Weighted Likelihoods for the NEF-QVF Family with Application

Malay Ghosh[1], James V Zidek[2], Tapabrata Maiti[3] and Rick White[2]

[1] Department of Statistics

University of Florida

Box 118545

Gainesville, Florida 32611-8545

[2] Department of Statistics

6356 Agricultural Road

University of British Columbia

Vancouver, British Columbia

Canada V6T 1Z2

[3] Department of Statistics

Iowa State University

Snedecor Hall

Ames, Iowa 50011-1210

ABSTRACT

The paper introduces a weighted likelihood concept for the estimation of parameters in natural exponential families with quadratic variance functions. The results are applied to both simulated and real data.

1 Introduction.

The likelihood has undoubtedly been one of the greatest contributions to statistics. Its associated estimators, tests and their large sample theory have helped to make it one of the most powerful tools in the practitioner's tool box. Moreover, the sample average as the MLE for estimating a normal mean, became the premier descriptive statistic in common use.

However, an increasing knowledge base and improvements in technology led to ever-expanding scales for experimentation. Consequently, the likelihood encountered hurdles that necessitated the development of new forms such as the profile- and partial-likelihood.

In spite of all adaptations and extensions that had shored up the likelihood, the issue of combining sample information from diverse sources seems to have remained outside the framework of its theory until the emergence of the weighted likelihood (WL). Hu and Zidek (2002) describe the origins of the WL that was popularized by Tibshirani and Hastie (1987) as the "local likelihood". The local likelihood contended with that problem, at least for "neighboring" samples in non-parametric regression when estimating the regression function. The local likelihood came to be very well-studied within the domain of that theory (e.g. Fan and Gijbels, 1996).

In Hu's thesis (Hu, 1994) the local likelihood is taken outside the domain of non-parametric regression and called "relevance weighted likelihood"(REWL). Moreover, Hu and Zidek (2000, 2002) show how Stein's famous estimator can be derived as a REWL estimator provided the weights are estimated from the sample. This work shows therefore that the likelihood framework can in fact be extended to address the issue of combining the information in samples from diverse populations.

The main objective of this article is to apply the technique of Hu and Zidek (2000, 2002) to generate simultaneous estimators of the means of several distributions each belonging to the one-parameter natural exponential family with quadratic variance function (NEF-QVF). Morris (1982, 1983) characterized the NEF-QVF distributions. The six basic distributions belonging to this family are: (i) the binomial; (ii) the Poisson; (iii) the negative binomial; (iv) the normal with known variance; (v) the gamma; (vi) the generalized hyperbolic secant. Morris (1982) considered probabilistic properties of these distributions, while Morris (1983) proposed frequentist and Bayesian inference for parameters of these distribution. We propose instead the WL methodology for inference about the parameters of these distributions and examine the advantages that accrue from using this alternative approach.

Sarkar and Ghosh (1998) obtained EB estimators simultaneously for several means assuming NEF-QVF populations. The EB approach requires modeling the parameters involved in the likelihood function. Instead, the present WL approach avoids modeling the parameters but nevertheless allows strength to be borrowed.

The WL differs from the classical likelihood. The usual assumption associated with the latter is that all the observations come from the same study population. Instead, the WL arises in parametric inference when in addition to the sample from the study population, relevant but independent samples from other populations are also available. This is clearly desirable in a compound decision setting. By down weighting these other samples according to their degree of relevance, the WL incorporates their information while downplaying their bias. In the process, strength is borrowed from all the samples.

The outline of the remaining sections is as follows. In Section 2, we first introduce the WL and then find the maximum WL estimators (the MWLE's) of the population means of NEF-QVF populations. We propose a data-based method of selecting the relevance weights. In particular, in Section 3, we derive these weights for an exchangeable model. We provide also an asymptotic expansion of the mean squared error of the vector of estimators in this section. In Section 4, we find these weights for slightly more general models. The most general version of these weights is derived in Section 5. A simulation study is undertaken in Section 6 to compare the three types of weighted likelihood estimators derived in Sections 3-5. Section 7 contains an illustrative application. Finally, some concluding remarks are made in Section 8. The proof of a technical result is deferred to in the Appendix.

2 The WL for the Exponential Family.

This section introduces the WL for the one-parameter exponential family without necessarily imposing the quadratic variance structure. Also, expressions for the MWLE's of the population means are found under the same general structure.

For the m populations, the independent sample vectors are denoted by $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T,\ i = 1, \ldots, m$ with $n_i \geq 0$. In practice, there is always, the possibility of no observations from a stratum. It is assumed that the y_{il} $(l = 1, \cdots, n_i)$ are generated from a one-parameter exponential family with pdf

$$f_i(y) = \exp\{\theta_i y - \psi(\theta_i) + h(y)\}.$$

For future reference we record that $E(y_i|\theta_i) = \psi'(\theta_i) = \phi_i$ (say), and $V(y_i|\theta_i) = \psi''(\theta_i)$. The regular likelihood is then given by $\prod_{i=1}^{m}\prod_{l=1}^{n_i} f_i(y_{il})$. In contrast, following Hu and Zidek (2000, 2002), the WL denoted by $WL(\theta) \equiv WL(\theta_1, \cdots, \theta_m)$ is given by $\prod_{i=1}^{m}\prod_{j=1}^{m}\prod_{l=1}^{n_i}[f_i(y_{jl})]^{\omega^*_{ij} n_j^{-1}}$, where the ω_{ij} are the weights. Writing $\bar{y}_j = n_j^{-1}\sum_{l=1}^{n_j} y_{jl}$, $j = 1, \cdots, m$, the above simplifies to

$$WL(\theta) = \exp[\sum_{i=1}^{m}\{\theta_i \sum_{j=1}^{m}\omega^*_{ij}\bar{y}_i - \psi(\theta_i)\sum_{j=1}^{m}\omega^*_{ij} + \sum_{j=1}^{m}\omega^*_{ij}n_j^{-1}\sum_{l=1}^{n_j}h(y_{jl})\}]. \quad (1)$$

It is clear that the WL differs from the usual likelihood in that any inference for the θ_j's based on the latter involves only the direct estimates in contrast to (1) which allows "borrowing strength" from other populations as well.

If the relevance weights, ω^*_{ij}, were specified, noting that $\partial \log WL(\theta)/\partial\theta_i = \sum_{j=1}^{m}\omega^*_{ij}\bar{y}_j - \phi_i\sum_{j=1}^{m}\omega^*_{ij}$ and $\partial^2 \log WL(\theta)/\partial\theta_i^2 = -\psi''(\theta_i)\sum_{j=1}^{m}\omega^*_{ij} < 0$ for all $i = 1, \ldots, m$, the MWLE's of the population means ϕ_i would be given by

$$\hat{\phi}_{iWL} = \sum_{j=1}^{m}\omega_{ij}\bar{y}_j, \quad (2)$$

where $\omega_{ij} = \omega^*_{ij}/\sum_{j=1}^{m}\omega^*_{ij}$. However, in practice, the weights ω^*_{ij} (and hence ω_{ij}) will not usually be easy to specify, and therefore may need to be estimated from the data. This is the subject of the next few sections where the estimators of the ω_{ij} are obtained under increasingly general model assumptions. Also, in this section the QVF structure of the exponential family distribution was not needed. However, this will become necessary in the subsequent sections once we start estimating the ω_{ij}.

3 The Exchangeable Model.

In this section we assume that $\omega_{ii} \equiv \omega,\ i = 1, \ldots, m$ while $\omega_{ij} = \omega^*, j \neq i = 1, \ldots, m$. This assumption embraces the belief that the samples from populations, different from i, are interchangeable with respect to the information they contain about i itself. It is precisely this assumption that leads to the James-Stein estimator in the normal case once the weight has been appropriately estimated using all the data.

Under this assumption, the WLE's of ϕ_i reduce to $\hat{\phi}_{iWL} = \omega y_i + \omega^*\sum_{j\neq i}\bar{y}_j$. Since we require $\sum_{j=1}^{n_i}\omega_{ij} = 1$ for each i, we must have $\omega + (m-1)\omega^* = 1$. Thus,

$$\hat{\phi}_{iWL} = \omega\bar{y}_i + (1-\omega)(m-1)^{-1}\sum_{j\neq i}\bar{y}_j. \quad (3)$$

To find the optimal ω, we minimize $g(\omega) = \sum_{i=1}^{m} E(\hat{\phi}_{iWL} - \phi_i)^2$ with respect to ω. Writing $u_i = V(\bar{y}_i) = \psi''(\theta_i)n_i^{-1}$, $\bar{u} = m^{-1}\sum_{i=1}^{m} u_i$ and $s_\phi^2 = (m-1)^{-1}\sum_{i=1}^{m}(\phi_i - \bar{\phi})^2$, after simplification

$$g(\omega) = m[\omega^2\bar{u} + (1-\omega)^2(m-1)^{-1}\bar{u} + m(1-\omega)^2(m-1)^{-1}s_\phi^2]. \tag{4}$$

Hence, $g(\omega)$ is minimized at $\omega = \omega_{opt}$, where

$$\omega_{opt} = (\bar{u} + ms_\phi^2)[m(\bar{u} + s_\phi^2)]^{-1}. \tag{5}$$

With the substitution of (5) in (3), we get after some simplification,

$$\hat{\phi}_{iWL}^{opt} = (1 - \beta_{opt})\bar{y}_i + \beta_{opt}\bar{y}, \tag{6}$$

where $\bar{y} = m^{-1}\sum_{i=1}^{m}\bar{y}_i$ and $\beta_{opt} = \bar{u}(\bar{u} + s_\phi^2)^{-1}$.

Once again, $\bar{u}$ and s_ϕ^2 are unknown and need to be estimated. To this end, let $s_y^2 = (m-1)^{-1}\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2$ and after simplification we get

$$E(s_y^2) = \bar{u} + s_\phi^2. \tag{7}$$

So far, we have not used the QVF assumption, that is $V(y_{jl}) = \nu_0 + \nu_1\phi_j + \nu_2\phi_j^2 = Q(\phi_j)$ say for all $l = 1, \ldots, n_j, j = 1, \ldots, m$. We now use this assumption to find an estimator of $\bar{u}$. First we note that $E[Q(\bar{y}_j)] = \nu_0 + \nu_1\phi_j + \nu_2[V(\bar{y}_j) + \phi_j^2]$ *i.e.*

$$\begin{aligned} E[Q(\bar{y}_j)] &= Q(\phi_j) + \nu_2 n_j^{-1} Q(\phi_j) \\ &= (n_j + \nu_2)u_j. \end{aligned} \tag{8}$$

Hence, $E[m^{-1}\sum_{j=1}^{m}(n_j + \nu_2)^{-1}Q(\bar{y}_j)] = \bar{u}$. Let

$$\hat{\bar{u}} = m^{-1}\sum_{j=1}^{m}(n_j + \nu_2)^{-1}Q(\bar{y}_j). \tag{9}$$

Then we estimate β_{opt} by $\tilde{\beta}$ defined by

$$\tilde{\beta} = \min\{\frac{\hat{\bar{u}} + m^{-1}}{s_y^2 + m^{-1}}, 1\}. \tag{10}$$

Now ϕ_i is estimated by

$$\tilde{\phi}_{iWL} = (1 - \tilde{\beta})\bar{y}_i + \tilde{\beta}\bar{y},\ i = 1, \ldots, m. \tag{11}$$

Remark 1. We introduce the coefficient m^{-1} to avoid zero both in the numerator and denominator in the discrete case.

Remark 2. Consider the special case when samples of equal size n are drawn independently from the $N(\theta_i, \sigma^2)$ distributions $i = 1, \ldots, m$ where $\sigma^2 > 0$ is known. Then $\nu_0 = \sigma^2$ and $\nu_1 = \nu_2 = 0$. Accordingly, $\hat{\bar{u}} = (nm)^{-1} m\sigma^2 = \sigma^2 n^{-1}$. Now dropping the component m^{-1} from both the numerator and denominator of (10) one gets the estimator

$$\tilde{\phi}^*_{iWL} = (1 - \frac{\sigma^2/n}{s_y^2})\bar{y}_i + \frac{\sigma^2/n}{s_y^2}\bar{y},\ i = 1, \ldots, m, \tag{12}$$

which is Lindley's modification of the James-Stein estimator.

Next we provide an asymptotic mean squared error expansion of the vector $(\tilde{\phi}_{1WL}, \ldots, \tilde{\phi}_{mWL})$ of estimators of $(\phi_{1WL}, \ldots, \phi_{mWL})$. This is the conditional mean squared error in the terminology of Rivest and Belmonte (2000) since the ϕ_i's are held fixed. Specifically, the following result is proved:

Theorem 1 *Assume that lim inf*$_{m\to\infty}(\bar{u} + s_\phi^2) = K > 0$. *Then*

$$m^{-1}\sum_{i=1}^{m} E(\tilde{\phi}_{iWL} - \phi_i)^2 = \frac{\bar{u}s_\phi^2}{\bar{u} + s_\phi^2} + O(m^{-1})$$

as $m \to \infty$.

Remark 3. As shown in the Appendix, $g(\omega_{opt})/m\sum_{i=1}^{m} E(\hat{\phi}_{iWL} - \phi)^2 = \frac{\bar{u}s_\phi^2}{\bar{u}+s_\phi^2} + \frac{\bar{u}^2}{m(\bar{u}+s_\phi^2)}$. Hence, the uncertainty due to estimation of β_{opt} is not reflected in the principal term in the asymptotic expansion of the mean squared error. The difference comes in the coefficient of $O(m^{-1})$.

The proof of the theorem is technical, and is deferred to the Appendix. Much of the proof depends on a special case of the following general result which may be of independent interest.

Proposition 1. Let $X_1, \cdots, X_m$ be mutually independent random variables. Consider a U-statistic with kernel h and degree $k(\leq m)$, that is $U_m = (k!(m-k)!/m!)\sum_{1\leq i_1<i_2\cdots<i_k\leq m} h(X_{i_1}, \cdots, X_{i_k})$. If $\sup_{m\geq 1}\max_{1\leq i_1<i_2\cdots<i_k\leq m} E|h(X_{i_1}, \cdots, X_{i_k})|^r < \infty$, for some $r \geq 2$, then $E|U_m - E(U_m)|^r = O(m^{-r/2})$ as $m \to \infty$.

Proof. Following 5.1.6 (page 180) of Serfling (1980), we express

$$U_m - E(U_m) = (m!)^{-1}\sum_{P} \tilde{W}(X_{i_1}, \cdots, X_{i_m}),$$

where each $\tilde{W}(X_{i_1}, \cdots, X_{i_m})$ is the average of $q = [m/k]$, independent random variables with zero means. Here q denotes the largest integer less than or equal to m/k. The summation extends over all permutations $(i_1, \cdots, i_m 0$ of the first m positive integers. Then by the c_r inequality,

$$\begin{aligned} E|U_m - E(U_m)|^r &\leq (m!)^{-r}(m!)^{r-1} \sum_P E|\tilde{W}(X_{i_1}, \cdots, X_{i_m})|^r \\ &\leq (m!)^{-r}(m!)^{r-1} m! O(q^{-r}) = O(m^{-r}), \end{aligned}$$

where the penultimate step, due to the moment assumption, follows from a standard result for the means.

Remark 4. For the iid case, the result is proved in Serfling (1980) and Sen and Ghosh (1981).

In the next section, the global exchangeability assumed in this section is replaced by local exchangeability.

4 Estimation for a Locally Exchangeable Model.

Suppose now $\omega_{ii} \equiv \omega_i, \omega_{ij} = \omega_{i*},\ j \neq i = 1, \ldots, m$. That is, suppose we drop the assumption of global exchangeability but retain that assumption within each local area. Then $\hat{\phi}_{iWL} = \omega_i \bar{y}_i + \omega_{i*} \sum_{j \neq i} \bar{y}_j$. Since we require $1 = \omega_i + (m-1)\omega_{i*}$, we get

$$\hat{\phi}_{iWL} = \omega_i \bar{y}_i + (1 - \omega_i)(m-1)^{-1} \sum_{j \neq i} \bar{y}_j. \tag{13}$$

To find optimal ω_i's, we now minimize $g(\omega_i) = E(\hat{\phi}_{iWL} - \phi_i)^2$ with respect to ω_i for each $i = 1, \ldots, m$. On simplification,

$$g(\omega_i) = \omega_i^2 u_i + (1 - \omega_i)^2 (m-1)^{-2}(m\bar{u} - u_i) + m^2 (m-1)^{-2}(1 - \omega_i)^2 (\phi_i - \bar{\phi})^2. \tag{14}$$

Noting that $g'(\omega_i) = 2\omega_i u_i - 2(1 - \omega_i)(m-1)^{-2}(m\bar{u} - u_i) - 2m^2(m-1)^{-2}(1 - \omega_i)(\phi_i - \bar{\phi})^2$ and $g''(\omega_i) > 0$, the optimal ω_i is found by solving $g'(\omega_i) = 0$. This gives

$$\omega_{i,opt} = \frac{m\bar{u} - u_i + m^2(\phi_i - \bar{\phi})^2}{(m-1)^2 u_i + m\bar{u} - u_i + m^2(\phi_i - \bar{\phi})^2}. \tag{15}$$

Accordingly, after some algebra,

$$\hat{\phi}_{iWL}^{opt} = (1 - \beta_{i,opt})\bar{y}_i + \beta_{i,opt}\bar{y}, \tag{16}$$

where

$$\beta_{i,opt} = \frac{(m-1)u_i}{(m-2)u_i + \bar{u} + m(\phi_i - \bar{\phi})^2}. \tag{17}$$

Remark 5. It can happen that $\beta_{i,opt} > 1$ when $u_i - \bar{u} > m(\phi_i - \bar{\phi})^2$. Thus, the expression given in (17) is not necessarily a convex combination of $\bar{y}_i$ and $\bar{y}$.

However, in practice the β_i's are unknown and need to be estimated. To this end we see that

$$\begin{aligned} E(\bar{y}_i - \bar{y})^2 &= V(\bar{y}_i - \bar{y}) + (\phi_i - \bar{\phi})^2 \\ &= m^{-1}[(m-2)u_i + \bar{u} + m(\phi_i - \bar{\phi})^2] \end{aligned} \tag{18}$$

Also $E[Q(\bar{y}_i)] = \nu_0 + \nu_1\phi_i + \nu_2(u_i + \phi_i^2) = (n_i + \nu_2)u_i$. Thus $E[Q(\bar{y}_i)]/(n_i + \nu_2) = u_i$. Hence, $\beta_{i,opt}$ is estimated by

$$\hat{\beta}_i = \min\left[\frac{m-1}{m}\frac{Q(\bar{y}_i)(n_i+\nu_2)^{-1} + m^{-1}}{(\bar{y}_i - \bar{y})^2 + m^{-1}}, 1\right]. \tag{19}$$

Correspondingly,

$$\hat{\phi}_{iWL} = (1 - \hat{\beta}_i)\bar{y}_i + \hat{\beta}_i\bar{y}. \tag{20}$$

5 Estimation Without Any Exchangeability Assumptions.

The most general situation allows the weights to be unit specific within each local area. In particular, if ϕ_i is estimated by $\sum_{j=1}^m \omega_{ij}\bar{y}_j$ then the optimal ω_{ij} are found by minimizing $E(\sum_{j=1}^m \omega_{ij}\bar{y}_j - \phi_i)^2$ with respect to ω_{ij} subject to the condition $\sum_{j=1}^m \omega_{ij} = 1$.

To this end, write $\Omega_i = (\omega_{i1}, \ldots, \omega_{im})^T$ and let $g(\Omega_i) = E(\sum_{j'=1}^m \omega_{ij'}\bar{y}_{j'} - \phi_i)^2 - 2\lambda_i(\sum_{j'=1}^m \omega_{ij'} - 1)$, where the λ_i's are the Lagrangian multipliers. Then on simplification,

$$g(\Omega_i) = \sum_{j'=1}^m \omega_{ij'}^2 u_{j'} + (\sum_{j'=1}^m \omega_{ij'}\phi_{j'} - \phi_i)^2 - 2\lambda_i(\sum_{j'=1}^m \omega_{ij'} - 1). \tag{21}$$

Thus

$$\frac{\partial g}{\partial \omega_{ij}} = 2\omega_{ij}u_j + 2(\sum_{j'=1}^m \omega_{ij'}\phi_{j'} - \phi_i)\phi_j - 2\lambda_i; \tag{22}$$

$$\frac{\partial^2 g}{\partial \omega_{ij}^2} = 2u_j + 2\phi_j^2 > 0; \tag{23}$$

$$\frac{\partial^2 g}{\partial \omega_{ij}\omega_{kl}} = 0,\ j \neq l. \tag{24}$$

Hence the Hessian matrix for g is positive definite and the optimal ω_{ij} may be found by solving $\partial g/\partial \omega_{ij} = 0,\ i,j = 1,\ldots,m$. In matrix notations these equations can be written as

$$\mathbf{M}\Omega_i = \lambda_i \mathbf{1}_m + \phi_i \Phi, \tag{25}$$

where $\mathbf{M} = \mathbf{D} + \Phi\Phi^T,\ \mathbf{D} = Diag(u_1,\ldots,u_m),\ \Phi = (\phi_1,\ldots,\phi_m)^T$ and $\mathbf{1}_m$ is the m-component column vector with each element equal to 1. Thus the optimal Ω_i is given by

$$\Omega_i^{opt} = \mathbf{M}^{-1}(\lambda_i \mathbf{1}_m + \phi_i \Phi). \tag{26}$$

Since $1 = \mathbf{1}_m^T \Omega_{i,opt} = \lambda_i \mathbf{1}_m^T \mathbf{M}^{-1}\mathbf{1}_m + \phi_i \mathbf{1}_m^T \mathbf{M}^{-1}\Phi$, one gets

$$\lambda_i = \frac{1 - \phi_i \mathbf{1}_m^T \mathbf{M}^{-1}\Phi}{\mathbf{1}_m^T \mathbf{M}^{-1}\mathbf{1}_m}. \tag{27}$$

Now from (26) and (27),

$$\Omega_{i,opt} = \frac{1 - \phi_i \mathbf{1}_m^T \mathbf{M}^{-1}\Phi}{\mathbf{1}_m^T \mathbf{M}^{-1}\mathbf{1}_m}(\mathbf{M}^{-1}\mathbf{1}_m) + \phi_i \mathbf{M}^{-1}\Phi. \tag{28}$$

The expression for $\mathbf{M}^{-1}$ can be somewhat simplified by the standard matrix inversion formula (Rao, 1973. p33, Exercise 2.8) as

$$\mathbf{M}^{-1} = \mathbf{D}^{-1} - \frac{\mathbf{D}^{-1}\Phi\Phi^T\mathbf{D}^{-1}}{1 + \Phi^T\mathbf{D}^{-1}\Phi}.$$

Now since $E(\bar{y}_i) = \phi_j$ and $E[(n_j + \nu_2)^{-1}Q(\bar{y}_j)] = u_j$, where $Q(\bar{y}_j) = \nu_0 + \nu_1\bar{y}_j + \nu_2\bar{y}_j^2$, writing $\hat{u}_j = (n_j + \nu_2)Q(\bar{y}_j)$, one gets

$$\hat{\Omega}_{i,opt} = \frac{1 - \bar{y}_i \mathbf{1}_m^T \hat{\mathbf{M}}^{-1}\mathbf{z}}{\mathbf{1}_m^T \hat{\mathbf{M}}^{-1}\mathbf{1}_m}(\hat{\mathbf{M}}^{-1}\mathbf{1}_m) + \bar{y}_i \hat{\mathbf{M}}^{-\mathbf{1}}\mathbf{z}, \tag{29}$$

where $\mathbf{z} = (\bar{y}_1,\ldots,\bar{y}_m)^T$ and $\hat{\mathbf{M}} = (\hat{D} + \mathbf{z}\mathbf{z}^T)^{-1}$, where $\hat{\mathbf{D}} = Diag(\hat{u}_1,\ldots,\hat{u}_m)$.

Remark 6. Note that the coefficients $\Omega_{i,opt}$ can themselves be negative which suggests that the present procedure is general enough to accommodate negative relevance (if any). The need to allow negative weights stems from the work of van Eeden and Zidek (2000, 2002) who exhibit classical estimators that successfully trade bias for precision and are MWLE's provided that such weights are allowed.

6 Simulation Study.

In this section,we do a simulation study involving a number of populations to compare the performance of the three types of weighted likelihood estimators derived in this paper. In particular, we compare

them against the respective maximum likelihood estimators for the individual populations. The latter, which are based on just the single sample drawn from that population, are unbiased and have many optimality properties among estimators based on just that sample. However, they are pitted against estimators that combine information from all the samples.

The response variable, a random count, is assumed to have a negative binomial distribution in each population. We chose that distribution for its convenience and, in part, to get away from the well-studied normal distribution. Moreover, unlike the Poisson distribution, this one can have differing means as well as variances and it can have quite heavy tails.

With potential applications such as small area estimation and disease mapping in mind, we chose a seemingly realistic number of 100 populations for our study. A total of 8 cases were considered as described in Table 1.

Except for Case 1 the where all population means were different from one population to another, the populations were clustered into somewhat homogeneous groups. There were 10 such groups in Cases 2-5, 3 in Case 6 (of sizes 30, 40 and 30) and just 2 in Case 7 (of sizes 50 each) and 1 in Case 8. Thus we were able to study the effect of the sort of clustering one might expect to see in a real application. Of primary interest was the effect of unequal population means. Would our estimators provide any gain in precise in this case?

Of secondary interest was the effect of unequal variances. By introducing heteorscedasticity, we were able to observe the effects of varying data-quality. Incorporating poor measurements (counts) with a high variance seems unlikely to produce much of a gain in precision as measured by normalized MSE, our primary performance criterion, when some of the data come from high variance populations.

In the same vein, we wondered about the effect of varying sample sizes so in some cases, we allowed that to happen as well. In the following figures we present our results based on running 5000 replicate sampling experiments for each of the 8 cases in Table 1. However, because of similarities noted below, figures and not provided for all cases.

As a performance criterion, we used a normalized mean squared error (MSE) criterion to insure the inter-comparability of precision between populations. Furthermore, it meant that we could meaningfully aggregate those MSEs across populations.

Thus, for each population $i = 1, c \ldots, 100$, we assessed the performance of any given estimator

Table 1: The Eight Cases Treated in Our Simulation Study. In each case, samples of the specified size were randomly drawn from a population of counts having a negative binomial population distribution with the given means and variances.

Case	# Clusters	Sample Sizes	Population Means	Population Variances
1	100	20	All different from a skewed distribution	20
2	10	20	10	110,60,14.4,35,30 30,8.7,3.4,22.5,3.5,20
3	10	20	4, 5.4,6.4,7.2,7.8 8.4,8.8,9.3,9.7,1	20
4	10	10,15,20,25,30 for each of two populations in each cluster.	10	110,60,14.4,35,30 30,8.7,22.5,3.5,20
5	10	10,15,20,25,3 for each of two populations in each cluster.	4,5.4,6.4,7.2,7.8 8.4,8.8,9.3,9.7,10	10
6	3	20	4,12.4,19.6	20
7	2	20	4,19.6	20
8	100	20	10	All different from a skewed distribution

$\bar{\mu}_i = WLE1, WLE2, WLE3, MLE$ of that population's mean by

$$E\left[\frac{(\hat{\mu}_i - \mu_i)^2}{\sigma_i^2/n_i}\right]. \tag{30}$$

where σ_i^2 and n_i denote the population and sample size respectively. This normalized MSE could readily be estimated using 5000 replicated sampling experiments for that population for any given case in Table 1.

Note that for $\hat{\mu} = MLE$, the ratio in Equation (30) is 1. However, we also estimated and plotted our estimates of that ratio alongside those for the other estimators. This provides some indication of the range of natural variability across populations induced by the sampling alone apart from that resulting from differences in performance.

For Case 1, Figure 1 depicts the results from the 5000 replicate sampling experiments. Here we see notched boxplots of the estimates of the normalized MSE in Equation (30), for each of the estimators under consideration. There, WLE1 represents the weighted likelihood based on the fully "exchangeable" weights model. Its estimated normalized MSE tends to be smaller than those for the other estimators, say WLE2, WLE3 and MLE, where WLE2 and WLE3 represent the weighted likelihoods for the locally and not exchangeable weights models, respectively.

However, notice that 9 of the 100 are "outliers" in that their normalized MSE estimates tend to be rather large. These correspond to the populations with very small true means obtained from their left-skewed distribution. WLE1, which tends to shrink toward the center of that distribution does not cope very well with these outlying populations.

In contrast, WLE3 tracks these populations well, as well in fact, as the MLE; their difference is imperceptible. In general WLE3 seems very robust but it seems to offer at best only small gains over the MLE.

WLE2's performance differs from both of its cousins. It does very well for the 25% of the populations corresponding to the lower whisker in its notched boxplot. However, overall, it does not do as well as any of the other 3 estimators for Case 1.

A different view of the performance of these estimators in Case 1 can be seen in Figure 2. The relationship between that performance and the population means can be seen more clearly. Note in particular, that only WLE1 proves to be susceptible to extreme bias for the populations with the smallest means.

Table 2: Simulation Experiment for Case 6 of Table 1. Each entry divided by 1000 represents the average over all replications and populations.

Type	MSE *1000	Variance*1000	Bias Squared*1000
Fully Exchangeable	972	947	25
Partially Exchangeable	830	805	25
Local	997	997	0
MLE	998	997	0

Notice also that WLE2 performs extremely well near the mean of the population means (14.33). We have no explanation for this phenomenon.

We omit Cases 2 and 3 from discussion since their notched boxplots are very similar to those for Cases 4 and 5 respectively to which we now turn.

In Cases 2 and 4, we see WLE1 at its best; the population means are identical. The variations in sample sizes as well as variances seen in Case 4 seem to have no deleterious effect on performance seen in Case 2 (but not depicted here). WLE1 proves robust against such variation.

Notice that both WLE1 and WLE2 do much better in these cases than WLE3 and MLE. Moreover, for all the estimators we see very little variation across populations in the performance assessment estimates. WLE3 does perform somewhat better than the MLE.

As an aside, we have observed empirically that the weights for population j in WLE2 (that must sum to 1) when we are estimating the mean of population i are approximately proportional to:

$$\frac{(k-1)}{k}\frac{n_j}{\sigma_j^2} + \delta_{ij}\frac{1}{k}\sum\frac{n'_j}{\sigma^2_{j'}},$$

where δ_{ij} represents Dirac's delta function. However, we do not a rigorous justification for this observation.

In Cases 3 and 5, where the population means vary, estimator performance mirrors that in Case 1, although there are only 10 rather than the 100 mean-clusters seen in Case 1. Curiously, the boxes in the plot tend to be somewhat more elongated in the latter case than in the former.

Cases 6 and 7 are somewhat similar in that they involve a small number of clusters with different means. Thus we illustrate our results for the former only.

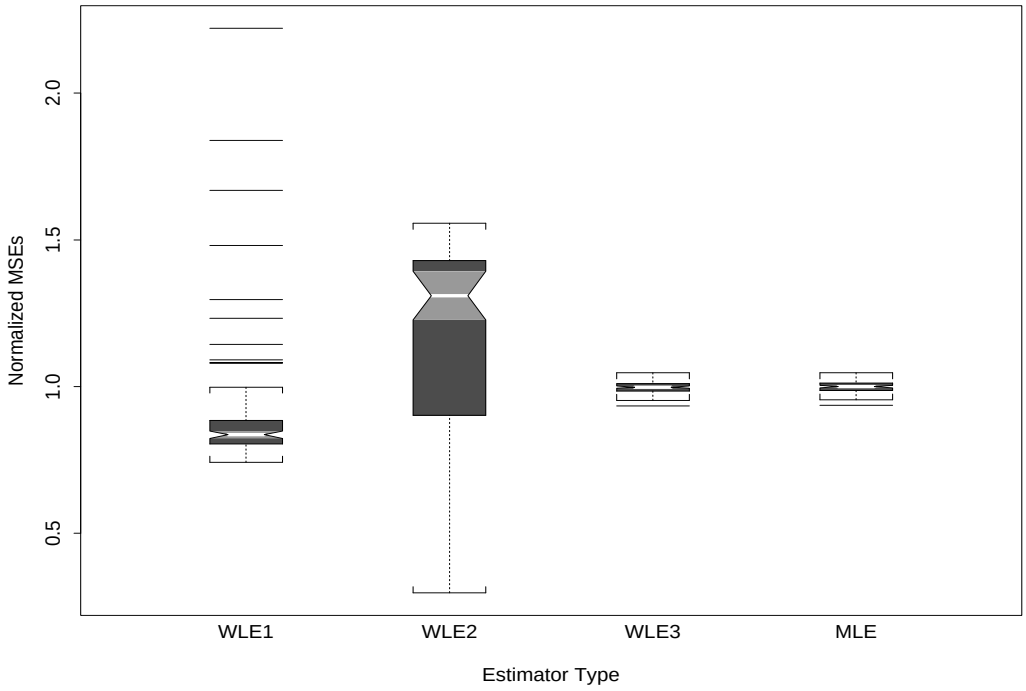

Figure 1: **Case 1. Notched Boxplots of the Estimated Normalized MSE's for All 100 Populations and Each of the Estimators, Based on 5000 Simulated Sampling Experiments.**

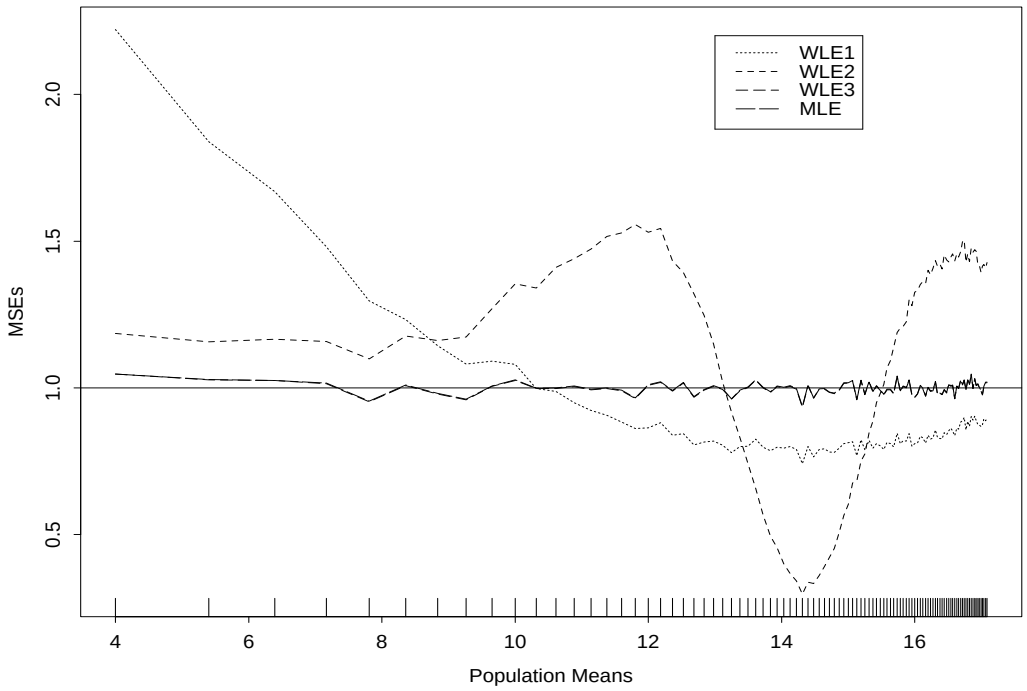

Figure 2: Case 1. Lineplot of the Estimated Normalized MSE's for All 100 Populations and Each of the Estimators, Based on 5000 Simulated Sampling Experiments. Here the Estimated MSE's Are Plotted Against Population Means that are Depicted in the Rugplot On the Horizontal Axis.

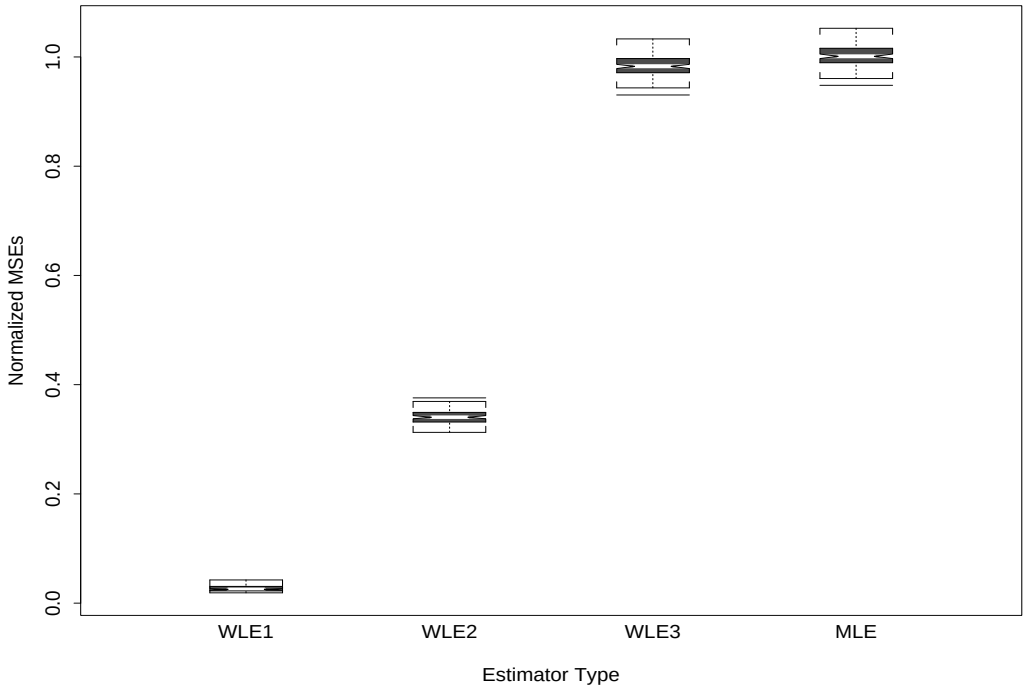

Figure 3: Case 4. Notched Boxplots of the Estimated Normalized MSE's for All 100 Populations and Each of the Estimators, Based on 5000 Simulated Sampling Experiments.

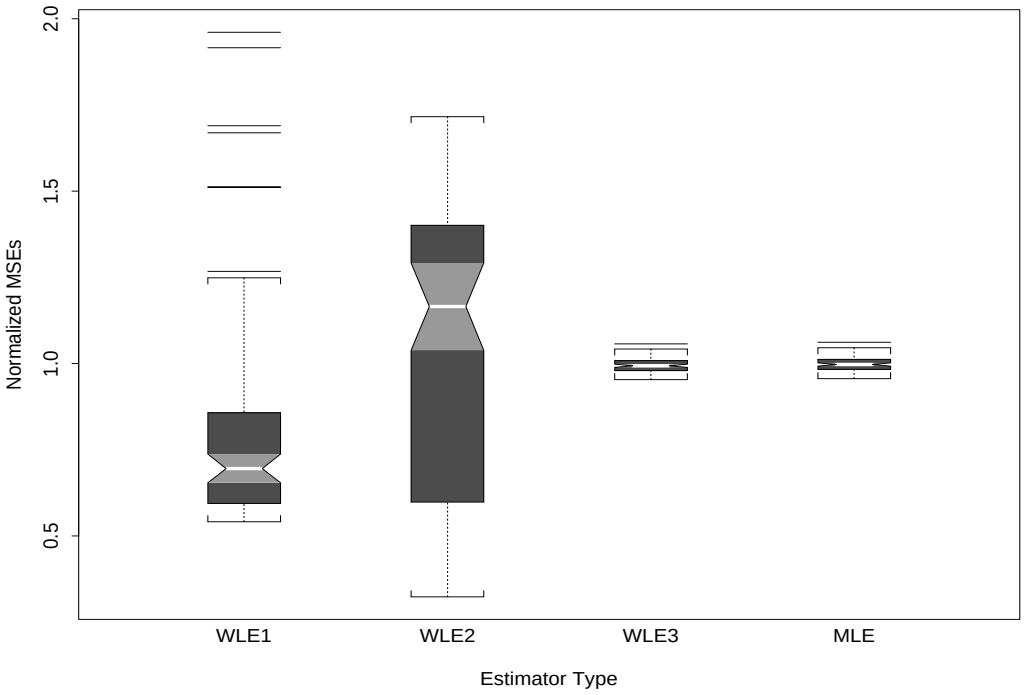

Figure 4: Case 5. Notched Boxplots of the Estimated Normalized MSE's for All 100 Populations and Each of the Estimators, Based on 5000 Simulated Sampling Experiments.

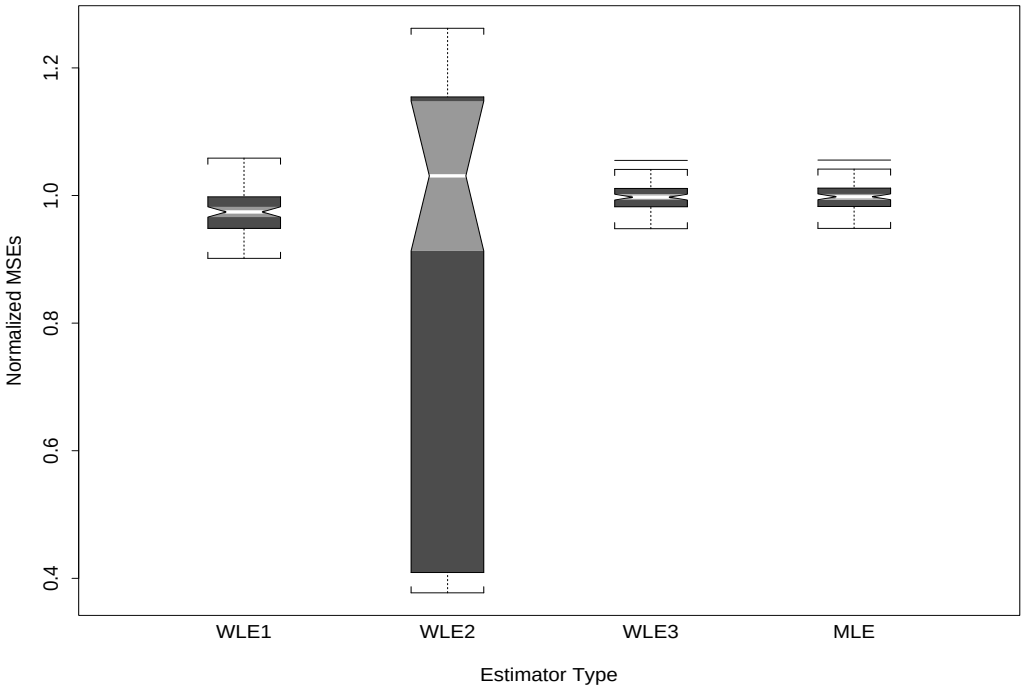

Figure 5: Case 6. Notched Boxplots of the Estimated Normalized MSE's for All 100 Populations and Each of the Estimators, Based on 5000 Simulated Sampling Experiments.

Center	n_i	$\hat{p}_i$	WL1	WL2	WL3	se($\hat{p}_i$)	se(WL1)	se(WL2)	se(WL3)
1	79	.316	.246	.253	.300	5.44	1.54	2.20	5.39
2	76	.329	.246	.256	.312	5.36	1.52	2.21	5.33
3	75	.240	.245	.244	.229	4.90	1.45	1.59	4.72
4	91	.209	.244	.241	.200	4.38	1.42	1.52	4.17
5	91	.209	.244	.241	.200	4.05	1.42	1.53	3.84
6	106	.198	.244	.240	.190	4.02	1.43	1.60	3.80
7	134	.269	.245	.246	.256	3.67	1.53	1.44	3.59
8	73	.137	.243	.225	.132	3.67	1.41	1.79	3.45
9	152	.263	.245	.246	.250	3.56	1.44	1.47	3.49
10	101	.277	.245	.248	.263	4.63	1.47	1.66	4.56

Table 3: This table presents the fraction ($\hat{p}$) of organ transplant complications that resulted in organ rejection during the period October 1, 1987 to December 31, 1989 at 121 US transplant centers. These fractions constitute the maximum likelihood estimators for the true probability of such complications. Presented as well, are the results from three weighted likelihood estimators along with estimators of their standard errors based on parametric bootstrap samples. Standard errors are multiplied by 100.

We see that in this case that none of the weighted likelihood estimators is able to gain substantially over the MLE overall. However the results for WLE2 vary widely and in nearly 50% of the cases it does seem to achieve a normalized MSE below the remaining cases. Moreover, this the only case in our study where we found that the overall, non-normalized MSE for WLE2 was actually lower than everyone of the remaining estimators. This performance can be seen in Table 2. This suggests that WLE2 may have some value in selected cases.

7 An Illustrative Application.

This section presents an illustrative example and demonstrates the weighted likelihood methodology. Our data come from the United Network for Organ Sharing (UNOS) Public-Use database. That database records the results of 3,688 transplant operations at 131 transplant centers during the period

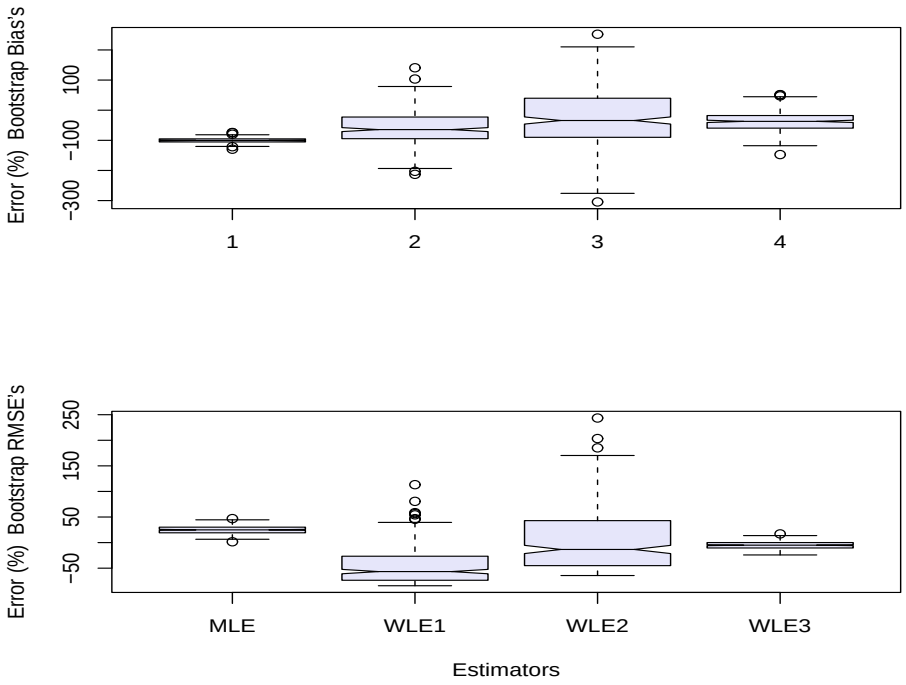

Figure 6: **Notched boxplots portray for Center 10, of the relative bias (top panel) and RMSE (bottom panel) based on 300 bootstrap estimates of the "true" fraction of organ transplant rejections for all estimators considered in this paper.**

October 1, 1987 to December 31, 1989. To avoid complexity that diverts more than it informs, we confine our demonstration to the 10 centers with the largest number of heart transplants. Outcomes are classified as '1' or '0' according as the patient developed short term problems leading to rejection of organ death or not. (Dey, Gelfand, Swartz and Vlachos (1998) present those data.) Here our objective is to estimate the proportion of patients who develop short term problems leading to rejection-of-organ-death for all the 10 centers. The estimators are presented in Table 3 along with their bootstrap standard error estimates.

To estimate those standard errors, simultaneous samples from all 10 centers were generated by the parametric bootstrap to yield 10 dimensional rejection proportion vectors, $\hat{\mathbf{p}}^{(k)},\ k = 1, \ldots, 300$. (We found 300 to be the largest feasible number given available computer memory.) In other words, rejection counts were independently generated for Center $i = 1, \ldots, 10$ from the $binomial(n_i, \hat{p}_i)$ distribution where n_i and $\hat{p}_i$ obtain from Table 3. Estimates were then computed for each of these 300

vectors and the sample standard deviations of these estimates yields the estimated standard errors.

Are these estimators accurate? To help answer this question, we used the "two-deep" parametric bootstrap. More precisely, for each k, we generated a further 300 vectors of fractions, $\hat{p}^{kh}$, $h = 1, \ldots, 300$ from binomial distributions, $binomial(n_i, \hat{p}_i^k)$ for all $i = 1, \ldots, 10$ and $k = 1, \ldots, 300$. Then $\sum_{h=1}^{300} \hat{p}^{kh}/300 - \hat{p}^k$ estimates the bias for all k. Similarly, bootstrap standard error estimates obtain from taking the standard deviation of the $\{\hat{p}_i^{kh}\}$. The distribution over k of these quantities provide some idea of how well the bootstrap estimates the corresponding quantities based on a one-deep analysis as indicated below.

However, bootstrap estimates of both estimator bias and root mean square errors (RMSE) of estimation are of interest. For bias, comparative performances of the estimators can be found by computing for all k, $100 * (biasboot_i^k - bias_i)/bias_i$ where $biasboot_i^k = \sum_{h=1}^{300}(\hat{\theta}_i^{kh} - \hat{p}_i^k)/300$ and $bias_i = \sum_{k=1}^{300}(\hat{\theta}_i^k - \hat{p}_i)]/300$. These indices expressed as relative error percentages, indicate how well the bootstrap estimator of bias works for the various estimators, $\hat{\theta}$, calculated with one-deep and two-deep bootstrap samples. Similar ratios can be computed for the RMSE. For Center 10, the results for both bias and RMSE appear in Figure 6.

Remark 7. The tables and figures, including a number not included here for brevity, show WLE1 and WLE2 on the one hand and MLE and WLE3 on the other, to be very similar in performance. For simplicity, we will refer to them as Pairs 1 and 2, respectively.

Remark 8. For Center 10 (that was arbitrarily chosen to be the example), Pair 1 varies little from one bootstrap sample to another unlike Pair 2 that varies a lot and produces extreme values in some cases. Figure 2 echoes these findings. Figure 1 suggests that Pair 1 is stochastically smaller than Pair 2, but Figure 3, a "group portrait" reveals that finding to be premature. In fact, there is quite a lot of variability in the level of bootstrap fraction estimates over the 10 centers in contrast to their standard errors that are almost identical over centers (see Figure 4).

Remark 9. Of more interest are figures that show the accuracy of bootstrap bias and RMSE estimates. For brevity, only Figure 6 is included here. These figures together show the bootstrap performance indicators for Pair 2 to be substantially more accurate than those for Pair 2. In particular, they indicate that the (rather large) standard errors seen in Table 3 accurately reflect their relatively poor performance. They show the members of Pair 1 performance about equally

well. In fact, Figure 6 suggests that WLE1 outperforms WLE2. More precisely, the error in the bias estimate for WLE2 can be as large as 50% in contrast to WLE1's 30%. Similarly, inaccuracies are plausible for their RMSE's at Center 10. Nevertheless, even allowing for these inaccuracies, the results in Table 3 suggest that both WLE1 and WLE2 will give more precise estimates of the true center fractions than MLE and WLE3.

Remark 10. The greater simplicity (and similar performance) of MLE compared to WLE3 and of WLE1 compared to WLE2, suggest the contest is really between MLE and WLE1. Our empirical results point to WLE1 as the overall winner, therefore.

8 Concluding Remarks

The paper has introduced the weighted likelihood approach for estimation of parameters in natural exponential family of distributions with quadratic variance functions. This class of distributions has been characterized by Morris (1982, 1983), and includes the binomial, Poisson and normal distributions as special cases.

Our simulation study shows that when the means of negative binomial populations are similar, and the weighted likelihood estimator takes advantage of this fact (WLE1), spectacular gains in performance over the MLE are achievable. At the same time, in more non-homogeneous cases WLE1's performance can be similar to that of the MLE and even worse in populations with means quite far away from the center of the mean values. However, even in these cases little seems to be lost from using WLE1's.

In contrast, the weighted likelihood based on local exchangeability can have highly variable performance across a group of negative binomial populations. The performance for some populations can be extremely good, while poor in others. In short, the results seem much less predictable than those for either of it to cousins. Still in one case (6), WLE2 does prove to be better overall than the others.

Finally, the third weighted likelihood estimator considered in our simulation study, WLE3, is highly robust but never seems to achieve much over the MLE. It could be worthy of consideration as a safe alternative to the MLE when small gains matter. For example, small improvements in estimating census under-counts, an application we have not yet studied, could have major implications for society. However, in general, the MLE would be much preferred to WLE3 because of its simplicity.

The illustrative example in Section 7 demonstrates a practical application of the theory presented in this paper. Moreover, it shows that the bootstrap can be used to estimate bias and root mean squared error of estimation and it even allows us to estimate the plausible range of errors in measures of those types of performance. Overall, WLE1 comes out the winner in this application, no doubt because of the similarity of their fractions and counts. However, it will not be the winner in every such application. An analysis like that above will be needed to find an the appropriate choice.

Acknowledgments

Malay Ghosh and Tapabrata Maiti's research was supported by NSF Grant Number SES-9911485. James V. Zidek's research was supported by Natural Sciences and Engineering Research Council of Canada. The paper has benefitted considerably from the constructive comments of the Editor, Associate Editor and two reviewers.

References

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications*. London:Chapman and Hall.

Hu, F. (1994). Relevance Weighted Smoothing and a New Bootstrap Method. Unpublished Ph.D. Dissertation. Department of Statistics. University of British Columbia.

Hu, F. and Zidek, J.V. (2000). The relevance weighted likelihood with applications. In *Empirical Bayes and Likelihood Inference*. Lecture Notes in Statistics. Eds. S.E. Ahmed and N. Reid. Springer Verlag, New York, pp 211-235.

Hu, F. and Zidek, J.V. (2002). The weighted likelihood. *Can J Statist*, 30, 347-371.

Morris, C.N. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics*, 10, 65-80.

Morris, C.N. (1983). Natural exponential families with quadratic variance functions: statistical theory. *Annals of Statistics*, 11, 515-529.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. New York:Wiley.

Rivest, Louis-Paul and Eve, Belmonte (2000). A conditional mean squared error of small area

estimators. *Survey Methodology*, 26, 67-78. Sarkar, S. and Ghosh, M. (1998). Empirical Bayes estimation of local area means for NEF-QVF superpopulations. *Sankhya B*, 60, 464-487.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probability*, V1. University of California Press, pp 197-206.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.

Van Eeden, C, and Zidek, JV (2000). Combining the data from two normal populations to estimate the mean of one when their means diffence is bounded. *J Mult Anal*. To appear. Available from http://hajek.stat.ubc.ca/ jim/pubs.

Van Eeden, C, and Zidek, JV (2002). Combining sample information in estimating ordered normal means. *Sankyha, Series A*, 64, 588-610.

9 Appendix

Proof of Theorem 1.

The following two lemmas are used repeatedly in proving the theorem, Lemma 1 being essentially an immediate application of Proposition 1.

Lemma 1. For every $r \geq 2$, $E|s_y^2 - E(s_y^2)|^r = O(m^{-r/2})$ as $m \to \infty$.

Proof. First we express s_y^2 as a U-statistic, namely $s_y^2 = [2/(m(m-1))] \sum_{1 \leq i_1 < i_2 \leq m} (\bar{y}_{i_1} - \bar{y}_{i_2})^2/2$. Next, by the c_r inequality,

$$E|\bar{y}_{i_1} - \bar{y}_{i_2}|^{2r} \leq E(\bar{y}_{i_1}^{2r} + \bar{y}_{i_2}^{2r}) \leq 2\text{max}_{1 \leq i \leq m} E(\bar{y}_i^{2r}).$$

For the NEF-QVF family of distributions, $\sup_{r \geq 1} E|\bar{y}_i|^r < \infty$. The result follows now immediately from Proposition 1.

Lemma 2. $E(s_y^2 + m^{-1})^{-p} = O(1)$ as $m \to \infty$. for every $p \geq 2$ as $m \to \infty$.

Proof. We write

$$\begin{aligned} E(s_y^2 + m^{-1})^{-p} &= E[(s_y^2 + m^{-1})^{-p} I_{[s_y^2 \geq c]} + (s_y^2 + m^{-1})^{-p} I_{[s_y^2 < c]} \\ &\leq (c + m^{-1})^{-p} + m^p P(s_y^2 < c)]. \end{aligned} \tag{31}$$

Here I is the usual indicator function and $c(>0)$ will be chosen appropriately. Recalling that $K = \liminf_{m\to\infty} E(s_y^2)$, choosing $c = K/2$, by Markov's inequalty and Lemma 1, one gets

$$\begin{aligned} P(s_y^2 < c) &\leq P(s_y^2 - E(s_y^2) < -K/2) \\ &\leq P(|s_y^2 - E(s_y^2)| > K/2) \\ &\leq (K/2)^{-r} E|s_y^2 - E(s_y^2)|^r = O(m^{-r/2}). \end{aligned} \tag{32}$$

Choosing $r \geq 2p$, the lemma follows from (31) and (32).

Back to the proof of Theorem 1, first we observe that

$$\begin{aligned} m^{-1}\sum_{i=1}^{m} E(\tilde{\phi}_{iWL} - \phi_i)^2 &= m^{-1}\sum_{i=1}^{m} E(\hat{\phi}_{iWL}^{opt} - \phi_i + \tilde{\phi}_{iWL} - \hat{\phi}_{iWL}^{opt})^2 \\ &= m^{-1}\sum_{i=1}^{m} E[(\hat{\phi}_{iWL}^{opt} - \phi_i)^2 + (\tilde{\phi}_{iWL} - \hat{\phi}_{iWL}^{opt})^2 \\ &\quad + 2(\hat{\phi}_{iWL}^{opt} - \phi_i)(\tilde{\phi}_{iWL} - \hat{\phi}_{iWL}^{opt})]. \end{aligned} \tag{33}$$

Now after some simplification,

$$\begin{aligned} m^{-1}\sum_{i=1}^{m} E(\hat{\phi}_{iWL}^{opt} - \phi_i)^2 &= m^{-1}\sum_{i=1}^{m} E[(1-\beta_{opt})\bar{y}_i + \beta_{opt}\bar{y} - \phi_i]^2 \\ &= \frac{\bar{u}s_\phi^2}{\bar{u}+s_\phi^2} + m^{-1}\frac{\bar{u}^2}{\bar{u}+s_\phi^2} = O(m^{-1}) \end{aligned} \tag{34}$$

as $m \to \infty$, where the last step follows by Assumption 1 and the fact that $\sup_{i\geq 1} E|\bar{y}_i|^r < \infty$ for every $r > 0$. Next letting $\tilde{\tilde{\beta}} = (\hat{\bar{u}} + m^{-1})/(s_y^2 + m^{-1})$, one gets

$$\begin{aligned} m^{-1}\sum_{i=1}^{m} E(\tilde{\phi}_{iWL} - \hat{\phi}_i^{opt})^2 &= (m-1)m^{-1}E[(\tilde{\beta} - \beta_{opt})^2 s_y^2] \\ &= (m-1)m^{-1}E[(\tilde{\tilde{\beta}} - \beta_{opt} + \tilde{\beta} - \tilde{\tilde{\beta}})^2 s_y^2] \\ &\leq 2(m-1)m^{-1}E[\{(\tilde{\tilde{\beta}} - \beta_{opt})^2 + (\tilde{\beta} - \tilde{\tilde{\beta}})^2.\}s_y^2] \end{aligned} \tag{35}$$

But

$$\begin{aligned} E[(\tilde{\tilde{\beta}} - \beta_{opt})^2 s_y^2] &= E\left[\left(\frac{\hat{\bar{u}} + m^{-1}}{s_y^2 + m^{-1}} - \frac{\bar{u}}{E(s_y^2)}\right)^2 s_y^2\right] \\ &= E\left[\left(\frac{\hat{\bar{u}} + m^{-1}}{s_y^2 + m^{-1}} - \frac{\bar{u}}{s_y^2 + m^{-1}} + \frac{\bar{u}}{s_y^2 + m^{-1}} - \frac{\bar{u}}{E(s_y^2)}\right)^2 s_y^2\right] \\ &\leq 2E\left[\frac{(\hat{\bar{u}} + m^{-1} - u)^2}{(s_y^2 + m^{-1})^2} s_y^2\right] \end{aligned}$$

$$
\begin{aligned}
&+ 2E\left[\bar{u}^2 \frac{(s_y^2 - Es_y^2 + m^{-1})^2}{(s_y^2 + m^{-1})^2 (E(s_y^2))^2} s_y^2\right] \\
\leq\ & 2E\left[\frac{(\hat{\bar{u}} + m^{-1} - u)^2}{s_y^2 + m^{-1}}\right] \\
&+ 2E\left[\bar{u}^2 \frac{(s_y^2 - Es_y^2 + m^{-1})^2}{(s_y^2 + m^{-1})(E(s_y^2))^2}\right]. \qquad (36)
\end{aligned}
$$

Next by the Cauchy-Schwarz inequality,

$$
E\left[\frac{(\hat{\bar{u}} + m^{-1} - \bar{u})^2}{s_y^2 + m^{-1}}\right] \leq E^{1/2}(\hat{\bar{u}} + m^{-1} - \bar{u})^4 E^{1/2}(s_y^2 + m^{-1})^{-2}. \qquad (37)
$$

By the c_r-inequality for the NEF-QVF family,

$$
\begin{aligned}
E(\hat{\bar{u}} + m^{-1} - \bar{u})^4 &\leq 2^3 E[(\hat{\bar{u}} - \bar{u})^4 + m^{-4}] \\
&= 8[O(m^{-2}) + m^{-4}]. \qquad (38)
\end{aligned}
$$

Also, by Lemma 2, $E(s_y^2 + m^{-1})^{-2} = O(1)$.

Combining (37) and (38) we obtain

$$
E\left[\frac{(\hat{\bar{u}} + m^{-1} - \bar{u})^2}{s_y^2 + m^{-1}}\right] = O(1). \qquad (39)
$$

Similarly, by Lemmas 1 and 2,

$$
E\left[\frac{(s_y^2 - Es_y^2 + m^{-1})^2}{s_y^2 + m^{-1}}\right] \leq E^{1/2}(s_y^2 - Es_y^2 + m^{-1})^4 E^{1/2}(s_y^2 + m^{-1})^{-2} = O(m^{-1}). \qquad (40)
$$

Again,

$$
\begin{aligned}
E[(\tilde{\tilde{\beta}} - \tilde{\beta})^2 s_y^2] &= E[(1 - \tilde{\tilde{\beta}})^2 s_y^2 I\{\hat{\bar{u}} > s_y^2\})] \\
&\leq E^{1/2}[(1 - \tilde{\tilde{\beta}})^4 s_y^4] P^{1/2}(\hat{\bar{u}} > s_y^2). \qquad (41)
\end{aligned}
$$

Now applying the c_r-inequality once again we obtain

$$
\begin{aligned}
E[(1 - \tilde{\tilde{\beta}})^4 s_y^4] &= E[(1 - \beta_{opt} + \beta_{opt} - \tilde{\tilde{\beta}})^4 s_y^4] \\
&\leq 8E[\{(1 - \beta_{opt})^4 + (\tilde{\tilde{\beta}} - \beta_{opt})^4\} s_y^4] \\
&\leq 8[Es_y^4 + E^{1/2}(\tilde{\tilde{\beta}} - \beta_{opt})^8 E^{1/2}(s_y^8)]. \qquad (42)
\end{aligned}
$$

Similar to (36),

$$
E[(\tilde{\tilde{\beta}} - \beta_{opt})^4 s_y^4] \leq 8E[\frac{\hat{\bar{u}} + m^{-1} - \bar{u})^4}{(s_y^2 + m^{-1})^{-2}} + 8E[\bar{u}^2 \frac{(s_y^2 - E(s_y^2) + m^{-1})^4}{(s_y^2 + m^{-1})^2 (E(s_y^2))^4}] = O(m^{-2}),
$$

by an application of Lemmas 1 and 2.

Also,

$$\begin{aligned} P(\hat{\bar{u}} > s_y^2) &= P[\hat{\bar{u}} - s_y^2 - E(\hat{\bar{u}} - s_y^2) > -E(\hat{\bar{u}} - s_y^2)] \\ &= P[\hat{\bar{u}} - s_y^2 - E(\hat{\bar{u}} - s_y^2) > s_\phi^4] \\ &\leq s_\phi^{-4r} E[\hat{\bar{u}} - s_y^2 - E(\hat{\bar{u}} - s_y^2)]^{2r} \\ &\leq s_\phi^{-4r} 2^{2r-1} [E(\hat{\bar{u}} - \bar{u})^{2r} + E(s_y^2 - E(s_y^2))^{2r}] = O(m^{-r}) \end{aligned}$$

for any arbitrary $r > 0$. Hence

$$E[(\tilde{\beta} - \tilde{\tilde{\beta}})^2 s_y^2] = O(m^{-r/2}) \tag{43}$$

for any $r > 0$. Thus, from (35), (36) (39), (40) and (43)

$$m^{-1} \sum_{i=1}^{m} E(\tilde{\phi}_{iWL} - \hat{\phi}_{iWL}^{opt})^2 = O(m^{-1}). \tag{44}$$

Finally,

$$\begin{aligned} m^{-1} \sum_{i=1}^{m} E[(\hat{\phi}_{iWL}^{opt} - \phi_i)(\tilde{\phi}_{iWL} - \hat{\phi}_{iWL}^{opt})] &= -m^{-1} \sum_{i=1}^{m} E[\{(1 - \beta_{opt})\bar{y}_i + \beta_{opt}\bar{y} - \phi_i\} \\ &\quad \times (\tilde{\beta} - \beta_{opt})(\bar{y}_i - \bar{y})] \\ &= -m^{-1} \sum_{i=1}^{m} E[(\tilde{\beta} - \beta_{opt})(\bar{y}_i - \phi_i)(\bar{y}_i - \bar{y})] \\ &\quad + m^{-1} \beta_{opt} \sum_{i=1}^{m} E[(\tilde{\beta} - \beta_{opt})(\bar{y}_i - \bar{y})^2]. \end{aligned} \tag{45}$$

The second term in the right hand side of (45) equals $(m-1)m^{-1}\beta_{opt} E[(\tilde{\beta} - \beta_{opt}) s_y^2]$. But

$$E[(\tilde{\beta} - \beta_{opt}) s_y^2] = E[(\tilde{\tilde{\beta}} - \beta_{opt}) s_y^2 + (\tilde{\beta} - \tilde{\tilde{\beta}}) s_y^2] \tag{46}$$

and

$$\begin{aligned} E[(\tilde{\tilde{\beta}} - \beta_{opt}) s_y^2] &= E\left[\left(\frac{\hat{\bar{u}} + m^{-1}}{s_y^2 + m^{-1}} - \frac{\bar{u}}{E(s_y^2)}\right) s_y^2\right] \\ &= E\left[\hat{\bar{u}} + m^{-1} - m^{-1} E\left(\frac{\hat{\bar{u}} + m^{-1}}{s_y^2 + m^{-1}}\right) - \bar{u}\right] = O(m^{-1}). \end{aligned} \tag{47}$$

Also, as before, one can make $E|(\tilde{\tilde{\beta}} - \tilde{\beta}) s_y^2| = O(m^{-r})$ for any arbitrary $r > 0$. Hence,

$$E(\tilde{\beta} - \beta_{opt}) s_y^2 = O(m^{-1}). \tag{48}$$

The first term in the right hand side of (45) equals

$$-m^{-1}\sum_{i=1}^{m} E[(\tilde{\beta}-\beta_{opt})\{(\bar{y}_i-\bar{y})-(\phi_i-\bar{\phi})\}(\bar{y}_i-\bar{y})] = -(m-1)m^{-1}E[(\tilde{\beta}-\beta_{opt})(s_y^2+s_{y\phi})], \quad (49)$$

where $s_{y\phi}=(m-1)^{-1}\sum_{i=1}^{m}(\bar{y}_i-\bar{y})(\phi_i-\bar{\phi})$. As shown already, $E[(\tilde{\beta}-\beta_{opt})s_y^2]=O(m^{-1})$. Again,

$$E[(\tilde{\beta}-\beta_{opt})s_{y\phi}] = E[\tilde{\beta}-\beta_{opt}]E[s_{y\phi}] + E[(\tilde{\beta}-\beta_{opt})(s_{y\phi}-Es_{y\phi})]. \quad (50)$$

But $E(s_{y\phi})=s_\phi^2$ and $E(\tilde{\beta}-\beta_{opt})=O(m^{-1})$. Further by the Cauchy-Schwarz inequality,

$$E[(\tilde{\beta}-\beta_{opt})(s_{y\phi}-Es_{y\phi})] \leq E^{1/2}(\tilde{\beta}-\beta_{opt})^2 V(s_{y\phi}) = O(m^{-1})V(s_{y\phi}).$$

Since

$$\begin{aligned} V(s_{y\phi}) &= V[(m-1)^{-1}\sum_{i=1}^{m}\bar{y}_i(\phi_i-\bar{\phi})] \\ &= (m-1)^{-2}\sum_{i=1}^{m} u_i(\phi_i-\bar{\phi})^2 = O(m^{-1}), \end{aligned} \quad (51)$$

one gets from (45)-(51),

$$m^{-1}\sum_{i=1}^{m} E[(\hat{\phi}_{iWL}^{opt}-\phi_i)(\tilde{\phi}_{iWL}-\hat{\phi}_{iWL}^{opt})] = O(m^{-1}). \quad (52)$$

The theorem follows now from (33), (34), (44) and (52).