tau\ =\ 0} = \frac{\delta_i\ \theta_i}{2}\ (\ \theta_i\ t_i^2\ -2\ t_i\ ) + (\ 1 - \delta_i\ )\ \frac{\theta_i^2\ t_i^2}{2}\ ,$$

and the elements of the information matrix are

$$W_{1i}\ = \ -\left\{\frac{\partial^2\ l_i}{\partial\ \theta_i^2}\right\}|\ _{\tau\ =\ 0} = \delta_i\ (\ \frac{1}{\theta_i^2}\ ),$$

$$W_{2i}\ = \ -\left\{\frac{\partial^2\ l_i}{\partial\ \theta_i\ \partial\tau}\right\}|\ _{\tau\ =\ 0} = \ -\ \delta_i\ t_i\ ,\ \text{and}$$

$$I_\tau\ = \ -\left\{\frac{\partial^2\ l_i}{\partial\ \tau_i^2}\right\}|\ _{\tau\ =\ 0} = \delta_i\ \frac{\theta_i^2}{4}\ (\ \theta_i\ t_i^2\ -2\ t_i)^2\ + (\ 1 - \delta_i)\ \frac{\theta_i^4\ t_i^4}{4}\ .$$

Table 3.2 Elements of the Mixture of Exponential Distributions.

	Overdispersed Exponential Model
θ	$\frac{1}{\mu}$
$a(\theta)$	θ
$g(\theta)$	$\log(\theta)$
$b(\theta)$	θ^2
τ	γ
$\mathcal{S}$	$e^{-\theta t}$
$E(T_i)$	$\frac{1}{\theta}$

3.2.2. The Double Exponential Formulation for an Exponential Distribution

We start with the exponential family with $n = 1$, or

$$g_{\mu,1}(t) = \frac{1}{\mu} e^{-\frac{1}{\mu} t} = \exp\left[-\log \mu - \frac{1}{\mu} t \right].$$

According to the terminology of Efron (1986), $\eta_i = -\frac{1}{\mu_i}$ and $\psi(\mu_i) = \log \mu_i$, where $\mu_i = \mu_i (\underline{x}, \underline{\beta})$; and n samples of $g_{\mu,1}$ become a gamma family $g_{\mu,n}$ with parameters μ and n such that

$$g_{\mu,n}(t) = \frac{1}{\Gamma(n)(\frac{\mu}{n})^n} t^{n-1} e^{-\frac{n t}{\mu}}.$$

To evaluate the constant c, we define $R_{n,\phi}$:

$$R_{n,\phi} = \frac{g_{y,n}(t)}{g_{y,n\phi}(t)} = \frac{\Gamma(n\phi)(\frac{n}{e})^n}{\Gamma(n)(\frac{n\phi}{e})^{n\phi}},$$

which does not depend on the random variable t. Efron (1986) has verified that the constant c is close to one for the gamma family using Stirling's formula,

$$c = \frac{1}{\phi^{1/2} R_{n,\phi}} \doteq 1 - \frac{1}{12\, n} \frac{1-\phi}{\phi};$$

and that the double exponential family starting from the gamma family $\int_A \widetilde{f}_{\mu,\phi,n}$ exactly equals $\int_A g_{\mu,n\phi}$, where $g_{\mu,n\phi}$ is the gamma family with n changed to $n\phi$. Therefore, we can use the facts stated in Chapter 2 to derive the survival function and the test statistic for testing the existence of overdispersion in censored data directly.

The double exponential formulation for the exponential distribution is

$$f_D\,(\,t_i\,) = \exp\,[\,\frac{1}{2}\,(\,\tau - e^{\tau}\,D(t_i,\,\mu_i))\,]\,g_{y,n},$$

where $D(t_i,\,\mu_i) = (\,\log\,y_i - \log\,\mu_i) - (\frac{1}{y_i} - \frac{1}{\mu_i})\,y_i$; and the overdispersed survival function is

$$S_D = 1 - \int_0^t f_{\mu,n,\phi}(y)d(y_\phi)$$

$$= 1 - \int_0^t \phi^{1/2}\, g_{\mu,n,\phi}\, R_{n,\phi}\, d(y_\phi)$$

$$= 1 - \phi^{1/2}\, e^{\phi-1} \int_0^t y^{\phi-1} e^{-y/\mu}\, dy\ ,$$

where the cumulative probability for the events up to time t is the integral of the gamma distribution (ϕ, μ) from time 0 to time t, and $n = 1$.

Finally, the score statistic is given by

$$U_i(\widehat{\mu_i}) = \frac{\partial l_\phi}{\partial \tau} \mid_{\tau = 0}$$

$$= \delta_i \left(- D(t_i, \mu_i) \right) + (1 - \delta_i)\, e^{t_i/\mu_i} \left\{ (1 - e^{-t_i/\mu_i}) (\log \mu_i - 1) \right.$$

$$\left. - e^{-t_i/\mu_i} [1 + \frac{t_i}{\mu_i} - \log (t_i)] + 1 - \log (t_i) \right\},$$

with the elements of the information matrix in Appendix C.

3.2.3. The Exponential-Gamma Distribution

Mixing an exponential distribution with a gamma distribution leads to the famous Pareto distribution. Let the parameter $\theta = \frac{1}{\mu}$ in the exponential distribution follow a gamma distribution with parameters $[\ \frac{1}{\gamma}, \gamma\theta]$. Then the density function of the mixture is

$$f_T\ (\ t\ ) = \frac{\frac{1}{\gamma}\ (\frac{1}{\gamma\ \theta})^{\frac{1}{\gamma}}}{(t + \frac{1}{\gamma\ \theta}\ )^{\frac{1}{\gamma}+1}}\ ,$$ and the survival and hazard functions are

$$S_T\ (\ t\ ) = \Big(\frac{1}{\gamma\ \theta\ t\ + 1} \Big)^{\frac{1}{\gamma}}\ , \quad h\ (\ t\ ) = \Big(\frac{1}{\gamma\ t + \frac{1}{\theta}} \Big)$$ respectively.

When $\gamma \rightarrow 0$, the distribution reduces to the exponential distribution with parameter θ. The likelihood function taking into account censoring is given by

$$\prod_{i=1}^{n} f_i^{\delta_i}\ S_i^{1-\delta_i},$$

and we know that $\frac{f_i}{S_i} = h_i\ (t)$, the hazard function, so the likelihood becomes

$\prod_{i=1}^{n} S_i \, h_i^{\delta_i}$, $\delta_i = 1$, if the event is observed, and $\delta_i = 0$ otherwise.

The log likelihood is

$$l = \sum_{i=1}^{n} \Big\{ - \frac{1}{\gamma} \log (\gamma \, \theta_i \, t_i + 1) - \delta_i [\log (\gamma \, t_i + \frac{1}{\theta_i})] \Big\} .$$

The dispersion parameter is γ. We use the score statistic to test the null hypothesis that $\gamma = 0$. Since γ is in the denominator, we cannot set it to zero, so we expand

$\log (\gamma \, \theta \, t_i + 1)$ in a series,

$$\log(1+x) = x - \frac{1}{2} x^2 + \frac{1}{3} x^3 - \frac{1}{4} x^4 + \cdots,$$

then derive the score statistic,

$$U_i(\hat{\mu}_i) = \frac{\partial \, l}{\partial \tau} \Big|_{\tau = 0} = - \delta_i \, \theta_i \, t_i + \frac{1}{2} \, \theta_i^2 \, t_i^2,$$

and the elements of the information matrix,

$$W_{1i} = - \Big\{ \frac{\partial^2 \, l_i}{\partial \, \theta_i^2} \Big\} \Big|_{\gamma = 0} = - \delta_i \, \theta_i^2 + 2 \, \theta_i^3 \, t_i ,$$

$$W_{2i} = - \Big\{ \frac{\partial^2 l_i}{\partial \mu_i \, \partial \, \gamma} \Big\} \Big|_{\gamma = 0} = - \delta_i \, t_i \, \theta_i^2 + \theta_i^3 \, t_i^2, \text{ and}$$

$$I_\gamma = - \Big\{ \frac{\partial^2 \, l_i}{\partial \, \gamma_i^2} \Big\} \Big|_{\gamma = 0} = - \delta_i \, \theta_i^2 \, t_i^2 + \frac{2}{3} \, \theta_i^3 \, t_i^3 .$$

3.5 Discussion

As mentioned before, for some applications, the existence of overdispersion is the main question. Generally, the presence of overdispersion indicates an inadequacy in the assumed model that must be rectified. For inference in a regression, where we are interested primarily in the effect of covariates, the use of estimating equations rather than the complete specification of an alternative distribution may be a viable, robust approach. However, in the presence of censoring, the full representation of the null survival function is still necessary.

The derivation of the null survival function is straightforward for the mixture of the exponential families, whereas double exponential families require some extra work to evaluate the value of the constant. We have included both situations in our examples. The derivation in the geometric distribution involves a constant not close to one, but the constant can be treated as one in the exponential distribution. Since survival functions and the observed information matrices are involved, the results look somewhat complicated in the formulation of the geometric distribution.

The full representation of the overdispersed distributions is used for comparison. However, the applications are not limited to the known mixtures. The mixture of exponential families does not require the full representation of the mixed distribution. Dean (1992) applied the models to extra Possion variates with (1) additive random effects by letting $\theta_i^* = \underline{X}_i^T \underline{\beta} + z_i$, where the z_i are iid random variables with $E\ (\ z_i\ ) = 0$ and $Var\ (\ z_i\ ) = \tau < \infty$, $i = 1\ ,\ldots,\ n.$; (2) multiplicative random effects by letting $\theta_i^* = \nu_i\ \mu_i$, where $E\ (\ \nu_i\ ) = 1$ and $Var\ (\ \nu_i\ ) = \tau < \infty$, and (3) a simple variance inflation by letting $\theta_i^* = \nu_i\ \mu_i$, where $E\ (\ \nu_i$

$) = 1$ and $Var\ (\ \nu_i\ ) = \tau\ /\ \mu_i$. Our models can deal with the same situations as long as we can define the θ, $a\ (\ \theta\ )$, $g\ (\ \theta\ )$, $b(\theta)$ and the survival function under the null hypothesis.

The double exponential families are similar to the third situation described above, which is with a variance inflation, since the variance is $Var\ (\ \mu_i\ )\ /\ (a\ \phi)$, where a is a known constant and ϕ is the dispersion parameter. The application of the double exponential family to data with censoring requires full representation of the survival function deriving from the double exponential family.

The analytic results obtained in this chapter will be used in the next chapter to study the behavior of the test statistics. In Chapter 4, we study the properties of the test statistics via simulation. The focus of the next chapter is on the power of these test statistics.

CHAPTER IV

SIMULATIONS

Score statistics for testing the existance of overdispersion have been derived in the previous chapter. For evaluation of the tests, our focus is on the robustness and the power ot the test statistics based on either model, rather than on the exact expressions for overdispersion. In this chapter simulation is used to explore the properties of the test statistics developed in Chapter 3. Due to the requirement of explicit representation of the survival function, only the behavior of the test statistics for some specific distributions are studied. The adjusted power (Zhang and Boos, 1994) of the test statistics for different sample sizes and the adjusted power for different degrees of censoring for the same test are compared.

4.1 Methodology

4.1.1. Generating Data

Simulation is nothing but samples drawn from a hypothetical statistical model. It is done by setting the population parameter to known values, selecting a random sample, then estimating the statistics using the sample. In this research, the statistics to be estimated are the score tests for the different models.

A random sample of observations drawn from the overdispersed geometric distribution is accomplished in two steps. The first is to draw a beta variate, then use it to generate a geometric variate.

The procedure for generating *Beta* $(a,\ b)$ for non-negative a and b is as follows (Fishman, 1973):

1. compute k_1 and k_2 to be the largest integers in a and b, respectively;
2. if $k_1 \neq 0$, then generate Y from *Gamma* $(a,1)$ and Z from *Gamma* $(b,1)$;
3. compute the beta variate $X = Y/(Y+Z)$;
4. if $k_1 = 0$ and $k_2 \neq 0$, repeat 2 and 3;
5. if $k_1 = 0$ and $k_2 = 0$, then generate variates U_j and U_{j+1} from uniform(0, 1), and calculate $Y = U_j^{1/a}$ and $Z = U_{j+1}^{1/b}$;
6. if $Y + Z \leq 1$, calculate the beta variate $X = Y/(Y+Z)$; otherwise repeat 5 and 6.

After generating the beta variate, it is used to generate a geometric variate. The geometric variate is generated based on the relationship between the exponential and geometric distributions (Fishman, 1973). Let K be from the exponential distribution with parameter β. Then it is noted that

$$Pr(r \leq K \leq r+1) = \frac{1}{\beta} \int_r^{r+1} e^{-K/\beta}\, dK$$

$$= e^{-r/\beta}\,(\,1 - e^{-1/\beta}), \tag{4.1}$$

which is the probability that a random variable from the geometric distribution $(q = e^{-1/\beta})$ assumes the value r. To generate the survival time T, (4.1) is used with the geometric $(p = 1 - e^{-1/\beta})$. Then $\beta = -1/\log\,(1 - X\,)$,

$$T = [K] + 1 = [-\beta \log U] + 1 = [\log U / \log (1 - X)] + 1,$$

where U is generated from the uniform $(0,1)$, X is the beta variate generated from the previous procedure, and the brackets indicate the largest integer.

It is assumed that the censoring is uniformly distributed throughout the survival time. Therefore, censoring is introduced using the UNIFORM function in SAS with a factor k which determines the percent of censoring. Having generated observations with censoring, we apply them to calculate the test statistics developed in Chapter 3. All of the calculations are carried out in SAS or SAS IML. See Appendix III for more details.

4.1.2. The Adjusted Power

The adjusted power (Zhang and Boos, 1994) is used to evaluate the power of the test statistics, since the asymptotic distribution for the test statistics is not available. The procedures are as follows. First, N_0 independent samples are drawn under the null hypothesis and N_1 samples under the alternative hypothesis. Then the test statistics $T_{01},\ldots,T_{0N_0}$ for the observations drawn from the null distribution and the test statistics $T_{a1},\ldots,T_{aN_1}$ for the samples drawn from the alternative distribution are calculated. In this way, the critical values C_α for the T_0's can be estimated from the $(1-\alpha)$ quantile of the null observed value by ordering the T_0's from the smallest to the largest, and locating the $(1-\alpha)N_0$th value. Finally, the adjusted power is calculated as

$$\widehat{p} = \frac{1}{N_1} \sum_{i=1}^{N_1} I\,(T_{ai} > C_\alpha), \quad \text{where } I \text{ is the indicator.}$$

4.2 Results

4.2.1 The Overdispersed Geometric Distribution

The beta-geometric distributions with $\gamma = 2$ and 0.1, with sample sizes 100, 250, 500, and 1000, and with 0%, 20% and 40% of censoring were simulated. For each situation, 1000 samples of different sample sizes were generated. The adjusted power of the test statistics for $\alpha = 0.05$ were calculated and presented in table 4.1.

Table 4.1 Power of Test Statistics Based on the Mixture of Geometric Distribution.

$\gamma = 2.0$	sample sizes			
% censoring	100	250	500	1000
0	1	1	1	1
20	0.95	1	1	1
40	0.44	0.91	0.99	1
$\gamma = 0.1$				
0	0.75	0.93	0.99	0.99
20	0.44	0.36	0.27	0.20
40	0.31	0.22	0.11	0.40
$\gamma = 0.0^*$				
0	0.050	0.046	0.064	0.048
20	0.059	0.063	0.058	0.059
40	0.050	0.060	0.059	0.074

$*$ The simulation of $\gamma = 0.0$ is to assure that the size of α is close to 0.05.

The estimated power of the tests based on the exact model is 1 when the sample size is 100. That is expected when the full expression of the distribution is available. Therefore, no further calculation for the tests based on the exact distribution is needed. We generate data with $\gamma = 0.0$ to varify the

appropriateness of C_α, which is produced following the steps described in the previous section. The results show that the α levels obtained from simulation are pretty stable.

Table 4.1 indicates that (1) the test statistics have sufficient power to reject the null hypothesis, or to detect the existence of overdispersion when the sample size is reasonably large (> 250) and censoring is mild ($= 40$ %); and (2) when the dispersion parameter is at the boundary of the overdispersion parameter space, i.e., $\gamma = 0.1$, the test statistics do not have the power to detect the existence of overdispersion with little amount of censoring ($\sim 20\%$).

The simulation results also show that normality holds when the dispersion parameter is inside the parameter space. When plotting the score statistics obtained from 1000 samples with $\gamma = 2$, the score statistics calculated under both the null distribution and the overdispersed distribution approach normality. The Shapiro-Wilk statistics for the distribution of the score statistics given by the UNIVARIATE procedure in SAS fail to reject the null hypothesis of a normal distribution (Table 4.2). But when the dispersion parameter is close to the boundary, i.e. $\gamma = 0.1$, the distribution of the test statistics is not normal (Table 4.2). However, the Shapiro-Wilk statistics are very sensitive to very little departure from normality. For example, the histogram of the data appears somewhat normal with a little shift in the peak, the Shapiro-Wilk statistic would reject the hypothesis of normality. When we plot the the score statistics, most of them for $\gamma = 0.1$ are skew to the right, i.e., they concentrate on the values away from the statistics calculated under the null hypothesis, but they have a long tail going towards the distribution of the statistics calculated under the null hypothesis.

Table 4.2 p Values of the Shapiro-Wilk Test of Normality For Tests Based on the Mixture of Exponential Family.

$\gamma = 2.0$	sample sizes			
% censoring	100	250	500	1000
0	0.29	0.46	0.10	0.34
20	0.70	0.42	0.67	0.42
40	0.70	0.20	0.59	0.21
$\gamma = 0.1$				
0	0.0001	0.0001	0.0001	0.0001
20	0.0001	0.0001	0.0001	0.0001
40	0.0001	0.0001	0.0001	0.0001

The power of the test statistics based on the double exponential families is not evaluated due to the complexity of the survival function.

4.2.2 The Overdispersed Exponential Distribution

Similarly, procedures determining the power of the test statistics derived for the overdispersed exponential distributions were carried out. The overdispersed exponential distribution was simulated from the exponential distribution mixed with a gamma distribution. The mean of the exponential distribution is 20, and the parameters of the gamma distribution are $(\frac{1}{4}, \frac{1}{5})$. The gamma variates can be generated from the existing GAMMA function in SAS. Then, the variates are substituted into the parameter of the exponential distribution to get the exponential variates. The exponential distribution is generated following the algorithm developed by Fishman(1973).

Let X have p.d.f.,

$$f_x(X) = \begin{cases} \frac{1}{\rho} e^{-x/\rho}, & 0 \le X \le \infty \\ 0, & X < 0. \end{cases}$$

The distribution function is:

$$F_x(X) = \frac{1}{\rho}\int_0^{\infty} e^{-u/\rho} du = 1 - e^{-x/\rho};$$

then we say X is from the exponential distribution (ρ). We generate a random number U and set

$$U = F_x(X) = 1 - e^{-x/\rho}, \text{ so that } X = -\rho \log (1 - U).$$

Since U is a uniform variate, it is easily seen that $1 - U$ is also a uniform variate. Hence we may save a step by choosing X as $X = -\rho \log (U)$, where ρ is the gamma variate generated in the previous procedure.

Again, we generate samples with different degrees of censoring and different sample sizes and calculate the test statistics for each sample. We apply the method of adjusted power to evaluate the power. Table 4.3 lists the adjusted power of the statistics derived from the double exponential family. The adjusted power of the test statistics derived from the mixture of exponential families or the Pareto distribution is one throughout all the situations.

Although normality is not necessary in determining the power in this study, we examine the distributions of the test statistics to get a better understanding of their behavior. Test statistics based on the mixture of exponential families are distributed normally in most of the situations simulated. On the other hand, the test statistics based on the double exponential family have some extreme values. The power for the statistics deriving from the double exponential family look pretty good at the 95% rejection level, i.e., $\alpha = 0.05$, but they often start with a very long tail from the left, concentrate around a certain point, then have one or two points running into the distribution of the test statistics under the null hypothesis.

Table 4.3 Power of Test Statistics Based on the Double Exponential Formulation of the Exponential Distribution.

$\gamma = 4.0$	sample sizes			
% censoring	100	250	500	1000
0	1	1	1	1
20	0.90	0.86	0.81	0.82
40	0.91	0.91	0.88	0.86
$\gamma = 1.0$				
0	1	1	1	1
20	0.97	0.97	0.97	0.97
40	0.97	0.97	0.97	0.97
$\gamma = 0.0^*$				
0	0.053	0.057	0.050	0.052
20	0.037	0.042	0.043	0.058
40	0.042	0.051	0.061	0.050

$*$ The simuation of $\gamma = 0.0$ is to assure that the size of α is close to 0.05.

4.3 Discussion

In summary, we have studied the power of the test statistics via simulation. The test statistics for the overdispersed geometric distribution have sufficient power when the overdispersion is moderate to large, sample size is reasonably large and censoring is moderate. Normality also holds for the situations mentioned above. The power drops, as well as normality, when the overdispersion parameter is close to zero. The evaluation of power for the test statistics based on the double exponential family is not available due to complications in deriving the elements of the information matrix involving the survival function.

Similar procedures are used for the overdispersed exponential distribution. The test statistics based on the mixture of exponential families have estimated

power equals to one in all the situation, even when the sample size is 100 and the percent of censoring is 40%. They are normally distributed too. The tests based on the double exponential family have high power in all situations simulated, but they are far from normally distributed in situations with 20% or greater censoring. They distribute normally when there is no censoring.

Having learned these properties, we apply the methods to real data. We apply them to both discrete and continuous survival time data in the next chapter.

CHAPTER V

DATA ANALYSIS

We apply the methods to three sets of data, the waiting time to conception, the survival time of patients with myocardial infarction and the duration of breast feeding. These three sets of data provide examples of applying the methods to different (1) types of survival time; (2) sample sizes, and (3) degrees of censoring. The results show that (1) all methods can detect the existence of overdispersion in these three sets of data; (2) the derivation and computation of tests based on the double exponential families is complicated when the constant is not close to one; and (3) the tests based on the double exponential family has higher power than other tests when the sample size is large and the constant is close to one. The benefit of using the proposed methods is that they are not limited to one kind of overdispersion, although the specific expression of the overdispersion represents only one kind of overdispersion.

5.1 Waiting Time to Conception

A convenient measure of fecundability is time (number of menstrual cycles) requires to achieve pregnancy. Couples attempting pregnancy are heterogeneous in the per-cycle probability of success. The methods are applied to a group of women with planned pregnancies (Weinberg and Gladen, 1986). Only women who had become pregnant within 24 months of trying were included. Since medical interventions tend to begin after 12 unsuccessful cycles, times reported beyond 12 cycles are treated as right-censored.

A total of 678 women with planned pregnancies were interviewed, of whom 654 became pregnant within 12 cycles after discontinuing contraception. Women were classified as current smokers if they reported smoking at least an average of 1 cigarette a day during at least the first cycle they were trying to get pregnant. This yielded 135 smokers.

Since we assume that the variation in pregnancy rates over time can be attributed to heterogeneity in fecundability among couples, and not to true time effects, it would be improper to include any group in which a true time effect was suspected. Weinberg and Gladen (1986) thereby excluded 92 women whose most recent method of contraception was given as the pill. This left 586 women contributing to a total of 1844 cycles. When assuming no overdispersion, the maximum likelihood estimates of waiting time to pregnancy taking censoring into account is 3.25 cycles for the whole group; 3.02 cycles for non-smokers and the 4.46 cycles for smokers. Table 5.1 lists three standardized test statistics from the proposed models and the exact distributions. All these large values of the score statistics suggest the existence of overdispersion. This is consistent with Weinberg and Gladen's (1986) finding that fecundability varies among couples, and that a simple geometric model is not appropriate in modeling fecundability. The test based on the double exponential family has less power and higher variance than the other tests. The reason is that the constant is not close to one and an approximation is used in evaluating it.

Table 5.1 Testing Overdispersion in Waiting Time to Conception

Method	standardized score	variance of the score
Mixture Exp	6.16	167.41
Double Exp	5.87	3611.58
Exact	16.82	122.54

5.2 Survival Time After Myocardial Infarction

The second set of data is the survival time after myocardial infarction used by Davis and Feldstein (1979). Appendix D lists survival times for 137 patients who were discharged alive from the hospital after suffering an ECG anterior wall myocardial infarction. These constituted the subset of 429 anterior infarction patients who had both documented left ventricular dysfunction in the coronary care unit and had one or more premature ventricular ectopis beats on a six hour, ambulatory, predischarging Hölter recording. Davis and Feldstein (1979) computed the Kaplan-Meyer product-limit estimates, $\widehat{S}$, of the survival function and plotted $-\log(\widehat{S})$ against t, survival time in months. The graph showed a monotone decreasing hazard. They observed the concave shape of the $-\log(\widehat{S})$ which implied the simple exponential distribution might not fit the data well.

Table 5.2 Tests of Overdispersion in Myocardial Infarction Survival Time.

Method	standardized score	variance of the score
Mixture Exp	6.76	46.91
Double Exp	8.19	20.97
Exact	9.52	25.14

Table 5.2 lists the testing results which indicate the existence of overdispersion. The findings provide formal tests for Davis and Feldstein's (1979) conclusion about overdispersion, and support their use of the Pareto distribution to fit the data.

5.3 Duration of Breast Feeding

It is the consensus of the international pediatric community that breast-feeding is the optimal form of infant nutrition. Benefits of human milk relative to artificial formula, such as protection against infection, allergy, diabetes, obesity and malnutrition, have been well studied (Brent et al., 1995). Recent research on infant dietary energy requirements has confirmed that exclusive breast-feeding provides sufficient nourishment for the average child until four to six months of age (Whitehead, 1995). The literature has documented that the incidence of breast-feeding is affected by mother's prenatal education, whereas the duration of breast-feeding is affected by postpartum management (Brent, et al., 1995). It is not clear if variations in the duration of breast-feeding are due to measured factors such as social-economic status and prenatal education or due to individual differences. We would like to see if heterogeneity exists in the duration of breast-feeding, controlling for some potential factors such as socio-economic status and some health care factors.

5.3.1 Data

Data for this application are from the National Maternal and Infant Health Survey conducted in 1988. The purpose of this nationally representative survey was to identify factors related to poor pregnancy outcome, such as adequacy of prenatal care; inadequate and excessive weight gain during pregnancy; maternal smoking, drinking, and drug use; and pregnancy and delivery complications. Therefore, African-American women and the low-birth-weight babies were over-sampled. A complex sampling strategy was used to allow for adequate statistical representation of the oversampled population. The survey sample was drawn

from 1988 vital records in 48 states (excluding Montana and North Dakota), Washington D.C. and New York City (Sanderson et al., 1991).

Questionnaires collecting information on demographic, economic, behavioral, health status, and health service from respondents about themselves, their families, and their infants were mailed to women sampled. Follow-up attempts for non-response included a second mailing of the questionnaire, a postcard reminder and telephone or personal interviews (Sanderson et al., 1991). Seventy four percent of the mothers with a live birth responded to the survey. Non-responders were more likely to be black and unmarried.

Because our focus is on breast-feeding duration and we have not developed the methods to deal with weighted data, we limited our sample to well born babies to white mothers. The inclusion criteria are (1) birth weight greater than 2250 grams and less than 4500 grams; (2) gestation age between 36 and 42 weeks; (3) singlet birth; (4) infant not transferred immediately after birth; (5) infant went to the mother's home from the hospital; (6) mother stayed three or fewer nights in the hospital for delivery; and (7) white mother. We have also restricted the sample to those women who ever breast fed their baby, since we are interested only in the duration of breast feeding. As a result, 1145 out of 9953 women with live birth remained in our sample. The distribution of the durations of breast-feeding for this group of women is in Appendix E. These durations ranged from 1 to 102 weeks. The average breast-feeding time was 20.8 weeks and the mode was 13 weeks. Those who breast-fed their babies longer than 42 weeks are considered censored, since babies need more food resources than milk after 6 months. This results in 418 instances of right censoring.

Assuming that the duration of breast feeding follows the exponential distribution when no overdispersion exists, we apply the tests derived in chapter three to the data. The results indicate the existence of overdispersion when covariates are not considered (Table 5.3). One may argue that some factors do affect the duration of breast-feeding. The next step is to see if overdispersion exists after controlling for covariates.

Table 5.3 Tests of Overdispersion in the Duration of Breast-Feeding.

Method	standardized score	variance of the score
Mixture Exp	8.53	6159.78
Double Exp	13.23	297.71
Exact	13.27	1820.65

5.3.2 Covariates

Three categories of factors are considered here to have effects on duration of breast-feeding. They are socio-demographic characteristics, economic factors, and health care factors. Each category contains more than one factor (Table 5.4). Behavioral variables, such as use of illicit drugs, are not included in the analysis, since we do not think they are reliable measurements.

A. Socio-demographic characteristics

The effects of Socio-demographic predictors on the incidence of breast-feeding are well established (Akin et al., 1981; Brent et al., 1995; Schwartz et al., 1995; Swigonski et al., 1995). These factors include maternal age and education, whether first birth or not, and geographic location. West (1980) found that the

duration of breast-feeding in Britain was significantly related to social class. It is quite possible that similar relationship between social-demographic factors and the duration of breast-feeding exist in this data set.

Maternal age is recorded in years. It is possible that women make different decisions about when to stop breast-feeding at different ages. To demonstrate the ability of the model to handle continuous variables as well as categorical variables, we keep maternal age as a continuous variable.

The mother's education may reflect her knowledge of the advantages of breast-feeding and of a longer length of breast-feeding. Over half (54%) of this group of women had greater than a high school education, about one-third (35%) only a school education, and about one-eighth (12%) less than that.

Regarding the effects of number of children, the more children, the more competition for the mother's attention. Also the physical condition of women with more children tends to be weaker. Thus, for either of these reasons, birth order may affect the duration of breast-feeding. About two-thirds (67%) of the babies of this group of women were first births.

We notice that different regions of the U.S. seem to discharge women after delivery differently (Margolis et al., 1995). Thus, in the South physicians tend to keep women in the hospital longer after delivery. The length of stay in the hospital after delivery affects instructions received by the women about caring the baby, including breast-feeding. The practice of prenatal care may be different from region to region as well. Therefore, the decision of how long to breast-feed the baby may differ among regions. The regions are grouped into North East and Atlantic, East and West Central, Mountain and Pacific, and South. 17% of the

women were from the North East and Atlantic, about 25% from East and West Central, about 28% from Mountain and Pacific, and 29% from the South.

B. Economic factors

Studies have found that women from low income groups, especially those who participate in the Supplemental Nutrition Program for Women and Children (WIC) have a much lower incidence of breast-feeding than their non-WIC counterparts (Brent et at., 1995; Schwartz et al., 1995). Questions about receiving WIC food for both mother and baby were asked. It is possible that those who received WIC for the baby (23% of our sample) stopped breast-feeding earlier than those who did not. One possible reason is that they participate WIC purposely to get baby formula.

Income levels are indicators of the health care, education and other resources these women had. Low proportion of the women (6%) reported less than $6,000 annual household income, most of them (37%) between $30,000 and $59,999; and the others were evenly distributed in other income groups (Table 5.4). The other economic factor is work status. Questions about work status before and after delivery were asked. No details about how long after delivery the mother went to work is available. More than half of the sample (60%) returned to work after delivery. Mothers who returned to work would more often have to stop breast-feeding earlier than those who did not.

Finally, the survey asked questions related to insurance status at the time of delivery in two ways. One was if the mother had insurance at the time of delivery, the other was how they paid for the delivery. We have combined

answers to these two questions into four categories of insurance: private insurance (77%), Medicaid (7%), government assistance (4%), which includes military, Indian health service and other government assistance other than Medicaid, and other (11%), which could be no insurance or other. Different insurance groups may provide prenatal care including breast-feeding advice differently, so the source of insurance may affect the duration of breast-feeding.

C. *Health care factors*

Kotelchuck's index (Kotelchuck, 1994) was calculated to evaluate the adequacy of prenatal care utilization. Since one important variable in calculating the index was missing from California, the indices for California were imputed using maternal age and education. The indices were combined into two scales: inadequate (23%) and adequate. Adequate prenatal care may provide proper information about breast-feeding, and thus lengthen the duration of breast-feeding.

Any advice for breast-feeding from either prenatal health care provider or WIC might influence the duration of breast-feeding. Advice given during prenatal care may play an important role in the incidence of breast-feeding. If a woman was encouraged to breast feed during prenatal care, she would be very likely to carry it out. About two thirds (66%) of the sample were given advice on breast-feeding. Attending birth class or not also has an effect similar to advice on the duration of breast-feeding. A large percent (78%) of the women attended birth classes.

The length of the hospital stay after the delivery may affect the advice received about breast-feeding a lot. Staying longer after delivery gives health workers longer time to make sure the mothers are comfortable with breast-feeding as well as with caring for the child. But there is a trend in reducing hospital stays due to the cost. We have grouped hospital stays into those who had a short stay in the hospital (0 or 1 night after delivery) and those who had a regular stay (2 or 3 nights).

The other factor related to health care is the type of attendants at delivery. Most women had physicians available at the delivery (96%). Thus, only a few of them had midwives for their delivery. It is possible that midwives give more individual attentions to the women. It is interested to see what kind of influence they may have on duration of breast-feeding.

5.3.3 Test Results

Having controlled for the possible factors affecting breast-feeding, the tests still show the existence of overdispersion (Table 5.5). So a simple exponential model may not be appropriate to represent the duration of breast-feeding.

Table 5.5 Tests of Overdispersion in Duration of Breast-Feeding With Covariates

Method	standardized score	variance of the score
Mixture Exp	6.39	2646.29
Double Exp	21.79	126.30
Exact	10.21	1037.91

Table 5.4 Factors Affecting the Duration of Breast-Feeding

Factors	Groups	Percent
Social-demographic Characteristics		
Age	no*	
Education	1: <HS	12
	2: = HS	35
	3: >HS (control)	54
Birth order	1: first birth (control)	67
	2: not first birth	34
Region	1: Northeastern & Atlantic	17
	2: East, West, North Central	25
	3: Mountain, Pacific	29
	4: South (control)	29
Economic Factors		
WIC for baby	1: yes (control)	23
	2: no	77
Household income	1: < $6,000 (control)	6
	2: $ 6,000-$11,999	9
	3: $12,000-$17,999	11
	4: $18,000-$29,999	25
	5: $30,000-$59,999	37
	6: > $60,000	13
Insurance	1: Medicaid	7
	2: government assistance	4
	3: other	11
	4: private insurance (control)	77
Return to work	1: yes (control)	60
	2: no	40
Health Care Factors		
Prenatal care index	1: inadequate (control)	22
	2: adequate	78
Any breast-feeding advice	1: yes (control)	66
	2: no	34
Nights in the hospital after delivery	1: 0 or 1	19
	2: 2 or 3 (control)	81
Attended birth class	1: yes (control)	78
	2: no	22
Attendants at delivery	1: midwife (control)	4
	2: physician	96

* The year of age is used directly as continuous variable

Knowing that the exponential distribution may not be appropriate to model the duration of breast-feeding, we fit the gamma distribution to demonstrate how much information is lost in fitting the exponential distribution. The reason for choosing the gamma distribution is that the double exponential formulation of the overdispersed exponential distribution leads to the gamma distribution anyway, and also the model is available in SAS.

Table 5.6 shows the log likelihood and the coefficients of the covariates. The gamma distribution maximizes the data well: the log likelihood for the exponential distribution is -1785.81, which for the gamma distribution it is -1719.10. The significance of the coefficients also changes from one model to another. The standard errors of the estimates are also listed in Table 5.6.

The model tells us that those who have higher education, do not return to work after delivery; do not receive WIC food for babies; have annual income between \$12,000 and \$29,999 or greater than \$60,000; or who live in the Mountain/Pacific area tend to breast-feed longer than others.

The results are consistent with West's (1980) finding. She conducted a follow-up study on a group of 239 mothers who were breast-feeding on leaving hospital. She found that duration of breast-feeding was significantly influenced by social class, and for parous mothers, previous breast-feeding success. She also reported two other reasons for stopping breast-feeding early (within 12 weeks): inadequacy of the milk supply and unsettled baby. We would like to see if these two reasons explain the heterogeneity if they were available.

Table 5.6 Coefficients of Factors Affecting Duration of Breast-Feeding

Covariates	Exponential Dist. Estimate (s.e)	Gamma Dist. Estimate (s.e)
Intercept	2.64 (0.339)*	1.90 (0.485)*
Age	0.06 (0.009)*	0.07 (0.012)*
Education (< HS)	−0.35 (0.134)*	−0.42 (0.201)*
(= HS)	−0.15 (0.089)	−0.25 (0.124)*
Birth order (not 1st)	0.20 (0.082)*	0.15 (0.114)
Region		
NE Atlantic	0.26 (0.117)*	0.26 (0.162)
E/W North Central	−0.14 (0.100)	0.0001 (0.143)
Mountain, Pacific	0.36 (0.102)*	0.41 (0.141)*
WIC for baby (no)	0.27 (0.112)*	0.33 (0.161)*
Household income		
$ 6,000-11,999	−0.028 (0.183)	0.40 (0.270)
$12,000-17,999	0.29 (0.190)	0.54 (0.266)*
$18,000-29,999	0.17 (0.175)	0.47 (0.249)
$30,000-59,999	0.014 (0.182)	0.34 (0.259)
$60,000 +	0.16 (0.212)	0.49 (0.295)
Insurance		
Gov't assist.	0.31 (0.192)	0.16 (0.268)
Medicaid	0.09 (0.160)	−0.17 (0.228)
Other	0.18 (0.128)	0.057 (0.178)
Return to work (no)	0.75 (0.082)*	0.82 (0.113)*
Prenatal care index adequate	−0.20 (0.098)*	−0.19 (0.134)
Any BF advice (no)	−0.13 (0.083)	−0.22 (0.114)
Hsp stay (short)	0.17 (0.106)	0.18 (0.143)
Birth class (No)	0.24 (0.099)*	0.12 (0.143)
Attendants (MD)	0.33 (0.180)	0.07 (0.262)

* indicates $p \leq 0.05$.

5.4 Discussion

In this chapter we have applied the tests derived in Chapter 3 to different types of survival time with different degrees of censoring. In the discrete case, we have an overdispersed geometric distribution. All the tests are able to detect overdispersion. The survival part of the test based on the double exponential families is complicated because the constant is not close to 1 and no simplification is available for the summation of survival probability to time t. Fortunately, censoring only occurred at one time point. We substitute the value of that time point to the get the values directly. Then, we put it back with the other parts to calculate the standardized score statistic.

For the continuous survival time, we have the overdispersed exponential distribution. In this case, the survival part of the test based on the double exponential family is not as complicated, because the constant can be set to 1 and the integration is not difficult. All the tests are able to detect overdispersion even when the censoring is heavy, for example 99 of the 137 patients censored in study of the survival time after myocardial infarction.

Heterogeneity exists in these three sets of data. The overdispersion of the first two sets of data has been well studied. The beta-geometric model was fitted to the waiting time to conception data (Weinberg and Gladen, 1986) and the Pareto model was fitted to the survival time after myocardial infarction (Davis and Feldstein, 1979). We fit a gamma model for the duration of breast-feeding, since it's available in SAS and the formulation of the double exponential family for exponential distribution leads to the
gamma distribution.

CHAPTER VI

DISCUSSION AND SUGGESTION FOR FUTURE STUDIES

6.1 Summary and Discussion

Score statistics for detecting the existence of overdispersion in survival time with censoring have been derived based on the mixture of exponential families and the double exponential family. They have been applied to overdispersed discrete and continuous survival time analytically and numerically. Simulation studies indicate that these test statistics derived from the mixture of exponential family approach normalilty when sample sizes are sufficiently large and the degree of censoring is mild. But normality does not hold for the test statistics derived from the double exponential families.

Simulation also show that the test statistics based on the mixture of exponential family have enough power to detect overdispersion when the sample sizes are sufficiently large, the degree of censoring is mild, and the overdispersion parameter in inside the parameter space. The test statistics based on the double exponential families always have enough power in detecting overdispersion in data with censoring.

The test statistics derived in this study are applied to real data. They are able to detect the existence of overdispersion. However, the requirement for explicit expression of the survival function has complicated the derivation, especially for tests based on double exponential families. When the constant in

the double exponential family can not be ignored, nor can the survival function be simplified, the calculation of the variance of the score statistic is extremely intensive. The other limitation of double exponential families is the requirement the distribution of the sum of a random variable, since the expression of the double exponential family is more or less the distribution of $n\phi$ samples of the original one-sample distribution.

Therefore, tests based on the mixture of exponential families are favored. Even though they include only the first two moments of the mixed distribution, it has sufficient power in detecting the existence of overdispersion in data with censoring. The fact that only the survival function under the null hypothesis is needed makes the calculation a lot easier. The only drawback is the information loss due to omitting third and higher moments in the mixing distribution.

In the procedure of examining the existence of overdispersion, we suggest to test the existence of overdispersion without covariates first. If the overdispersion exists, then incorporate covariates and see if the covariates can explain the overdispersion. If the overdispersion still exists, we conclude that the assumed model is not appropriate for the data and that models accommodating overdispersion is needed.

6.2 Suggestions for Future Study

The main purpose of the statistics outlined in this research is to provide a tool to examine the adequacy of a given survival time model. Knowing the existence of overdispersion, we would like to identify the information loss under the wrong model is an area to be investigated. We have derived the information

matrix and evaluated it under the null hypothesis. With some extra work, we should be able to determine the amount of information loss due to mis-specification of the model.

The other issue is to find a better model. Now that we have found the simple model is not appropriate for the data. What model should be used? Models predicting overdispersed mortality as a function of time and the covariates have been proposed (Engel, 1984; O'Hara Hines and Lawless, 1993). Our interest is in predicting survival time as a function of covariates adjusted for overdispersion. Theoretically, both models used in this research are capable of this task. The goodness-of-fit of the models and the properties of the estimates are to be investigated.

We have used specific alternatives to illustrate the derivation and calculation of the statistics. However, in many cases no specific form for the alternative model is postulated. Dean (1993) proposed a simple random effect model which is somewhat attractive. He let the parameter be

$\theta_i = \mu_i = X_i^T\beta$, and $\theta_i^* = X_i^T\beta + z_i$,

where the z_i are iid random variables with $E(z_i) = 0$ and $Var(z_i) = \tau < \infty$, $i = 1,\ldots,n$; the z_i may be considered random effect. The derivation of the score statistic is feasible for mixtures of exponential families. The complication will occur in deriving the survival part of the test based on double exponential families, since a random effect is involved in the survival function.

The heterogeneity can arise from covariates. Ganio et al. (1993) let the dispersion parameter ϕ_i be a function of γ_i and

$$\gamma_i = \lambda + z_i\,\alpha,$$

where z_i are $q \times 1$ vectors of the explanatory variables, and are taken to be centered. In our case, we have re-defined the dispersion parameter $\phi_i = e^{\tau_i}$. If some covariates are involved in the overdispersion parameter τ_i, we have

$$\log \phi_i = \tau_i = \lambda + z_i \alpha.$$

Three sets of variables are involved in the derivation of the score statistic: $\underset{\sim}{\lambda}$, $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$. In the case of double exponential families, the part taking censoring into account can be very messy.

REFERENCES

Akin, J., R. Bilsborrow, D. Guilkey, and B. M. Popkin (1981) The determinants of breast-feeding in Sri Lanka. Demography 18(3): 287-307.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993) Statistical Models Based on Counting Processes. New York, Springer-Verlag, 767p, pp 660-674.

Albert, J. H. and P. A. Pepple (1989) A Bayesian approach to some overdispersion models. Canadian Journal of Statistics (17): 333-344.

Brent, N. B., B. Redd, A. Dworetz, F. D'Amico and J. J. Greenberg (1995) Breast-feeding in a low-income population. Archives of Pediatric and Adolescent Medicine. (149) 7: 789-803.

Breslow, N. E. (1990) Test of hypotheses in overdispersed Poisson regression and others quasi-likelihood models. Journal of the American Statistical Association (85)410: 565-571.

Breslow, N. E. (1989) Score tests in overdispersed GLM's. In The Proceedings of GLIM 89 and the Fourth International Workshop on Statistical Modeling eds. A. Decarli, B. J. Francis, R. Gilchrist and G. U. H. Seber, New York, Springer-Verlag, pp 64-74.

Breslow, N. E. (1984) Extra-Poisson variation in long-linear models. Applied Statistics (33)1: 38-44.

Collings, B. J. and B. H. Margolin (1985) Testing goodness of fit for the Poisson assumption when observations are not identically distributed. Journal of the American Statistical Association (80)390: 411-418.

Consul, P. C. and F. Famoye (1992) Generalized Poisson regression model. Communication Statistics (21)1: 89-109.

Cox, D. R. (1983) Some remarks on overdispersion. Biometrika (70)1: 269-274.

Cox, D. R. and D. V. Hinkley (1974) Theoretical Statistics. Chapman and Hall, London.

Cox, D. R. and D. Oakes (1988) Analysis of Survival Data. Chapman and Hall, London.

Davis, H. T. and M. L. Feldstein (1979) The generalized Pareto law as a model for progressively censored survival data. Biometrika (66) 2: 299-306.

Dean, C. B. (1992) Testing for overdispersion in Poisson and binomial regression models. Journal of the American Statistical Association (87)418: 451-457.

Dean, C. B. and J. F. Lawless (1989) Tests for detecting overdispersion in

Poisson regression models. Journal of the American Statistical Association (84)406: 467-472.

Efron, B. (1986) Double exponential families and their use in generalized linear regression. Journal of the American Statistical Association (81): 709-721.

Engel, J. (1984) Models for response data showing extra-Poisson variation. Statistica Neerlandica (38)3: 159-167.

Fishman, G. S. (1973) Concepts and Methods in Discrete Event Digital Simulation. John Wiley and Sons, New York.

Ganio, L. M. and D. W. Schafer (1992) Diagnostics for overdispersion. Journal of the American Statistical Association (87) 419: 795-804.

Garren, S, R. Smith and W. Piegorsch (1994) Bootstrap goodness-of-fit tests for the beta-binomial model. Mimeo Series # 2314 Department of Statistics, University of North Carolina at Chapel Hill.

Heckman, J. J. and B. Singer (1982) Population heterogeneity in demographic models. in K. C. Land and A. Rogers (ed.) Multidimensional Mathematical Demography. Academic Press, New York.

Heckman, J. J., R. Robb, and J. R. Walker (1990) Testing the mixture of exponential hypothesis and estimating the mixing distribution by the method of moments. Journal of the American Statistical Association (85)410: 582-589.

Heckman, J. J. and J. R. Walker (1990) Estimating fecundability from data on waiting time to first conception. Journal of the American Statistical Association (85)410: 283-294.

Hougaard, P. (1986) Survival models for heterogeneous population derived from stable distributions. Biometrika (73) 2: 387-396.

Hougaard, P. (1984) Life table methods for heterogeneous population: distributions describing the heterogeneity. Biometrika (71) 1: 75-83.

Jewell, N. P. (1982) Mixture of exponential distributions. The Annals of Statistics (10)2: 479-484.

Kalbfleisch, J. D. and R. L. Prentice (1980) The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York.

Keiding, N., P. K. Andersen and K. Frederiksen (1990) Modeling excess mortality of the unemployed: choice of scale and extra Poisson variability. Applied Statistics (39) 1: 63-74.

Kim, B. S. and B. H. Margolin (1992) Testing goodness of fit of a multinomial model against overdispersion alternatives. Biometrics (48): 711-719.

Kotelchuck M. (1994) An evaluation of the Kessner adequacy of prenatal care index and a proposed adequacy of prenatal care utilization index. American Journal of Public Health (84) 9: 1414-1420.

Kupper, L. L. and J. K. Haseman (1978) The use of a correlated binomial models for the analysis of certain toxicological experiments. Biometrics (34): 69-76.

Liang, K-Y and P. McCullagh (1993) Case studies in binary dispersion. Biometrics (49): 623-630.

Margolin, B. H., N. Kaplan and E. Zeger (1981) Statistical analysis of the Ames Salmonella/microsome test. Proceedings of the National Academy of Sciences: Genetics. USA (78) 6: 3779-3783.

Margolis, L., M. Kotelchuck and H-Y Chang (1995) Factors associated with early discharge after delivery. Submitted to American Journal of Public Health.

McCullagh, P. and J. A. Nelder (1989) Generalized Linear Models (2nd edition). Chapman and Hall, London.

Mendenhall, W. and R. J. Hader (1958) Estimation of parameters of mixed exponential distributed failure time distributions from censored life test data. Biometrika 45: 504-520.

Moran, P. A. P. (1971) Maximum likelihood estimation in non-standard conditions. Proceedings of the Cambridge Philosophical Society. (70): 441 -450.

Nelder, J. A. and Y. Lee (1992) Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. Journal of Royal Statical Society-B (54): 273-284.

Nelder, J. A. and D. Pregibon (1987) An extended quasi-likelihood function. Biometrika (74): 221-232.

O'Hara Hines, R. J. and J. F. Lawless (1993) Modeling overdispersion in toxicological mortality data grouped over time. Biometrics (49): 107-121.

Pack, S. E. (1986) Hypothesis testing for proportions with overdispersion. Biometrics (42): 967-972.

Paul, S. R., K. Y. Liang, and S. G. Self (1989) On testing departure from the binomial and multinomial assumptions. Biometrics (45) 1: 231-236.

Paul, S. R. and R. L. Plackett (1978) Inference sensitivity for Poisson mixtures. Biometrika (65) 3: 591-602.

Prentice, R. L. (1986) Binary regression using an extended Beta-binomial

distribution, with discussion of correlation indices by covariate measurement errors. Journal of the American Statistical Association (81) 394: 321-327.

Sanderson, M., P. J. Placek, and K. G. Keppel (1991) The 1988 national maternal and infant health survey: design, content, and data availability. Birth (18): 26-32.

Schwartz, J. B., B. M. Popkin, J. Tognetti, and N. Zohoori (1995) Does WIC participation improve breast-feeding practices? American Journal of Public Health (85) 5: 729-731.

Sheps, M. C. and J. A. Menken (1973) Mathematical Models of Conception and Birth. University of Chicago Press, Chicago.

Suchindran, C. M. (1972) Estimators of Parameters in Biological Models of Human Fertility. Ph. D. Dissertation, Department of Biostatistics, University of North Carolina at Chapel Hill.

Swigonski, N. L., C. S. Skinner, and F. D. Wolinsky (1995) Prenatal health behavior as predictors of breast-feeding, injury, and vaccination. Archive of Pediatric and Adolescent Medicine (149) 4: 380-385.

Vaupel, J. W., K. G. Manton and E. Stallard (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography (16)3: 439-454.

Wedderburn, W. M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton methods. Biometrika (61) 3: 439-447.

Weinberg, C. R. and B. C. Gladen (1986) The beta geometric distribution applied to comparative fecundability studies. Biometrics (42): 547-560.

West, C. P. (1980) Factors influencing the duration of breast-feeding. Journal of Biosocial Science (12): 325-331.

Whitehead, R. G. (1995) For how long is exclusive breast-feeding adequate to satisfy the dietary energy needs of the average young baby? Pediatric Research (37)2: 239-243.

Williams, D. A. (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics (31): 945 -952.

Zeger, S. L. and S. L. Edelstein (1989) Poisson regression with a surrogate X; an analysis of vitamin A and Indonesian children's mortality. Applied Statistics (38) 2: 309-318.

Zhang, J. and D. D. Boos (1994) Adjusted power estimates in Monte Carlo experiments. Communications in Statistics: Simulation and Computation (23) 1: 165-173.

APPENDIX A

THE CONSTANT OF THE OVERDISPERSED GEOMETRIC DISTRIBUTION BASED ON THE DOUBLE EXPONENTIAL FAMILIES

$$\frac{1}{c} = \frac{e^{\tau(e^{\tau}+\frac{1}{2})}\ (\mu-1)\ e^{\tau}(e^{\tau}-1)!\ (\mu e^{\tau})^{(e^{\tau}-1)}\ q_t{}^{(e^{\tau}\mu-1)}}{(e^{\tau}\mu-1)!\ (1-\frac{1}{e^{2\tau}\mu}\ )^{e^{\tau}(\mu-)}}$$

$$\{\ 1+\frac{1}{2}\ e^{3\tau}q_t\ \mu^2\ [\Psi(1,e^{\tau}(\mu-1)+1)\ +(\Psi(\ e^{\tau}(\mu-1)+1)\ )^2$$

$$-2\ \Psi(1,e^{\tau}(\mu-1)+1)\ \Psi(e^{\tau}\mu)\ +\frac{2}{e^{\tau}\mu}\ \Psi(e^{\tau}(\mu-1)+1)\ (e^{\tau}-1)$$

$$+2\ \Psi(e^{\tau}(\mu-1)+1)\ (\ \log\ q_t\ +\frac{\mu\ e^{\tau}-1}{\mu^2 e^{2\tau}\ q_t}\ )$$

$$-2\ \Psi(e^{\tau}(\mu-1)+1)\ (\ \log\ (1-\frac{1}{\mu\ e^{2\tau}})+\frac{\mu-1}{\mu^2\ e^{2\tau}(1-\frac{1}{\mu e^{2\tau}})})$$

$$-\ \Psi(1,e^{\tau}\mu)+(\Psi(e^{\tau}\mu))^2-\frac{2\ (e^{\tau}-1)\Psi(e^{\tau}\mu)}{\mu\ e^{\tau}}$$

$$-2\ \Psi(e^{\tau}\mu)\ (\ \log\ q_t\ +\frac{\mu\ e^{\tau}-1}{\mu^2 e^{2\tau}\ q_t}\ )$$

$$+2\ \Psi(e^{\tau}\mu)(\ \log\ (1-\frac{1}{\mu\ e^{2\tau}})+\frac{\mu-1}{\mu^2\ e^{2\tau}(1-\frac{1}{\mu e^{2\tau}})})-\frac{(e^{\tau}-1)}{\mu^2 e^{2\tau}}$$

$$+\frac{(e^{\tau}-1)^2}{\mu^2 e^{2\tau}}+\frac{2\ (e^{\tau}-1)}{\mu\ e^{\tau}}(\ \log\ q_t\ +\frac{\mu\ e^{\tau}-1}{\mu^2 e^{2\tau}\ q_t}\ )$$

$$-2\,\frac{(e^{\tau}-1)}{\mu\,e^{2}}\,(\,\log\,(1-\frac{1}{\mu\,e^{2\tau}})+\frac{\mu-1}{\mu^{2}\,e^{2\tau}(1-\frac{1}{\mu e^{2\tau}})})$$

$$+\frac{1}{q_t}\,(\,\frac{2}{\mu^{2}\,e^{2\tau}}-\frac{2\,(\mu\,e^{\tau}-1)}{\mu^{3}e^{3\tau}}\,-\frac{\mu\,e^{\tau}-1}{\mu^{4}e^{4\tau}q_t}\,)$$

$$+(\,\log\,q_t\,+\frac{\mu\,e^{\tau}-1}{\mu^{2}e^{2\tau}\,q_t}\,)^{2}$$

$$-2\,(\,\log\,q_t\,+\frac{\mu\,e^{\tau}-1}{\mu^{2}e^{2\tau}\,q_t}\,)(\,\log\,(1-\frac{1}{\mu\,e^{2\tau}})+\frac{\mu-1}{\mu^{2}\,e^{2\tau}(1-\frac{1}{\mu e^{2\tau}})})$$

$$-\frac{1}{(1-\frac{1}{\mu\,e^{2\tau}})}\,(\,\frac{2}{\mu^{2}e^{3\tau}}-\frac{2\,(\mu-1)}{\mu^{3}\,e^{3\tau}}-\frac{(\mu-1)}{\mu^{4}\,e^{5\tau}(1-\frac{1}{\mu\,e^{2\tau}})}\,)$$

$$+\,(\,\log\,(1-\frac{1}{\mu\,e^{2\tau}})+\frac{\mu-1}{\mu^{2}\,e^{2\tau}(1-\frac{1}{\mu e^{2\tau}})})^{2}\,]\,\},$$

where $\Psi(x)=\frac{\partial}{\partial x}\,\mathrm{Gamma}(x)$, $\Psi(n,x)$ is the nth derivative of $\Psi(x)$, that is

$\frac{\partial}{\partial x^{(n)}}\,[\mathrm{Gamma}(x)]^{(n)}$, and $q_t=(1-\frac{1}{\mu\,e^{\tau}})$.

APPENDIX B

EXAMPLE OF MAPLE PROGRAM FOR DERIVATION

```
rg:=y**(theta-1)*u**(-theta)*(1-1/y)**(y-1)/(1-1/(theta*y))**(y-theta)*
     (1-1/(theta*u))**(y-theta);          /*      define the function of rg     */;
s:=sum(rg,y=2..t);                        /*      define the survival function */;
theta:=exp(tau);
st:=diff(s,tau);
st2:=diff(s,tau$2);
su2:=diff(s,u$2);
stu:=diff(st,u);
su:=diff(s,u);
tau:=0;                                   /* evaluate under null hypothesis     */;
print(su);
print(stu);
print(st2);
print(su2);
print(st);
quit
```

APPENDIX C

ELEMENTS OF INFORMATION MATRIX FOR OVERDISPERSED EXPONENTIAL DISTRIBUTION BASED ON DOUBLE EXPONENTIAL FAMILY

$$W_{1i} = -\left\{\frac{\partial^2 l_i}{\partial \mu_i^2}\right\}|_{\tau=0} = \frac{\delta_i}{2}\frac{1}{\mu_i^2} - (1-\delta_i)\left\{\frac{1}{S_0}\left\{\frac{1}{\mu_i^2}\left[-2\left(1-e^{-\frac{t_i}{\mu_i}}\right)\right.\right.\right.$$

$$\left. + 2\left(e^{-\frac{t_i}{\mu_i}}\left(1-\frac{t_i}{\mu_i}\right)+1\right) + e^{-\frac{t_i}{\mu_i}}\frac{t_i^2}{\mu_i^2}\right\} + \frac{1}{S_0^2}\left\{\frac{1}{\mu_i}\left(1-e^{-\frac{t_i}{\mu_i}}\right)\right.$$

$$\left.\left. -\left[e^{-\frac{t_i}{\mu_i}}\left(-1-\frac{t_i}{\mu_i}\right)+1\right]\right\}^2\right\},$$

$$W_{2i} = -\left\{\frac{\partial^2 l_i}{\partial \mu_i\, \partial\, \tau}\right\}|_{\tau=0} = \frac{\delta_i}{2}\frac{(t_i-\mu_i)}{\mu_i^2} - (1-\delta_i)\left\{\frac{1}{S_0}\left\{\frac{1}{\mu_i}\left(1-e^{-\frac{t_i}{\mu_i}}+\frac{t_i}{\mu_i}\right)\right.\right.$$

$$+\log(\mu_i)\frac{t_i}{\mu_i}e^{-\frac{t_i}{\mu_i}} + \frac{1}{\mu_i}\left(1-e^{-\frac{t_i}{\mu_i}}\right) + \frac{1}{\mu_i}\left[e^{-\frac{t_i}{\mu_i}}\left(1+t_i-\log(t_i)\right)\right.$$

$$\left. -1+\log(t_i)\right] - \frac{1}{\mu_i}\left[e^{-\frac{t_i}{\mu_i}}\left(1+\frac{t_i}{\mu_i}+\frac{t_i^2}{\mu_i^2}-\log(t_i)-\frac{t_i}{\mu_i}\log(t_i)\right)-1\right.$$

$$\left.\left. +\log(t_i)\right]\right\} + \frac{1}{S_0^2}\left[\left(1-e^{-\frac{t_i}{\mu_i}}\right)(\log(\mu_i)-1) - \left(e^{-\frac{t_i}{\mu_i}}\left(1+\frac{t_i}{\mu_i}-\log(t_i)\right)\right.\right.$$

$$\left.\left. -1+\log(t_i)\right)\frac{t_i}{\mu_i^2}e^{-\frac{t_i}{\mu_i}}\right]\Bigg\}, \text{ and}$$

$$I_\tau = -\left\{\frac{\partial^2 l_i}{\partial \tau_i^2}\right\}\Big|_{\tau=0} = \frac{\delta_i}{2} D(y_i, \mu_i) - (1-\delta_i)\Big\{ \frac{1}{S_0}\Big\{2(1- e^{-\frac{t_i}{\mu_i}})$$

$$(\log(\mu_i) - 1) - 2[\, e^{-\frac{t_i}{\mu_i}}(1 + \frac{t_i}{\mu_i} - \log(t_i)) - 1 + \log(t_i)\,]$$

$$+ (1 - e^{-\frac{t_i}{\mu_i}} \frac{\pi}{6}) - (1 - e^{-\frac{t_i}{\mu_i}})(\log(\mu_i))^2 + 2[\, e^{-\frac{t_i}{\mu_i}} \log(\mu_i)$$

$$(1 + \frac{t_i}{\mu_i} - \log(t_i)) - 1 + \log(t_i)\,] - [\, e^{-\frac{t_i}{\mu_i}}(1 - \frac{t_i}{\mu_i} + \log(t_i) - \frac{t_i^2}{\mu_i^2}$$

$$+ 2\frac{t_i}{\mu_i}\log(t_i) - (\log(t_i))^2\,] + 1 - \log(t_i) + (\log(t_i))^2]$$

$$- \frac{1}{S_0^2}[\,(1 - e^{-\frac{t_i}{\mu_i}})(\log(\mu_i) - 1) - e^{-\frac{t_i}{\mu_i}}(1 + t_i - \log(t_i)) - 1$$

$$+ \log(t_i)\,]^2 \Big\}.$$

APPENDIX D

SURVIVAL TIMES (IN MONTHS) OF HIGH RISK PATIENTS FROM HEART RESEARCH FOLLOW-UP STUDY (Davis and Feldstein, 1979).

1.067	0.633	0.700	0.700	0.733	0.967	1.000
4.333	1.267	2.300	2.367	2.933	3.733	3.833
6.467	4.600	5.633	5.833	5.833	5.867	6.100
13.267*	7.300	7.800	8.900	9.300	12.300*	12.567*
14.800*	13.600*	14.033*	14.067*	14.267*	14.467*	14.500
18.033*	14.933	15.000*	17.267	17.433*	17.500*	17.967*
19.533*	18.533	19.067*	19.133*	19.333*	19.367*	19.433*
21.900*	20.233	20.267*	20.267*	20.667*	20.967*	21.600*
24.933*	22.133*	22.800*	23.833	24.367*	24.433*	24.433*
26.933*	25.600*	25.667	26.167	26.267*	26.367*	26.900*
27.567*	26.967*	27.200*	27.367*	27.400*	27.533*	27.533*
31.267*	28.167*	28.267*	28.667*	29.867*	29.967*	30.233
36.833*	31.733*	32.133*	32.500*	34.100	34.133*	36.200*
41.867*	37.333*	38.000*	38.733*	39.167*	39.267	41.067*
44.967*	41.967*	42.000*	42.100*	42.267*	42.367	43.667*
49.833*	45.333*	46.067*	46.167*	46.167	48.367*	49.433*
51.833*	50.100*	50.167*	50.200*	50.500*	50.500*	50.967*
55.367*	52.367*	52.533*	52.633*	52.767*	53.533*	54.567*
58.033*	55.700*	56.533*	56.867*	57.467*	57.800*	57.967*
58.033*	58.233*	59.467*	60.167*			

Patient months are measured from date entered into the study, roughly two weeks after their infarction. The 137 patients did not enter the study at the same time. Censoring is denoted by $*$.

APPENDIX E
DISTRIBUTION OF DURATION OF BREAST-FEEDING

Distribution of Duration of Breast-Feeding (weeks).

Weeks	counts	%	cum %	Weeks	counts	%	cum %
1	55	4.8	4.8	29	1	0.1	72.6
2	22	1.9	6.7	31	53	4.6	77.2
3	61	5.3	12.1	33	1	0.1	77.3
4	33	2.9	14.9	35	58	5.1	82.4
5	63	5.5	20.4	38	1	0.1	82.5
6	23	2.0	22.4	39	47	4.1	86.6
7	52	4.5	27.0	43	40	3.5	90.0
8	8	0.7	27.7	46	1	0.1	90.1
9	103	9.0	36.7	48	24	2.1	92.2
10	2	0.2	36.9	52	41	3.6	95.8
11	8	0.7	37.6	56	7	0.6	96.4
12	2	0.2	37.8	61	9	0.8	97.2
13	139	12.1	49.9	65	13	1.1	98.3
14	1	0.1	50.0	69	8	0.7	99.0
16	4	0.3	50.3	73	1	0.1	99.1
18	86	7.5	57.8	78	7	0.6	99.7
20	3	0.3	58.1	86	1	0.1	99.8
22	59	5.2	63.2	91	1	0.1	99.9
26	106	9.3	72.5	102	1	0.1	100.0

APPENDIX F
EXAMPLES OF SAS PROGRAMS FOR SIMULATION

i. Program for Overdispersed Geometric Distribution.

```
*-------------------------------------------------------------;
*purpose: generating a beta-geometric distributon             ;
*         with p=0.4 and r=2                                  ;
*         calculating standardized score using Dean's method  ;
*         with no covariate, with censoring                   ;
*                                                             ;
*programmer: Hsing-Yi Chang                                   ;
*                                                             ;
*Date: 1/2/95                                                 ;
*                                                             ;
*Filename: /sas/b_g_dsc1.sas                                  ;
*                                                             ;
*Modified by:                              Date:              ;
*-------------------------------------------------------------;
libname out '~';
options ls=80 ps=59 pageno=1;
proc iml;
%let seed=5071996;
aa={0.2}; bb={0.3};
k1=int(aa); k2=int(bb);

start beta_geo;
create s var{sc,pi,ti,v2,ci};
do j=1 to 1000;
sc=0; v2=0;
   free b_g;
   do i=1 to 500;
         free x; free t;
 if k1=0 then do;
```

```
         if k2=0 then do;
         run unidist;
      end;
      else do;
      run gamdist;
      end;
      end;
      run geom;
      if (t > 600) then t=600;
      cens=round(500#ranuni(&seed),1.);
      delta=(t <= cens);
      b_g=b_g//(x||t||delta);
   end;

pp=b_g[,1];
t_=b_g[,2];
ti=t_[+,1]/i;
i1=j(500,1,1);
pi=pp[+,]/i;
p_i=0.4#i1;
d=b_g[,3];

a1=-1/(i1-p_i);
a2=-1/(i1-p_i)##2;
g1=-1/(p_i#(i1-p_i));
g2=(i1-2#p_i)/(p_i##2#(i1-p_i)##2);
b=p_i#(i1-p_i);
b1=(i1-2#p_i);
a3=-2/(i1-p_i)##3;
g3=-2#(i1-3#p_i+3#p_i##2)/(p_i##3#(i1-p_i)##3);
w1a=d/(i1-p_i)##2#((i1-2#p_i)/p_i##2+t_);
w2a=d/2#(b#((g3-a3#t_)-2#(a1#t_-g1)#(a2#t_-g2))
   -b1#((a2#t_-g2)+(a1#t_-g1)##2));
i_ra=d#b##2/4#((a1#t_-g1)##2+(a2#t_-g2))##2;
```

```
w1b=(i1-d)#t_/(i1-p_i)##2;
w2b=(i1-d)/2#(-b#t_##2#(t_-i1)/(i1-p_i)##3-((b1#t_#(t_-i1)#(i1-p_i)-
    b#t_#(t_-i1)#(t_-2#i1))/(i1-p_i)##3));
i_rb=(i1-d)#b##2/4#t_##2#(t_-i1)##2/(i1-p_i)##4;
w1=w1a+w1b;
w2=w2a+w2b;
i_r=i_ra+i_rb;

sc_0=d/(2#(i1-p_i))#(p_i#t_##2-p_i#t_-2#t_+2#i1)+
     (i1-d)#b#t_#(t_-i1)/(i1-p_i)##2;
i11=ginv(w1);
v2=i_r[+,]-w2'*diag(i11)*w2;
v2=v2[+,];
sc=sc_0[+,1]/sqrt(v2);
c=i1-d;
ci=c[+,]/i;

s=sc||pi||ti||v2||ci;
append;
end;
finish;
*-------------------------------------------------------------;
*    generating beta variates from uniform dist.              ;
*-------------------------------------------------------------;
start unidist;
      y=1; z=1;
      do while (1 < y+z);
      u1=ranuni(&seed);
      u2=ranuni(&seed);
      y = u1 ## (1/aa);
      z = u2 ## (1/bb);
      end;
      x=y/(y+z);
finish;
```

```
*-------------------------------------------------------------;
*      generating beta variates from gamma dist.              ;
*-------------------------------------------------------------;
start gamdist;
      y=rangam(&seed, aa);
      z=rangam(&seed, bb);
      x=y/(y+z);
finish;
*-------------------------------------------------------------;
*     generating geometric variates based x                   ;
*-------------------------------------------------------------;
start geom;
tt=ranuni(&seed);
if (x < 0.0001) then t=999;
else if (x=1) then t=1;
else do;
t1=log(1-tt)/log(1-x);
t=int(t1)+1;
end;
finish;
do;
run beta_geo;
end;
proc means data=s;
var sc ti pi v2 ci;
title1 "generated beta-gemonetric w/ theta=0.4";
title2 "r=2, with ~20% censoring";
title3 "standardized score: Dean's method";
proc univariate data=s plot normal;
var sc;
run;
```

ii. Program for Overdispersed Exponential Distribution.

```
options pageno=1 ls=80 ps=59;
*------------------------------------------------------------;
*DATE:3/18/1996                    PROGRAMER:HSING-YI CHANG   ;
*                                                             ;
*PURPOSE: generating exponential gamma distribution, with     ;
*         aplha=1/4, beta=1/5 & the mean of exp is 20         ;
*                                                             ;
*REQUESTED BY: HSING-YI CHANG (FOR DISSERTATION)              ;
*                                                             ;
*FILENAME:a:exp_gam1.sas                                      ;
*                                                             ;
*MODIFIED BY:                      DATE:                      ;
*------------------------------------------------------------;
%macro exp_gam(iter);
%do j=1 %to &iter;
data gamma;
     U=20; THETA=1/U;
     retain seed1 0 seed2 0 seed 0;
     a=1/4;
     i=0;
do until (i=100);
     call rangam(seed1, a, x);
     mu= 1/5*x;
     call ranuni(seed2, au);
     y = - 1/mu* log(au);
     if (y > 104) then y=104;
     cens=round(90.5*uniform(seed),1.);
     delta=(cens <= y);
     i=i+1;
output;
end;
data new; set gamma;
```

```
*************************************************************;
*SCORE STATISTIC: DEAN''S METHOD                             ;
*************************************************************;
SC_D=delta*THETA/2*(THETA*y**2-2*y)+(1-delta)/2*THETA**2*y**2;
W1_D=delta/THETA**2;
W2_D=-delta*y;
IR_D=delta*THETA**2/4*(THETA*y**2-2*y)**2+(1-delta)/4
     *(THETA**4*y**4);
*************************************************************;
*SCORE STATISTIC: DOUBLE EXPONENTIAL FAMILY                  ;
*************************************************************;
T=y; D=delta;
SC_DE = D*(3.0/2.0+log(T)-T/U)+(1-D)*(U*exp(-T/U)/2-U/2-T*exp(-T/U)+
            U*exp(-T/U)*log(T)-U*log(T))/(1+U*exp(-T/U)-U);

IR_DEN = D*(log(T)-T/U+1)+(1-D)*(-5.0/4.0*U+2*U*exp(-T/U)*log(T)-2*T
    *exp(-T/U)-2*U*log(T)+5.0/4.0*U*exp(-T/U)+U*exp(-T/U)*log(T)**2-
   U*log(T)**2-2*T*exp(-T/U)*log(T)+T**2/U*exp(-T/U))/(1+U*exp(-T/U)
   -U)-(1-D)*(U*exp(-T/U)/2-U/2-T*exp(-T/U)+U*exp(-T/U)*log(T)-U*
   log(T))**2/(1+U*exp(-T/U)-U)**2;

W1_DEN=-2*D*T/U**3+(1-D)/U**3*T**2*exp(-T/U)/(1+U*exp(-T/U)-U)-(1-D)
       *(-exp(-T/U)-1/U*T*exp(-T/U)+1)**2/(1+U*exp(-T/U)-U)**2;

W2_DEN = D*T/U**2+(1-D)*(exp(-T/U)/2+1/U*T*exp(-T/U)/2-1/2-
         T**2/U**2*exp(-T/U)+exp(-T/U)*log(T)+1/U*T*
         exp(-T/U)*log(T)-log(T))/(1+U*exp(-T/U)-U)+(1-D)*
         (U*exp(-T/U)/2-U/2-T*exp(-T/U)+U*exp(-T/U)*
       log(T)-U*log(T))/(1+U*exp(-T/U)-U)**2*(-exp(-T/U)-1/U*T
       *exp(-T/U)+1);
W1_DE=-W1_DEN; IR_DE=-IR_DEN; W2_DE=-W2_DEN;
*************************************************************;
*SCORE STATISTIC: EXACT METHOD (PARETO)                      ;
*************************************************************;
```

```
SC_P=-delta*y/U+1/2*y**2/U**2;
W1_P=-delta/U**2+2*y/U**3;
W2_P=-delta*y/U**2+y**2/U**3;
IR_P=-delta*y**2/U**2+2/3*y**3/U**3;

proc means data=new sum noprint;
var SC_D W1_D W2_D IR_D SC_DE W1_DE W2_DE IR_DE
    SC_P IR_P W1_P W2_P;
output out=outsc sum=ssc_d sw1_d sw2_d sir_d
    ssc_de sw1_de sw2_de sir_de ssc_p sir_p sw1_p
    sw2_p;
data sdsc; set outsc;
     v2_d=sir_d-sw2_d**2/sw1_d;
     sc_dn=ssc_d/sqrt(v2_d);
     v2_de=sir_de-sw2_de**2/sw1_de;
     sc_den=ssc_de/sqrt(v2_de);
     v2_p=sir_p-sw2_p**2/sw1_p;
     sc_pn=ssc_p/sqrt(v2_p);
proc append base=sc data=sdsc;
%end;
%mend exp_gam;
%exp_gam(1000);
proc univariate data=sc normal plot;
title1 'sample size=100', 1000 samples;
title2 'alpha=1/4, beta=5, 60% censoring ';
var sc_dn sc_den sc_pn;
run;
```