

# Statistical Inference Based on Pooled Data: A Moment-Based Estimating Equation Approach

Howard D. Bondell

Department of Statistics, Rutgers University, Piscataway, NJ 08854, U.S.A.

Aiyi Liu\* and Enrique F. Schisterman

Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, Department of Health and Human Services, 6100 Executive Blvd., Bethesda, MD 20852, U.S.A.

\**email*: liua@mail.nih.gov

**SUMMARY.** We consider statistical inference on parameters of a distribution when only pooled data are observed. A moment-based estimating equation approach is proposed to deal with situations where likelihood functions based on pooled data are difficult to work with. We outline the method to obtain estimates and test statistics of the parameters of interest in the general setting. We demonstrate the approach on the family of distributions generated by the Box-Cox transformation model, and in the process, construct tests for goodness of fit based on the pooled data.

**KEY WORDS:** Pooling biospecimens; Set-based observations; Moments; Box-Cox transformation; Goodness of fit; Lognormal distribution.

## 1. Introduction

The classical approach to conducting statistical inference on an unknown vector of parameters  $\theta$  characterizing a distribution  $F_\theta$  of a random variable  $X$  is based on a random sample of size  $np$  (with both  $n$  and  $p$  being integers) from  $F_\theta$ , say,  $X_1, \dots, X_{np}$ ,

independent and identically distributed according to  $F_\theta$ . Suppose that, in order to reduce the cost of the study, the subjects that yield the  $np$  samples are randomly grouped into  $n$  sets, each of size  $p$ . Subsequently, instead of observing each individual  $X$ , the average of the  $X$ s in each set is observed, yielding  $n$  observations,  $X_j^* = \sum_{i=p(j-1)+1}^{jp} X_i/p$ ,  $j = 1, \dots, n$ , which are often called set-based observations. The  $X^*$ s are also independent and identically distributed, but each following the distribution  $F_\theta^*$  of the average of  $p$  random draws from  $F_\theta$ . We are concerned with inference on  $\theta$  based on the pooled data  $X_1^*, \dots, X_n^*$ .

The data framework described above, often called group testing as a result of testing for a microorganism in pooled biospecimens, appeared initially in the context of screening with dichotomous outcomes (Dorfman, 1943; Sobel and Groll 1959), and was later further developed by Gastwirth and Johnson (1994) and Litvak, Tu and Pagano (1994) in HIV contamination, Sobel and Elashoff (1975), Chen and Swallow (1990), Farrington (1992), Hughes-Oliver and Swallow (1994), and Tu, Litvak and Pagano (1995) in estimating population prevalence, and Barcellos et al. (1997) in localizing disease genes. Recently Weinberg and Umbach (1999) proposed a set-based logistic model to explore the association between a disease and exposures when only the pooled exposure values are available. Farragi, Reiser and Schisterman (2003) and Liu and Schisterman (2003) considered evaluation of diagnostic biomarkers based on pooled specimens whose measurements are assumed to follow normal or gamma distributions. Other areas where pooling biospecimens has been found useful include gene microarray experiments where mRNA samples are often pooled across subjects (Jin et al., 2001; Enard et al., 2002; Kendzierski et al., 2003).

Although the strategy of pooling specimens has been used in practice, methods for analysis of set-based data from such experiments have not been fully and well

developed in the literature, except for certain special cases. This is, perhaps, partly because for a general distribution,  $F_\theta$ , the likelihood methods based on set-based data may not be feasible since the distribution of the averages involves convolution of  $p$  random variables of  $F_\theta$ . The purpose of the present paper is to develop a general methodology for reasonably efficient estimation and testing under a broad class of distributional assumptions, including the family of distributions generated by the Box-Cox transformation model. The context is that  $F_\theta$  possesses and is fully determined by its first several moments, which may be estimated by converting the estimates of the moments of the set-based random variable  $X^*$ .

The paper is arranged as follows. In §2 we shall describe the method under the assumption that  $F_\theta$  can be parameterized by no more than its first three moments. We obtain estimating equations to yield estimates of  $\theta$ . The method can be readily extended to distributions with more than three parameters, but we omit the details at the present time. In §3 we derive the large sample distribution of the estimates. We then apply the method in §4 to the family of distributions generated by the Box-Cox transformation model, which is an extremely versatile class of distributions that includes the normal, lognormal, and (non-central)  $\chi^2$  distributions as special cases. As an extension, we discuss several procedures to test goodness of fit based on the pooled data. The methods are exemplified in §5 using data from a study evaluating oxidative stress and antioxidants on cardiovascular disease in upstate New York. Some comments and discussions appear in §6.

## **2. Estimating Equations Based on the Moments: A General Method**

Our aim is to conduct inference on a  $k$ -parameter vector,  $\theta$ , that characterizes the distribution  $F_\theta$  of interest, based on the set-based observations  $X^*$ s. Except for certain special cases, the likelihood function based on  $X^*$ s will be extremely difficult to derive.

Alternatively we can obtain and connect the moments of  $X^*$  with that of  $X$ , and construct inference based on moment estimates. For this purpose, we assume that  $F_\theta$  has at least  $2k$  moments of which the first  $k$  moments will be used to estimate  $\theta$  and the rest to estimate the variance.

We illustrate the method by assuming that  $\theta$  is at most three-dimensional. Define  $\mu_1 = E(X)$ ,  $\mu_1^* = E(X^*)$ ,  $\mu_r = E\{(X - \mu_1)^r\}$ , and  $\mu_r^* = E\{(X^* - \mu_1^*)^r\}$ , for  $r > 1$ ; these are all functions of  $\theta$ . Then it is straightforward to show that the first three central moments of  $X^*$ , being the mean of  $p$  independent, identically distributed variables of  $X$ , are

$$\mu_1^* = \mu_1, \quad \mu_2^* = \mu_2/p, \quad \mu_3^* = \mu_3/p^2. \quad (1)$$

Replacing the left-hand sides of the equations by their corresponding set-based sample moments and solving for  $\theta$  will then yield an estimator of  $\theta$  based on which inference can be conducted. Putting the above into an estimating equation framework allows us to conveniently utilize the general asymptotic theory on estimating equations (See Serfling, 1980, and §3 below). Define

$$\Psi(w; \theta) = \begin{pmatrix} w - \mu_1(\theta) \\ p\{w - \mu_1(\theta)\}^2 - \mu_2(\theta) \\ p^2\{w - \mu_1(\theta)\}^3 - \mu_3(\theta) \end{pmatrix}, \quad (2)$$

then a consistent estimator,  $\tilde{\theta}$ , of  $\theta$  is the solution to the equation:

$$n^{-1} \sum_{j=1}^n \Psi(X_j^*; \tilde{\theta}) = \mathbf{0}, \quad (3)$$

or equivalently  $\tilde{\theta} = \arg \min_{\theta} n^{-1} \sum_{j=1}^n \Psi(X_j^*; \tilde{\theta})^T \Psi(X_j^*; \tilde{\theta})$ . Note that if  $\theta$  is of dimension  $k < 3$ , we only require the first  $k$  components of  $\Psi$ .

We have restricted our attention to at most three parameters which are sufficient for most practical needs. If more parameters are required, the approach described above

can be readily extended to accommodate the additional parameters, though the higher order moments may have less simple formulas. (See next section below.)

### 3. Distribution Theory for Estimates

Clearly, there is no simple exact distribution theory for the estimator  $\tilde{\theta}$ , since it will depend on the distribution  $F_{\theta}^*$  of  $X^*$ , which, as mentioned earlier, may not be feasible to work with, unless the original distribution is of a particularly convenient form, such as a normal or gamma. Here we derive asymptotic theory for  $\tilde{\theta}$ , on which statistical inference may be based.

In order to obtain the asymptotic variance of the estimator, we also need the next three higher-order moments. Straightforward manipulation leads to the following relationships (for  $p > 1$ ):

$$\begin{aligned}\mu_4^* &= p^{-3}\{\mu_4 + 3(p-1)\mu_2^2\}, \quad \mu_5^* = p^{-4}\{\mu_5 + 10(p-1)\mu_3\mu_2\}, \\ \mu_6^* &= p^{-5}\{\mu_6 + 15(p-1)\mu_4\mu_2 + 10(p-1)\mu_3^2 + 15(p-1)(p-2)\mu_2^3\},\end{aligned}\quad (4)$$

again, all are functions of  $\theta$ .

Following standard asymptotic theory of estimating equations (for example, Serfling, 1980, Chapter 7), we can show, via Taylor expansion, that  $n^{1/2}(\tilde{\theta} - \theta) \xrightarrow{d} N_3(0, \Sigma)$ , where  $\Sigma = ABA^T$ , with,

$$A^{-1} = -E_{\theta} \left\{ \frac{\partial}{\partial \theta} \Psi(X^*; \theta) \right\}, \quad B = E_{\theta} \left\{ \Psi(X^*; \theta) \Psi(X^*; \theta)^T \right\}. \quad (5)$$

For the particular form of the estimating equations (3), we can explicitly express the matrices  $A$  and  $B$  in terms of the central moments of the original distribution, by using the moment relationships given in (1) and (4). Putting  $\theta = (\theta_1, \theta_2, \theta_3)^T$ , we then

have

$$A^{-1} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \theta_1} & \frac{\partial \mu_1}{\partial \theta_2} & \frac{\partial \mu_1}{\partial \theta_3} \\ \frac{\partial \mu_2}{\partial \theta_1} & \frac{\partial \mu_2}{\partial \theta_2} & \frac{\partial \mu_2}{\partial \theta_3} \\ 3p\mu_2 \frac{\partial \mu_1}{\partial \theta_1} + \frac{\partial \mu_3}{\partial \theta_1} & 3p\mu_2 \frac{\partial \mu_1}{\partial \theta_2} + \frac{\partial \mu_3}{\partial \theta_2} & 3p\mu_2 \frac{\partial \mu_1}{\partial \theta_3} + \frac{\partial \mu_3}{\partial \theta_3} \end{pmatrix} \quad (6)$$

and

$$B = \begin{pmatrix} \mu_2/p & \mu_3/p & p^2 \mu_4^* \\ \mu_3/p & p^2 \mu_4^* - \mu_2^2 & p^3 \mu_5^* - \mu_3 \mu_2 \\ p^2 \mu_4^* & p^3 \mu_5^* - \mu_3 \mu_2 & p^4 \mu_6^* - \mu_3^2 \end{pmatrix}. \quad (7)$$

We will need estimates of  $\Sigma$  to construct confidence intervals and test statistics for hypotheses regarding a function of  $\theta$ . Here we propose two approaches to construct estimates of  $A$  and  $B$  which then yield estimates of  $\Sigma$ . One approach is to “plug-in”  $\tilde{\theta}$  for  $\theta$  in the expressions (6) and (7), respectively to obtain estimates of  $A$  and  $B$ .

Alternatively, we may obtain “semi-empirical” estimates for  $A$  and  $B$ , using a two-stage strategy. First replace  $\theta$  by  $\tilde{\theta}$  in (5), and then estimate the resulting functions by their empirical sample means, we have

$$\tilde{A}^{-1} = -n^{-1} \sum_j \frac{\partial}{\partial \theta} \Psi(X_j^*; \tilde{\theta}), \quad \tilde{B} = n^{-1} \sum_j \Psi(X_j^*; \tilde{\theta}) \Psi(X_j^*; \tilde{\theta})^T. \quad (8)$$

Both approaches lead to consistent estimators of  $\Sigma$ . We may then conduct approximate inference on any function of  $\theta$ , using the asymptotic normality of the statistics. For example, a  $100(1 - \alpha)$  per cent confidence interval for a linear function  $l^T \theta$  is  $l^T \tilde{\theta} \pm z_{1-\alpha/2} (l^T \tilde{A} \tilde{B} \tilde{A}^T l / n)^{1/2}$ , with  $z$  being the upper percentile of the standard normal distribution.

## 4. The Box-Cox Transformation Family

### 4.1 Preliminaries

In their seminal paper, Box and Cox (1964) developed a widely used transformation

family for the linear regression model.

$$Y = \begin{cases} (X^\lambda - 1)/\lambda, & \lambda \neq 0, \\ \log X, & \lambda = 0, \end{cases}, \quad (9)$$

where  $\lambda$  is a real-valued, unknown parameter, and  $X$  is the original data, assumed to be strictly positive (in order that all real  $\lambda$  yield real values).

Based on this transformation, a diverse family of distributions for  $X$  is generated. The Box-Cox power transformation assumes that there is some member of the power family of transformations such that when applied to the data, the transformed data are normally distributed. Hence the original data can take on a wide range of possible distributions, and in most practical situations, there exists some member of this family of distributions that is a reasonable model for the data generating mechanism. Three important special cases of this family are the normal ( $\lambda = 1$ ), lognormal ( $\lambda = 0$ ), and (non-central)  $\chi^2$  ( $\lambda = 1/2$ ).

It is now assumed that  $Y$  follows a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . We would like to conduct inference on the parameter vector  $\theta = (\mu, \sigma^2, \lambda)^T$  based on observing  $X^*$ , which as before, are averages of  $p$  random observations of  $X$ .

## 4.2 Inference

### 4.2.1 Estimation of parameters

The standard approach to inference in the Box-Cox model is via maximum likelihood, while other approaches have also been proposed (See Sakia, 1992, for a review). However, when only pooled data are observed, these methods are extremely difficult or impossible to carry out, except in special cases. We shall instead estimate the three parameters based on the pooled data via the moment-based estimating equation method described in §2 and §3.

To proceed, we obtain expressions for the central moments of the distribution of  $X$ , the transformed normal. Once the expressions are obtained, it is then computationally straightforward to derive the estimator and its estimated large sample covariance matrix.

It should be pointed out that the claim that  $Y$  has a normal distribution is not exactly true for  $\lambda \neq 0$ , since  $Y$  is bounded by  $-1/\lambda$  from below (above) for  $\lambda > (<)$  0. This effect in general practice is assumed to be negligible (and usually is), but must be accounted for in our expression for the moments.

We first deal with  $\lambda \geq 0$ , in which case all moments exist for  $X$ . For  $\lambda > 0$ , define  $U = X^\lambda = \lambda Y + 1$ , whose density is a truncated (at zero) normal density:

$$f(u; \mu, \sigma, \lambda) = \frac{1}{|\lambda|\sigma\Phi(\delta)} \phi\left(\frac{u}{|\lambda|\sigma} - \delta\right), \quad (u > 0), \quad (10)$$

where  $\delta = (\lambda\mu + 1)/(|\lambda|\sigma)$ ,  $\phi$ , and  $\Phi$  are the standard normal density and distribution functions. Note that the absolute value sign is used in order to unify the density expressions for both  $\lambda > 0$  and  $\lambda < 0$  (See below).

The moments of  $X$  as functions of  $\mu$ ,  $\sigma$ , and  $\lambda$  are thus given by

$$\mu_1(\mu, \sigma, \lambda) = \frac{1}{|\lambda|\sigma\Phi(\delta)} \int_0^\infty u^{1/\lambda} \phi\left(\frac{u}{|\lambda|\sigma} - \delta\right) du, \quad (11)$$

$$\mu_r(\mu, \sigma, \lambda) = \frac{1}{|\lambda|\sigma\Phi(\delta)} \int_0^\infty \left\{u^{1/\lambda} - \mu_1(\mu, \sigma, \lambda)\right\}^r \phi\left(\frac{u}{|\lambda|\sigma} - \delta\right) du, \quad (r > 1). \quad (12)$$

Note that as  $\lambda \rightarrow 0$ , the moments converge to the moments of the lognormal distribution, for which known formulas are available (See, for example, Aitchison and Brown, 1957), and thus explicit expressions for the estimates of  $\mu$  and  $\sigma^2$  based on the pooled data may be obtained.

For the case  $\lambda < 0$ , we notice from (11)-(12) that if  $X$  is bounded from above then all moments exist; otherwise, only the moments of order  $r < -\lambda$  exist. To ensure feasible

execution of the proposed moment-based procedure, we assume that  $X \leq x_0$  for some  $x_0 > 0$ . If we define  $U = X^\lambda - x_0^\lambda = \lambda Y + 1 - x_0^\lambda$ , then the density of  $U$  and the moments of  $X$  are still given respectively by (10)-(12), but with  $\delta = (\lambda\mu + 1 - x_0^\lambda)/(|\lambda|\sigma)$ , and  $u$  in the integrand being replaced by  $u + x_0^\lambda$ .

Using these moment formulas and the estimating equations (3), we can obtain estimates of  $\theta = (\mu, \sigma^2, \lambda)^T$ . The derivatives required for the asymptotic distribution given by (5)-(7) may be computed by differentiating under the integral sign, or may be computed numerically using the estimated parameters. We may then plug in the empirical moment estimates to obtain the estimated asymptotic covariance matrix in order to construct confidence intervals and test statistics.

Note that if we assume  $\lambda$  to be known, then only the first two moments  $\mu_1$  and  $\mu_2$  are needed to obtain estimates of  $\mu$  and  $\sigma^2$ , and the next two higher moments,  $\mu_3$  and  $\mu_4$  to derive asymptotic variance. See §5 for an explicit example in the lognormal case ( $\lambda = 0$ ).

#### 4.2.2 *Computation and inference regarding $\mu$ and $\sigma^2$*

Often, we are only interested in inference on  $\mu$  and  $\sigma^2$ , and the transformation parameter is regarded as being a nuisance. We adopt the following convenient approach to conduct the inference.

Write  $\mu(\lambda)$  and  $\sigma^2(\lambda)$  to denote the dependency on the transformation scale. For a fixed  $\lambda$ , we obtain  $(\tilde{\mu}(\lambda), \tilde{\sigma}^2(\lambda))$  by using only the first two moment relationships.  $\tilde{\lambda}$  is then found via a grid search using the third moment. Our limited simulation results show that the third moment equation is monotone in  $\lambda$ , when considered as a function of  $(\tilde{\mu}(\lambda), \tilde{\sigma}^2(\lambda), \lambda)$ , and hence the grid search should yield a unique estimate  $\tilde{\lambda}$  of  $\lambda$ .

Once  $\tilde{\lambda}$  is determined from the data, we then proceed, just as in the standard data transformation situation, as if  $\lambda$  were known (to be  $\tilde{\lambda}$ ), to estimate  $\mu$  and  $\sigma^2$  and the

asymptotic variance with the last row and column of  $A^{-1}$  and  $B$  in (6) and (7) being removed.

The appropriateness of such “conditional” inference has been a subject of much debate in the literature (Bickel and Doksum, 1981; Box and Cox, 1982; Doksum and Wong, 1983; Hinkley and Runger, 1984, among others). Since  $\tilde{\lambda}$  is a consistent estimate of  $\lambda$  under the Box-Cox model, the asymptotic equivalency of the “conditional” transformed two-sample t-statistic with the “unconditional” one as in Doksum and Wong (1983) would hold for the t-statistic based on the moment estimates as well. Since formulas for the asymptotic variances are available for both the  $\lambda$  known, and the  $\lambda$  estimated situations, the appropriateness of treating  $\lambda$  as known for problems other than those treated by Doksum and Wong (1983) deserves further investigation.

### 4.3 *Testing Goodness-of-fit*

When inference procedures depend heavily on the distributional assumptions, it is important to justify these assumptions before conducting the inference. Below we propose several goodness-of-fit tests concerning the distribution of  $X$  based on the pooled data  $X^*$ .

#### 4.3.1 *One sample test*

One standard approach to testing goodness-of-fit is to imbed the distribution in question into a larger family of distributions indexed by one or more additional parameters and then test hypotheses regarding these parameters. Since the Box-Cox family is a diverse family that includes many important special cases, a natural extension to the estimation problem is the test for goodness-of-fit based on the pooled data to a hypothesized distribution.

Using the estimate of  $\lambda$ , derived by using the estimating equations (3) with  $\theta = (\mu, \sigma^2, \lambda)^T$  and the moments given in (11) and (12), we may test the fit of the underlying

data to a desired distribution of  $X$ , based solely on the pooled data. For example, testing for goodness-of-fit to a lognormal distribution based on the pooled data is accomplished simply by testing  $H_0 : \lambda = 0$  vs.  $H_1 : \lambda \neq 0$ , using the asymptotically standard normal test statistic  $Z = \tilde{\lambda}/s$ , where  $s$  is the estimated standard error of  $\tilde{\lambda}$  based on the asymptotic distribution derived in §3.

#### 4.3.2 Two-sample extension

The Box-Cox transformation family has been used in the receiver operating characteristic (ROC) curve analysis to evaluate the accuracy of a medical diagnostic test or biomarker that yields continuous outcomes (Faraggi and Reiser, 2002; Zou and Hall, 2000, 2002). A key assumption to warrant the use of the Box-Cox transformation theory in such analysis is that the transformation parameter  $\lambda$  be the same for both diseased and non-diseased outcomes. Below we propose a method of testing this key assumption based on the pooled data.

We observe two independent pooled samples from  $np$  diseased and  $mq$  nondiseased subjects,  $X_j^* = \sum_{i=(j-1)p+1}^{jp} X_i/p$ , ( $j = 1, \dots, n$ ), and  $Y_k^* = \sum_{i=(k-1)q+1}^{kq} Y_i/q$ , ( $k = 1, \dots, m$ ). The individual  $X$ s and  $Y$ s are not observed. We assume that for certain unknown parameters  $\lambda_X$  and  $\lambda_Y$ ,  $(X^{\lambda_X} - 1)/\lambda_X$  and  $(Y^{\lambda_Y} - 1)/\lambda_Y$  both follow (truncated) normal distributions. The null hypothesis to be tested in this situation is  $H_0 : \lambda_X - \lambda_Y = 0$  vs.  $H_1 : \lambda_X - \lambda_Y \neq 0$ .

We may test this hypothesis in the following manner. Let  $\tilde{\lambda}_X$  and  $\tilde{\lambda}_Y$  be the estimates of  $\lambda_X$  and  $\lambda_Y$ , respectively, obtained by solving the estimating equations (3), and  $s_X^2$  and  $s_Y^2$  be their estimated variances derived from the methods described in §3. We then reject  $H_0$  at significance level  $\alpha$  if  $|\tilde{\lambda}_X - \tilde{\lambda}_Y| > z_{1-\alpha/2}s$ , where  $s = (s_X^2 + s_Y^2)^{1/2}$ .

If we do not reject the null hypothesis, we may then feel comfortable in combining the two estimates to obtain a weighted estimate of the common value of  $\lambda$ , use the

common estimate as the true  $\lambda$  to estimate  $\mu$  and  $\sigma^2$  for each group, and then proceed with the analysis as proposed by Zou and Hall (2000, 2002).

### 4.3.3 *General goodness-of-fit*

The goodness-of-fit test based on an estimate of  $\lambda$  in §4.3.1 is technically valid only under the assumption that the true distribution is actually a member of the Box-Cox family. We may also adapt other readily available techniques to test the distributional assumptions without assuming membership in the Box-Cox family.

We will assume that the two distributions,  $F_\theta$  of  $X$  and  $F_\theta^*$  of  $X^*$ , are uniquely determined by each other, which then implies that testing the hypothesis that the unobserved data  $X$  follow the distribution  $F_\theta$  is equivalent to testing the hypothesis that the observed pooled data  $X^*$  follow the distribution  $F_\theta^*$ . While there are certain exceptions to this uniqueness characterization, we suspect that it holds for most of the distributions actually used in practice. We comment on this further in §6.

One simple method is to draw a Q-Q plot of the pooled data versus a hypothetical distribution  $F^*$  of  $X^*$ . Suppose we want to test the hypothesis that a distribution  $F_\theta$  generates the unobserved data from which the pooled data are observed. Using the moment based technique we have developed in the previous sections, we obtain an estimator  $\tilde{\theta}$  of  $\theta$  based on the pooled data  $X_j^*$ , ( $j = 1, \dots, n$ ).

If the quantiles of  $F^*$  are difficult to compute, which is the case in general, We may generate a large number of observations from  $F_{\tilde{\theta}}$ , and group them into sets of size  $p$ , to yield the distribution,  $F_{\tilde{\theta}}^*$ . We then plot the quantiles of this large sample empirical distribution versus the quantiles of the empirical distribution of the observed data, and check for linearity in the plot.

Another approach would be to use a formal goodness-of-fit hypothesis test. One of the most common goodness-of-fit tests of a parametric assumption is the Kolmogorov-

Smirnov test. This test is based on the statistic,

$$D = \sup_x |F_n^*(x) - F_{\hat{\theta}}^*(x)|,$$

the largest difference in cumulative distribution functions between the empirical and theoretical distributions. The distribution of  $D$  under the null hypothesis that the data follows the hypothesized distribution is complicated by the fact that we are using an estimate of  $\theta$ , and not the true parameter. Hence, the critical regions of the standard Kolmogorov-Smirnov test are not valid.

Based on the results of Romano (1988), the following bootstrap method will determine critical values that will yield tests with correct asymptotic significance levels.

*Step 1.* Based on the estimate,  $\tilde{\theta}$ , generate a random sample of size  $np$  from  $F_{\tilde{\theta}^*}$ , and then group into sets of size  $p$  to obtain the pooled sample.

*Step 2.* Compute the empirical distribution,  $F_n^*$ , for this sample.

*Step 3.* Generate the empirical distribution  $F_{\tilde{\theta}^*}^*$  based on a large sample grouped into sets of size  $p$ , as in the Q-Q plot.

*Step 4.* Calculate  $D^* = \sup_x |F_n^*(x) - F_{\tilde{\theta}^*}^*(x)|$  for this sample.

*Step 5.* Repeat a large number of times, and use the frequency distribution of  $D^*$  as the null distribution of  $D$  to find the critical region.

## 5. An example

A population-based sample of randomly selected residents of New York State's Erie and Niagara counties, 35 to 79 years of age, was the focus of this investigation. The New York State Department of Motor Vehicles drivers' license rolls were utilized as the sampling frame for adults between the ages of 35 and 65; whereas the elderly sample (age 65 to 79) was randomly selected from the Health Care Financing Administration database.

A total of 72 men and women were selected for the analyses. Personal history of myocardial infarction and angina pectoris was ascertained by self-reported questionnaire. Participants were asked if they had been diagnosed with angina pectoris confirmed by angiogram or with myocardial infarction. Medical charts were reviewed by a physician for outcome verification and were defined as having coronary heart disease. Participants provided a 12-hour fasting blood specimen for biochemical analysis. A number of parameters were examined in fresh blood samples, including routine Vitamin E levels. We assume that the distribution of Vitamin E concentrations is a member of the Box-Cox family. Fig 1(a) and 1(b) shows the normal Q-Q plots for the original and log-transformed data, respectively.

\*\*\* (Insert Figure 1 here) \*\*\*

**Figure 1.** Normal Q-Q plot of Vitamin E concentration.

A lognormal distribution appears to be a reasonable fit to the data. Based on the full (un-pooled) data, the standard Kolmogorov-Smirnoff test rejects the normal assumption (p-value=0.0023), while not rejecting the lognormal assumption (p-value=0.5). A 95% confidence interval for  $\lambda$  based on the maximum likelihood estimate, which is obtainable when full data are available, is found to be (-0.6924, 0.3334), overlapping zero, further confirming the lognormal assumption.

We now randomly group the subjects into groups of 2 and take the average as the pooled observation. Based on this pooled sample, the moment-based estimate of  $\lambda$  is 0.1096 with a standard error of 0.5565, yielding a confidence interval of (-0.9811, 1.2003). This again indicates lognormality. Moreover, the simulated Q-Q plots of the

pooled data (Figure 2) also supports that data are lognormally distributed. We will therefore proceed under the lognormal assumption.

\*\*\* (Insert Figure 2 here) \*\*\*

**Figure 2.** Lognormal Q-Q plot of Vitamin E concentration with pooled data.

For the lognormal distribution, the four required central moments are given by (Aitchison and Brown, 1957):

$$\begin{aligned}\mu_1 &= \exp\left(\mu + \frac{1}{2}\sigma^2\right), \quad \mu_2 = \mu_1^2\omega^2, \\ \mu_3 &= \mu_1^3\omega^4(\omega^2 + 3), \quad \mu_4 = \mu_1^4\omega^4(\omega^8 + 6\omega^6 + 15\omega^4 + 16\omega^2 + 3)\end{aligned}$$

where  $\omega^2 = \exp(\sigma^2) - 1$ .

We find that

$$\mu = \log\left\{(\mu_1^2 + \mu_2)^{-1/2}\mu_1^2\right\}, \quad \sigma^2 = \log\left\{1 + \mu_1^{-2}\mu_2\right\}.$$

We thus obtain  $(\tilde{\mu}, \tilde{\sigma}^2)$  by respectively replacing  $\mu_1$  and  $\mu_2$  by their moment estimates:

$$\tilde{\mu}_1 = \frac{1}{n} \sum_{i=1}^n X_i^*, \quad \tilde{\mu}_2 = \frac{p}{n} \sum_{i=1}^n (X_i^* - \tilde{\mu}_1)^2,$$

where the  $X^*$ s are the pooled observations. Using the explicit formulas for the moments and the asymptotic variance, some straightforward but tedious algebra yields the asymptotic variances and covariance as:

$$\begin{aligned}n \operatorname{Var}(\tilde{\mu}) &\doteq (4p\gamma^2)^{-1} \{\gamma^6 - 8\gamma^4 + 16\gamma^3 + (2p - 11)\gamma^2 - 4(p - 1)\gamma + 2(p - 1)\}, \\ n \operatorname{Var}(\tilde{\sigma}^2) &\doteq (p\gamma^2)^{-1} \{\gamma^6 - 4\gamma^4 + 4\gamma^3 + (2p - 3)\gamma^2 - 4(p - 1)\gamma + 2(p - 1)\}, \\ n \operatorname{Cov}(\tilde{\mu}, \tilde{\sigma}^2) &\doteq -(2p\gamma^2)^{-1} \{\gamma^6 - 6\gamma^4 + 8\gamma^3 + (2p - 5)\gamma^2 - 4(p - 1)\gamma + 2(p - 1)\},\end{aligned}$$

where  $\gamma = 1 + \omega^2$ .

Notice that the above formulas depend only on  $\sigma^2$ , the variance of the underlying normal distribution, along with the pooling size  $p$ . We may plug in  $\tilde{\gamma} = \exp(\tilde{\sigma}^2)$  to yield consistent estimates of the variances and covariance, and thus to construct test statistics and confidence intervals.

Applying these formulas we obtain the estimates of  $(\mu, \sigma^2)$  and their estimated standard errors, for both the pooled and un-pooled data. The results are presented in Table 1. For comparison, estimates based on pooled data with group size of 3 and 4 are also given in the Table.

**Table 1**

Estimate ( $\pm$  standard error) of the lognormal mean  $\mu$  and variance  $\sigma^2$ .

$p$	$n$	$\tilde{\mu}$	$\tilde{\sigma}^2$
1	72	2.6421 ( $\pm 0.0498$ )	0.1733 ( $\pm 0.0373$ )
2	36	2.6408 ( $\pm 0.0519$ )	0.1757 ( $\pm 0.0465$ )
3	24	2.6396 ( $\pm 0.0540$ )	0.1781 ( $\pm 0.0545$ )
4	18	2.6405 ( $\pm 0.0554$ )	0.1764 ( $\pm 0.0603$ )

To evaluate the performance of these estimates, we also computed the maximum likelihood estimates  $\hat{\mu}$  of  $\mu$  and  $\hat{\sigma}^2$  of  $\sigma^2$ , using the un-pooled data. It turned out that  $\hat{\mu} = 2.6466$  and  $\hat{\sigma}^2 = 0.1609$ , with standard errors 0.0473 and 0.0270, respectively. We observe that, due to the small value of  $\sigma^2$ , which is common in lognormal data, there is not a great deal of efficiency loss in the moment based estimates as compared with the maximum likelihood estimates, especially in the estimation of  $\mu$ . In addition, there is also only a small loss of efficiency as we pool the data. This small efficiency

loss is in agreement with previous studies on the merits of pooling data, under normal and gamma assumptions, to reduce costs associated with bioassays. See Faraggi et al. (2003), Liu and Schisterman (2003) and Weinberg and Umbach (1999).

## 6. Discussion

### 6.1 *Comments on Goodness-of-fit Test*

In the goodness-of-fit testing problem, we are testing a hypothesis regarding the underlying distribution of the unpooled data. It is implicitly assumed that the distribution of the convolution is in one-to-one correspondence with the distribution of the individual observations. This is true under general regularity conditions. For example, a nonvanishing characteristic function is a sufficient condition for this one-to-one correspondence. For other characterization conditions see Prokhorov and Ushakov (2002) and the references therein. Regardless of the uniqueness of the characterization, the type I error of the test is unaffected. However, the test will be unable to detect the difference between any two original distributions that may yield the identical convolution distribution. An additional question then arises, if the characterization is not unique, how different are the generating distributions with respect to an underlying distance such as Kolmogorov-Smirnoff, or (symmetrized) Kullback-Leibler?

### 6.2 *Other Comments And Further Directions*

In this paper we proposed inference on pooled data under parametric assumptions on the individual observations. We also suggested methods to test these parametric assumptions. One of the problems inherent in this type of set-based data is created by the central limit theorem. The pooled data tend to a normal distribution as the pooling size increases, and even for small to moderate pooling sizes, much of the skewness (and higher moments) of the original distribution is lost in the set-based distribution. While the loss in variability is linear in the pooling size, this loss of skewness is quadratic,

as can be seen by the moments (1). This hampers the ability to detect differences in distributional shape even for modest pooling sizes.

To our knowledge, the current paper is the first to present a general methodology for dealing with set-based data under a broad class of parametric distributional assumptions. As the pooling of data becomes a more common procedure, particularly in the area of evaluation of disease biomarkers, more research on methods to deal with this form of data needs to be done. For example, under a parametric assumption, we may be able to use Edgeworth expansions to write out an approximate likelihood function for the set-based data and proceed via likelihood methods. The accuracy of inference based on these approximations would be of interest.

Non-parametric methods for set-based data may also be appropriate. A possible alternative to the parametric models proposed in this paper, would be an approach based on density deconvolution, which again appears to be technically and computationally challenging.

#### ACKNOWLEDGEMENTS

The authors thank W. Jack Hall and Kai F. Yu for helpful discussion and suggestions.

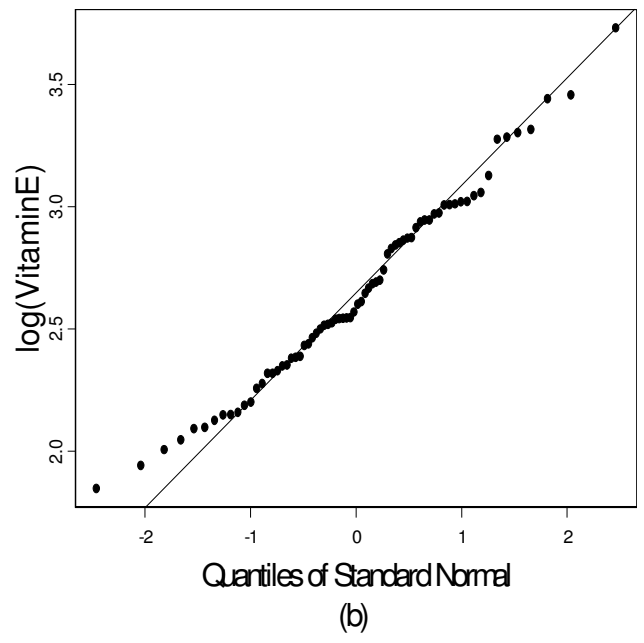
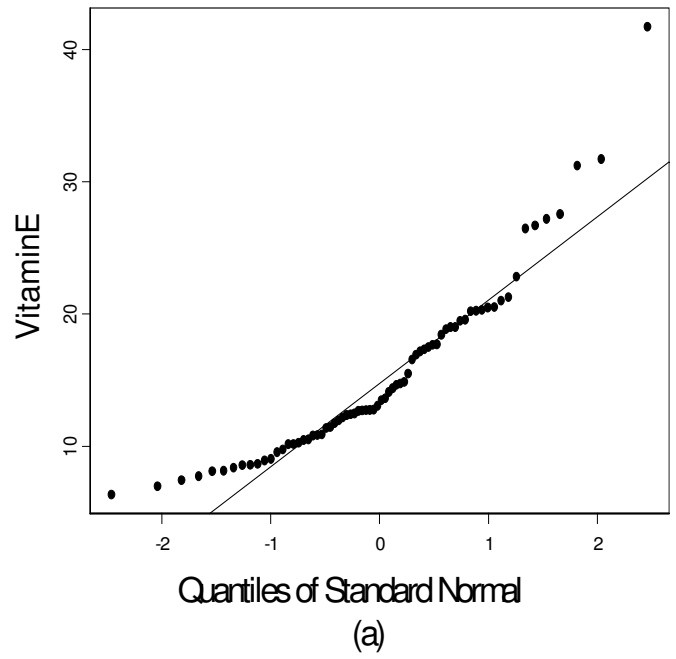
#### REFERENCES

- Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution*. Cambridge: the University Press.
- Barcellos, L. F., Klitz, W., Field, L. L., Tobias, R., Bowcock, A. M., Wilson, R., Nelson, M. P., Nagatomi, J., Thomson, G. (1997). Association mapping of disease loci, by

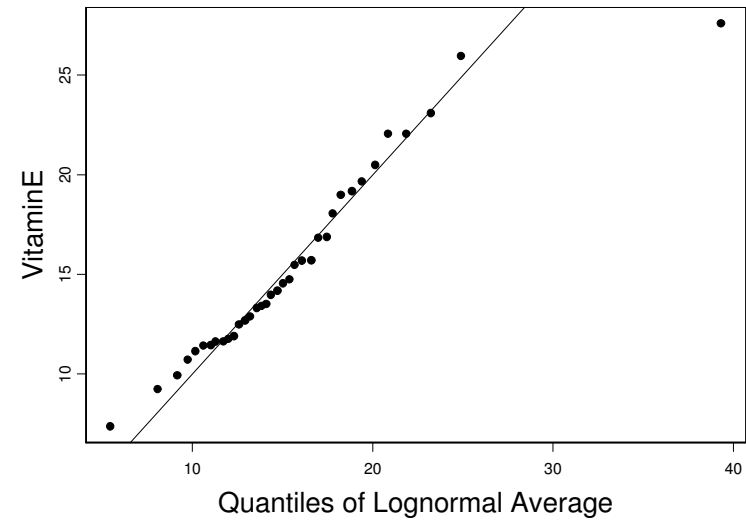
- use of a pooled DNA genomic screen. *American Journal of Human Genetics* **61**, 734-47.
- Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association* **76**, 296-311.
- Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B* **26**, 211-52.
- Box, G. E. P. and Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association* **77**, 209-10.
- Chen, C. L. and Swallow, W. H. (1990). Using group testing to estimate a proportion, and to test the binomial model. *Biometrics* **46**, 1035-46.
- Doksum, K. A. and Wong, C-W. (1983). Statistical tests based on transformed data. *Journal of the American Statistical Association* **78**, 411-7.
- Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics* **14**, 436-40.
- Enard, W., Khaitovich, P., Klose, J., Zollner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G. M., Bontrop, R. E. and Paabo, S. (2002). Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340-3.
- Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine* **21**, 3093-106.
- Faraggi, D., Reiser, B. and Schisterman, E. F. (2003). ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine* **15**, 2515-27.
- Farrington, C. (1992). Estimation prevalence by group testing using generalized linear models. *Statistics in Medicine* **11**, 1591-7.
- Gastwirth, J. and Johnson, W. (1994). Screening with cost-effective quality control:

- Potential applications to HIV and drug testing. *Journal of the American Statistical Association* **89**, 972-81.
- Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association* **79**, 302-20.
- Hughes-Oliver, J. M. and Swallow, W. H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association* **89**, 982-93.
- Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G. and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* **29**, 389-95.
- Kendzioriski, C. M., Zhang, Y. , Lan, H. and Attie, D. (2003). The efficiency of pooling m RNA in microarray experiments. *Biostatistics* **4**, 465-77.
- Litvak, E. , Tu, X. M. and Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association* **89** , 424-34.
- Liu, A. and Schisterman, E. F. (2003). Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal* **45**, 631-44.
- Prokhorov, A. V. and Ushakov, N. G. (2002). On the problem of reconstructing a summands distribution by the distribution of their sum. *Theory of Probability and its Applications* **46**, 420-30.
- Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association* **83**, 698-708.
- Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician* **41**, 169-78.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

- Sobel, M. and Groll, P. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *The Bell System Technical Journal* **38**, 1179-252.
- Sobel, M. and Elashoff, R. (1975). Group testing with a new goal: Estimation. *Biometrika* **62**, 181-93.
- Tu, X. M., Litvak, E. and Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika* **82**, 287-97.
- Weinberg C. R. and Umbach, M. (1999). Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* **55**, 718-26.
- Zou, K. H. and Hall, W. J. (2000). Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* **27**, 621-31.
- Zou, K. H. and Hall, W. J. (2002). Semiparametric and parametric transformation models for comparing diagnostic markers with paired design. *Journal of Applied Statistics* **29**, 803-16.



**This is Figure 1.**



**This is Figure 2.**