Biometrika (2005), **92,** 3, *pp.* 543–557

Printed in Great Britain

Smooth quantile ratio estimation

BY FRANCESCA DOMINICI

Department Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 21205-2179, U.S.A.

fdominic@jhsph.edu

LESLIE COPE

Department of Oncology, The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, 550 North Broadway, Baltimore, Maryland 21205-2011, U.S.A.

cope@jhu.edu

DANIEL Q. NAIMAN

Department of Applied Mathematics and Statistics, The Whiting School of Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218-2682, U.S.A.

daniel.naiman@jhu.edu

AND SCOTT L. ZEGER

Department Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 21205-2179, U.S.A.

szeger@jhsph.edu

SUMMARY

We propose a novel approach to estimating the mean difference between two highly skewed distributions. The method, which we call smooth quantile ratio estimation, smooths, over percentiles, the ratio of the quantiles of the two distributions. The method defines a large class of estimators, including the sample mean difference, the maximum likelihood estimator under log-normal samples and the *L*-estimator. We derive asymptotic properties such as consistency and asymptotic normality, and also provide a closed-form expression for the asymptotic variance. In a simulation study, we show that smooth quantile ratio estimation has lower mean squared error than several competitors, including the sample mean difference and the log-normal parametric estimator in several realistic situations. We apply the method to the 1987 National Medicare Expenditure Survey to estimate the difference in medical expenditures between persons suffering from the smoking attributable diseases, lung cancer and chronic obstructive pulmonary disease, and persons without these diseases.

Some key words: Comparing means; Health expenditure; Log-normal; Order statistic; Q–Q plot; Regression spline; Smoking.

1. Introduction

This paper is motivated by the question of how to compare medical expenditures between cases, who are persons with lung cancer or chronic obstructive pulmonary disease, and controls, who are persons without a major smoking attributable disease in a given year; that is, we seek to estimate the difference $\Delta = E(Y_1) - E(Y_2)$, where Y_1 and Y_2 are random variables representing the expenditures for a case and a control, respectively. We estimate Δ from the 1987 National Medical Expenditure Survey (National Center for Health Services Research, 1987), which provides data on annual medical expenditures and disease status for a representative sample of U.S. non-institutionalised adults.

Two special features of these data are that, in each group, the distribution of the nonzero medical expenditures is highly positively skewed, see Fig. 1, and that there are far fewer cases than controls, the two sample sizes being 118 and 2262.

Other concerns that arise in studying expenditure data include the existence of a significant fraction of zero expenditures, right censoring and lack of independence among observations within clusters (Lipscomb et al., 1999). The general problem of comparing costs among two or more groups is discussed by for example Duan (1983), O'Brien (1988), Fenn et al. (1996), Lin et al. (1997), Hlatky et al. (1997), Lin (2000) and Tu & Zhou (1999).

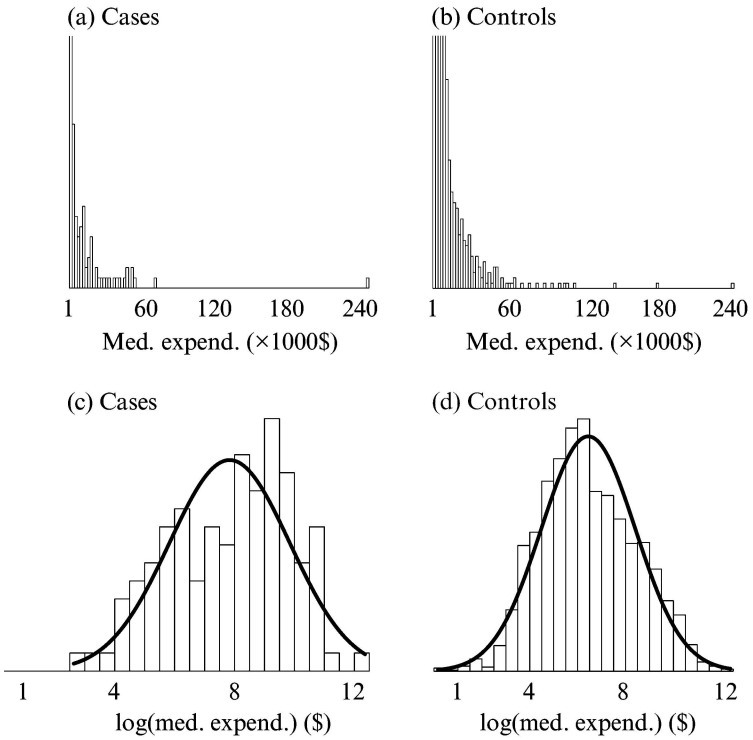

Fig. 1. Histrograms of nonzero Medicare medical expenditures for the 1987 National Medical Expenditure Survey with and without a logarithm transformation, and for individuals in the case and control groups. For clarity of exposition the histogram of the expenditures has been truncated at the top. The solid curves in (c) and (d) are density functions from Normal distributions with means $\hat{v}_1 = n_1^{-1} \sum_{i=1}^{n_1} \log y_{1i}$ and $\hat{v}_2 = n_2^{-1} \sum_{i=1}^{n_2} \log y_{2i}$, and variances $\hat{\sigma}_1^2 = n_1^{-1} \sum_{i=1}^{n_1} (\log y_{1i} - \hat{v}_1)^2$ and $\hat{\sigma}_2^2 = n_2^{-1} \sum_{i=1}^{n_2} (\log y_{2i} - \hat{v}_2)^2$, for the case and control groups, respectively.

Let $y_{11}, \ldots, y_{1n_1}$ and $y_{21}, \ldots, y_{2n_2}$ be the observed nonzero costs in the case and control groups. An obvious estimator of Δ is the difference in sample means $\bar{y}_1 - \bar{y}_2$ where $\bar{y}_g = n_g^{-1} \sum_{i=1}^{n_g} y_{gi}$ $(g = 1, 2)$. However, with highly skewed distributions, this unbiased estimator suffers from sensitivity to extremely large observations.

An obvious approach is to try a log-normal model, in which $\log y_{gi} \sim N(v_g, \sigma_g^2)$, for $i = 1, \ldots, n_g$ and $g = 1, 2$, and now $\Delta = \exp(v_1 + \sigma_1^2/2) - \exp(v_2 + \sigma_2^2/2)$ (Aitchison & Shen, 1980). The maximum likelihood estimator of Δ is biased (Zellner, 1971), but has reduced variability relative to the sample mean difference. Zhou et al. (1997) and Zhou & Gao (1997) have studied methods for testing the null hypothesis that $\Delta = 0$ under the log-normal model.

However, in most applications involving expenditures, including the particular context that has motivated this work, the symmetry implicit in the log-normal model is not based upon any meaningful mechanism and is not likely to be realistic: the shape of the left-hand tail is determined by administrative actions that control access to care, small charges that can occur for prescriptions, and minor preventative services; the shape of the right-hand tail is determined by occurrence of major diseases and traumatic events such as myocardial infarctions, strokes and so on. Since distinct processes influence each tail, we should not a priori expect them to have the same shape.

Deviations from the log-normal model are plainly seen in a quantile–quantile plot. Under the log-normal model, the logarithms of the quantiles from each distribution satisfy the linear equation

$$\log Q_1(p) = \left(v_1 - \frac{\sigma_1}{\sigma_2} v_2 \right) + \frac{\sigma_1}{\sigma_2} \log Q_2(p), \tag{1}$$

where, for $0 < p < 1$, $Q_1(p)$ and $Q_2(p)$ are the quantile functions for the case and control groups respectively. Figure 2 displays the sample Q–Q plot of the log expenditures for the cases versus those for the controls, as well as the bold straight line corresponding to the maximum likelihood estimates of the log-normal parameters for each sample. The Q–Q plot clearly deviates from linearity.

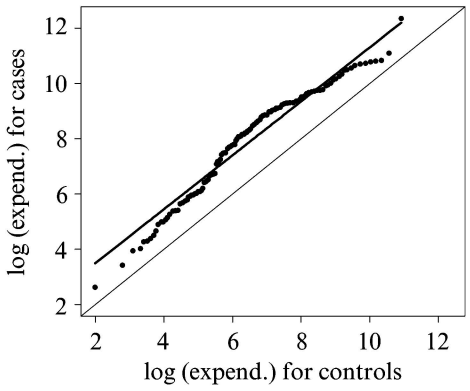

Fig. 2. Quantile–quantile plot of log nonzero Medicare expenditures for cases and controls. The bold straight line is the Q–Q plot if each sample is assumed to come from a log-normal model with parameter values estimated by maximum likelihood.

In general, we might assume that $Q_1(p)$ is an arbitrary function of $Q_2(p)$, that is $Q_1(p) = g\{Q_2(p)\}$ or equivalently $F_1(y) = F_2\{h(y)\}$, where $F_g(y)$ $(g = 1, 2)$ are the cumulative distribution functions of Y_1 and Y_2. Doksum & Sievers (1976) define $h(.)$ as the amount of 'shift' needed to bring the first sample Y_1 up to the second Y_2 in distribution; for example, we might assume $Q_1(p)$ to be a smooth function of $Q_2(p)$ with λ degrees of freedom, $Q_1(p) = s\{Q_2(p), \lambda\}$, where s is a parametric or nonparametric smoother.

Instead, we assume that the log quantile ratio is a smooth function of the percentile p with λ degrees of freedom:

$$\log\left\{\frac{Q_1(p)}{Q_2(p)}\right\} = s(p, \lambda) \quad (0 < p < 1). \tag{2}$$

The basic idea of smooth quantile ratio estimation is to replace the empirical quantiles $\hat{Q}_1(p)$ and $\hat{Q}_2(p)$ with smoother and less variable versions obtained by smoothing the log-transformed ratio of the two quantile functions across percentiles. This produces an estimator of Δ that tends to be less variable than the sample mean difference but with small bias. For different distributional assumptions, shapes of $s(p, \lambda)$ and choices of λ, smooth quantile ratio estimation encompasses a rich class of estimators including the sample mean difference, the maximum likelihood estimator under log-normal samples and L-estimators.

The method has three possible advantages over the shift estimator when estimating Δ. First, the procedure of the shift estimator does not treat the two quantile functions symmetrically, as would be natural when the target for inference is Δ. Secondly, the smooth function s would take arguments on the positive real line making the choice of λ critical. Instead, smooth quantile ratio estimation 'spends' its degrees of freedom λ over the interval $(0, 1)$ rather than over the real line, and hence imposes stronger smoothness constraints in the tails where little information is available in our smaller sample. Thirdly, if we then use the fitted values from the smoothed Q–Q plot to calculate $\hat{\Delta}$, this estimator is asymptotically equivalent to the difference in the sample means. This is because, for large samples, the fitted values of the smoothed Q–Q plot are estimates of the empirical quantile functions and therefore the average of these fitted values will reproduce the sample mean.

2. Smooth quantile ratio estimation

2·1. *Definition*

Let Y_1 and Y_2 be two positive random variables, with cumulative distribution functions F_1 and F_2, and define Q_1 and Q_2 to be the corresponding quantile functions so that $Q_g(p) = F_g^{-1}(p)$ and $F_g\{Q_g(p)\} = \mathrm{pr}\{Y_g \leqslant Q_g(p)\} = p$, for $g = 1, 2$ and $0 < p < 1$. Our goal is to estimate

$$\Delta = E(Y_1) - E(Y_2) = \int_0^1 \{Q_1(p) - Q_2(p)\}dp \tag{3}$$

under the assumption that the ratio of the quantiles is a smooth function of the percentiles with λ degrees of freedom:

$$\log\left\{\frac{Q_1(p)}{Q_2(p)}\right\} = s(p, \lambda) \quad (0 < p < 1). \tag{4}$$

Equations (3) and (4) lead to

$$\Delta = \int_0^1 Q_1(p)[1 - \exp\{-s(p, \lambda)\}]dp = \int_0^1 Q_2(p)[\exp\{s(p, \lambda)\} - 1]dp. \tag{5}$$

2·2. *Estimation approach*

Let $y_1 = (y_{11}, y_{12}, \ldots, y_{1n_1})$ be a random sample of size n_1 from F_1, and let $y_2 = (y_{21}, y_{22}, \ldots, y_{2n_2})$ be a random sample of size n_2 from F_2. We define $y_{(g)} = (y_{g(1)}, y_{g(2)}, \ldots, y_{g(n_g)})$ to be the order statistics for the sample from F_g. We first estimate Δ for the case $n_1 = n_2 = n$, and then extend our definition to the more common situation, $n_1 \ll n_2$.

First, we define a regression model for $s(p, \lambda)$ and we use it to smooth the observed log ratios $\log(y_{1(1)}/y_{2(1)}), \ldots, \log(y_{1(n)}/y_{2(n)})$ across percentiles. This is the parametric part. Secondly, we estimate Δ by using the smoothed quantile ratios and nonparametric estimates of F_1 and F_2. The two steps are detailed below.

Step 1. We impose a smoothness assumption for $s(p, \lambda)$ by assuming a regression model:

$$\log\left(\frac{y_{1(i)}}{y_{2(i)}}\right) = s(p_i, \beta) + \varepsilon_i \quad (i = 1, \ldots, n), \tag{6}$$

where $s(p_i, \beta) = \sum_{j=0}^{\lambda} B_j(p_i)\beta_j$, $p_i = i/(n+1)$, and $B_j(p)$ are orthonormal basis functions, with $B_0(p) = 1$. We estimate β by $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_\lambda)$ by ordinary least squares, although alternative methods could be substituted.

Step 2. We define

$$u_1 = (u_{11}, \ldots, u_{12n}) = (y_{1(1)}, \ldots, y_{1(n)}, y^*_{1(1)}, \ldots, y^*_{1(n)}),$$

$$u_2 = (u_{21}, \ldots, u_{22n}) = (y_{2(1)}, \ldots, y_{2(n)}, y^*_{2(1)}, \ldots, y^*_{2(n)})$$

to be samples of size $2n$, where $y^*_{1(i)} = y_{2(i)} \exp\{s(p_i, \hat{\beta})\}$, $y^*_{2(i)} = y_{1(i)} \exp\{-s(p_i, \hat{\beta})\}$, and $s(p_i, \hat{\beta})$ are the fitted values for the regression model (6). We estimate Δ by

$$\begin{aligned}\hat{\Delta}_{\mathrm{SQ}}(u_1, u_2, \lambda) &= \bar{u}_1 - \bar{u}_2 \\ &= \frac{1}{2n}\sum_{i=1}^{n} y_{1(i)}[1 - \exp\{-s(p_i, \hat{\beta})\}] + \frac{1}{2n}\sum_{i=1}^{n} y_{2(i)}[\exp\{s(p_i, \hat{\beta})\} - 1].\end{aligned} \tag{7}$$

The estimator $\hat{\Delta}_{\mathrm{SQ}}(u_1, u_2, \lambda)$ is then the sample mean difference between the two extended samples u_g $(g = 1, 2)$, by which we mean the vector of actual observations $y_{(g)}$ augmented with the transformed values from the other sample $y^*_{(g)}$. Two desirable properties are immediately evident. The estimator is symmetric in the two samples: $\hat{\Delta}_{\mathrm{SQ}}(u_1, u_2, \lambda) = -\hat{\Delta}_{\mathrm{SQ}}(u_2, u_1, \lambda)$. Furthermore, $\hat{\Delta}_{\mathrm{SQ}}(u_1, u_2, \lambda)$ can be viewed as a linear combination of order statistics, but with weights estimated from the data, and thus it is related to *L*-estimation (Huber, 1996, pp. 16–20; Serfling, 1980, Ch. 8).

To simplify the notation, we will denote $\hat{\Delta}_{\mathrm{SQ}}(u_1, u_2, \lambda)$ by $\hat{\Delta}_{\mathrm{SQ}}(\lambda)$. Note that if $\lambda = n$ then the basis functions in (6) can be chosen so that $s(p, \hat{\beta})$ interpolates the values $\log(y_{1(i)}/y_{2(i)})$. In this case, we treat the two samples as independent, and $\hat{\Delta}_{\mathrm{SQ}}(\lambda)$ reduces to the difference in sample means $\bar{y}_1 - \bar{y}_2$.

In the motivating application, n_1 is much smaller than n_2 so the order statistics from the two samples do not line up perfectly. In this case we calculate $\hat{\Delta}_{SQ}(\lambda)$ with y_2 replaced by q_2, the linear interpolant of the order statistics $y_{2(i)}$ at the grid of points $p_{1i} = i/(n_1 + 1)$, for $i = 1, \ldots, n_1$, with a similar modification if $n_1 > n_2$.

2·3. *Special cases*

For different shapes of $s(p, \beta)$, choices of the basis functions $B_j(p)$ and specifications of the parametric cumulative distribution functions, smooth quantile ratio estimation encompasses a large class of estimators.

Example 1. If $Y_g \sim \mathrm{Un}[0, \theta_g]$, for $g = 1, 2$, then $Q_1(p)/Q_2(p) = \theta_1/\theta_2$ and $\Delta = (\theta_1 - \theta_2)/2$. The smooth quantile ratio estimator of Δ, denoted by $\hat{\Delta}_{SQ}(\mathrm{Un}, \lambda = 0)$, is obtained by fitting the regression model (6) with $B_0(p) = 1$ and $B_1(p) = 0$, and using $s(p_i, \hat{\beta}) = \hat{\beta}_0 = \bar{l}_1 - \bar{l}_2$, where $l = \log y$ in equation (7). This leads to

$$\hat{\Delta}_{SQ}(\mathrm{Un}, \lambda = 0) = \tfrac{1}{2}[\bar{y}_1\{1 - \exp(-\hat{\beta}_0)\} - \bar{y}_2\{1 - \exp(\hat{\beta}_0)\}].$$

Note that $\hat{\Delta}_{SQ}(\mathrm{Un}, \lambda = 0)$ is not the maximum likelihood estimator of Δ, which is equal to $(y_{1(n)} - y_{2(n)})/2$.

Example 2. If $Y_g \sim \mathrm{LN}(v_g, \sigma_g)$, for $g = 1, 2$, then $\log\{Q_1(p)/Q_2(p)\} = \beta_0 + \beta_1 \Phi^{-1}(p)$, where $\Phi^{-1}(p)$ is the quantile function of the $N(0, 1)$ random variable, $\beta_0 = (v_1 - v_2)$, $\beta_1 = (\sigma_1 - \sigma_2)$ and $\Delta = \exp(v_1 + \sigma_1^2/2) - \exp(v_2 + \sigma_2^2/2)$. The smooth quantile ratio estimator of Δ, denoted by $\hat{\Delta}_{SQ}(\mathrm{LN}, \lambda = 1)$, is obtained by fitting the regression model (6) with $B_0(p) = 1$ and $B_1(p) = \Phi^{-1}(p)$, and using $s(p_i, \hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 \Phi^{-1}(p_i)$ in equation (7). Note that $\hat{\Delta}_{SQ}(\mathrm{LN}; 1)$ is not the maximum likelihood estimator of Δ, which instead is defined as $\mathrm{LN} = \exp(\bar{l}_1 + h_1^2/2) - \exp(\bar{y}_2 + h_2^2/2)$, where $l = \log y$ and h is the standard deviation of the log-transformed data. Also note that, if $\sigma_1 = \sigma_2$, then $s(p, \lambda)$ is constant in p and equal to β_0.

3. ASYMPTOTIC PROPERTIES

The smooth quantile ratio estimator is nearly an L-estimator except that the weight function applied to the empirical quantile function is stochastic instead of deterministic. In this section we derive the asymptotic distribution of smooth quantile ratio estimation by extending standard results from L-statistic theory.

THEOREM 1: *Consistency and asymptotic normality of* $\hat{\beta}$. *Assume that* $n_1, n_2 \to 0$ *and there exist* M, b_1, b_2 *and* $\delta > 0$ *such that*

(i) $|\log F_g^{-1}(p)| \leqslant M p^{-\frac{1}{2} + b_1 + \delta}(1 - p)^{-\frac{1}{2} + b_2 + \delta}$, *for* $g = 1, 2$;

(ii) *the basis functions* $|B_j(p)|$ *are continuously differentiable on* $(0, 1)$ *and* $|B_j(p)| \leqslant M p^{-b_1}(1 - p)^{-b_2}$.

Then $\hat{\beta}_j$ *is strongly consistent for* β_j, *for* $j = 0, 1, \ldots, \lambda$. *If, in addition,*

$$\lim_{n_1, n_2 \to \infty} \{n_1/(n_1 + n_2)\}$$

exists and is in the interval $(0, 1)$, *then* $\hat{\beta} - \beta$ *asymptotically has a multivariate normal distribution with mean* 0 *and covariance matrix* $\Sigma = (\sigma_{ij})$, *where*

$$\sigma_{ij} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \int_0^1 \int_0^1 \{\min(p, q) - pq\} B_i(p) B_j(q) dp\, dq.$$

Remark 1. For consistency alone, the condition (i) can be replaced by the less stringent condition

(i′) $\int |\log Y_g|^r dF_g(x) < \infty$, $|\log F_g^{-1}(p)| \leqslant Mt^{-1+b_1+\delta}(1-t)^{-1+b_2+\delta}$, for $g = 1, 2$.

Proof of Theorem 1. The consistency and asymptotic normality of the $\hat{\beta}_j$'s follow as a corollary to the *L*-statistic results of Shorack (1972) and Wellner (1977). The Cramer–Wold device is applied to show that $\hat{\beta} - \beta$ has an asymptotic multivariate normal distribution. □

THEOREM 2: *Asymptotic normality of smooth quantile ratio estimation. Assume that* $n_1, n_2 \to \infty$ *and that* $\lambda_g = \lim_{n_1,n_2\to\infty}\{n_1/(n_1+n_2)\}$ *exists and lies in* $(0, 1)$. *Suppose there exist* M, b *and* $\delta > 0$ *such that the following conditions hold for* $g = 1, 2$, $j = 1, 2, \ldots, \lambda$ *and all* $p \in (0, 1)$:

(i) $F_g^{-1} \leqslant M\{p(1-p)\}^{-b+\delta}$ *and* $|\log F_g^{-1}| \leqslant M\{p(1-p)\}^{-\frac{1}{2}+\delta}$;

(ii) $\exp\{(-1)^g \sum_{j=1}^{\lambda} \beta_j B_j(p)\} \leqslant M\{p(1-p)\}^{-\frac{1}{2}+b}$;

(iii) $|B_j(p)| \leqslant M\{p(1-p)\}^{-\delta/(\delta+2)}$.

Then $\sqrt{n}(\hat{\Delta} - \Delta)$ *is asymptotically normal with mean* 0 *and variance* σ^2, *where*

$$\sigma^2 = \int_{p=0}^{1}\int_{q=0}^{1} \{\min(p, q) - pq\}\{\lambda_1\eta_1(p)\eta_1(q) + \lambda_2\eta_2(p)\eta_2(q)\}dp\,dq, \tag{8}$$

$$\eta_g(p) = \frac{F_g^{-1}(p) + \frac{1}{2}[F_1^{-1}(p) + F_2^{-1}(p) - \int_0^1 \sum_{j=1}^{\lambda} B_j(q)\{F_1^{-1}(q) + F_2^{-1}(q)\}dq]}{(-1)^g F_g^{-1}(p) f_g\{F_g^{-1}(p)\}}.$$

A sketch of the proof is found in the Appendix; details are available in L. Cope's 2003 Ph.D. thesis from Johns Hopkins University.

Expression (8) appears unwieldy but is straightforwardly calculated. In § 4 we use adaptive quadrature to calculate this asymptotic variance expression in one special case, and we demonstrate that it gives a good approximation.

All of the conditions in Theorem 2 are easily interpreted. They simply require that the quantile functions, the log quantile functions, the quantile ratio, its reciprocal and the basis functions do not grow too rapidly as $p \to 0$ and $p \to 1$. A similar comment can be made about the conditions in Theorem 1. These assumptions are all made in order to ensure the square integrability of the smoothed estimates of the quantile functions, which involves a second moment condition on the random variables Y and $\log Y$. The final assumption in Theorem 1 ensures that the two sample sizes converge in a smooth way.

Many distributions with finite second moments and relatively smooth quantile functions satisfy these criteria, including the log-normal family. It can be easily shown that a finite mixture distribution will satisfy these conditions if and only if all of the individual distributions do. The log quantile function of an exponential distribution is not integrable, and therefore does not satisfy the conditions. However, simulation studies in L. Cope's thesis indicate that when both samples are drawn from the exponential distribution it is possible to calculate the smooth quantile ratio estimator and its asymptotic variance, with good agreement for large samples.

Example 3. If both samples are drawn from log-normal distributions, then the smooth quantile ratio estimator is consistent and asymptotically normal. In this case,

$$F_g = \Phi\left(\frac{\log x - \mu_g}{\sigma_g}\right), \quad F_g^{-1} = \exp\{\mu_g + \sigma_g\Phi^{-1}(p)\}, \quad \log F_g^{-1} = \mu_g + \sigma_g\Phi^{-1}(p).$$

The log quantile ratio is a linear function of $\Phi^{-1}(p)$, so that a natural orthonormal basis is $B_0 \equiv 1$ and $B_1 = \Phi^{-1}(p)$.

Example 4. If both samples are drawn from Pareto distributions, for which $F(x) = 1 - b^a x^{-a}$ and $a > 2$, then the smooth quantile ratio estimator is strongly consistent and asymptotically normal. The Pareto distribution is an interesting example for the method because it is very heavy-tailed, and has a finite kth moment only if the shape parameter $a \geqslant k$. Its density function is $f(x) = ab^a x^{-a-1}$, where $x \geqslant 1$ and $a, b > 0$. The log quantile function is given by $\log F^{-1}(p) = \log b - \log(1-p)/a$, leading to the two basis functions $B_0(p) \equiv 1$ and $B_1(p) = \log(1-p)$. Note that the orthonormalised version of B_1 is equal to $\{\log(1-p)+1\}/\sqrt{3}$.

4. Simulations and data analysis

The simulations in this section demonstrate that $\hat{\Delta}_{\mathrm{SQ}}(\lambda)$ often has substantially lower mean squared error and bias than commonly used estimators of Δ, such as the maximum likelihood estimator for log-normal populations and the sample mean difference. We also investigate the performance of the asymptotic variance of the smooth quantile ratio estimator in equation (8). At the end of the section we apply the method to the data represented in Figs 1 and 2.

As mentioned above, when $\lambda = n$ the smooth quantile ratio estimator is asymptotically equivalent to the difference between sample means. To obtain improved mean squared error, the smoothing parameter λ must be small compared to n. So far we have treated λ as a prespecified parameter, but in practice one may prefer to estimate the value. To estimate λ we use a B-fold crossvalidation method (Efron & Tibshirani, 1993, p. 240) which minimises

$$\mathrm{CV}(\lambda) = \sum_{b=1}^{B} \{(\bar{y}_1^{(b)} - \bar{y}_2^{(b)}) - \hat{\Delta}_{\mathrm{SQ},\lambda}^{(-b)}\}^2, \tag{9}$$

where $(\bar{y}_1^{(b)} - \bar{y}_2^{(b)})$ is the sample mean difference applied to the two bth random subvectors for the cases and the control, as the training sets, and $\hat{\Delta}_{\mathrm{SQ},\lambda}^{(-b)}$ is the smooth quantile ratio estimate obtained from the rest of the data. We choose $B = 10$ and we minimise $\mathrm{CV}(\lambda)$ for $\lambda = 1, 2, 4, 6, 8$.

For the simulations, data are generated under five scenarios, A–E. Under each scenario, we compare bias and variance properties of the following six estimators of Δ: $\hat{\Delta}_{\mathrm{SQ}}(\hat{\lambda})$, where $\hat{\lambda}$ is estimated by minimising $\mathrm{CV}(\lambda)$ in equation (9); $\hat{\Delta}_{\mathrm{SQ}}(\lambda = 2)$; $\hat{\Delta}_{\mathrm{SQ}}(\lambda = 4)$; the smooth quantile ratio etimator under the assumption that the two populations are log-normal $\hat{\Delta}_{\mathrm{SQ}}(\mathrm{LN}, 1)$; the maximum likelihood estimator under the log-normal model LN; and the sample mean difference $\bar{y}_1 - \bar{y}_2$. Table 1 and Fig. 3 summarise the five scenarios studied. In scenarios A, B and C, the population-2 distribution is log-normal with normal mean $v_2 = 7$ and normal standard error $\sigma_2 = 1{\cdot}5$. These parameters were chosen to approximate roughly the sample statistics from the medical expenditures datasets for non-diseased subjects. In scenario A, population 1 is also log-normal, with larger parameter values $v_1 = 7{\cdot}5$ and $\sigma_1 = 1{\cdot}75$. In scenarios B and C, population 1 differs from population 2 by the functions $s(p)$ shown in Fig. 3, chosen to represent a range of plausible shapes. We next studied $\hat{\Delta}_{\mathrm{SQ}}(\lambda)$'s performance for the real-data application. Scenario D, the log quantile functions of which are pictured in Fig. 3(d) with a dark solid line, contrasts the distributions of nonzero medicare expenditures for cases and controls. In scenario E, we assume that both populations have Gamma distributions with finite second moments.

Table 1. *Description of the sampling mechanisms used under each simulation study scenario. In scenario* D, $\hat{F}_g$ $(g = 1, 2)$ *are the empirical cumulative distribution functions of the nonzero Medicare expenditures for patients in the case and control groups, and, in scenarios* B *and* C, $g(y) = \exp\{7 + \Phi^{-1}(y)1{\cdot}5\}$

Scenario	Population 1	Population 2	n_1	n_2
A	LN(7·5, 1·75)	LN(7, 1·5)	100	1000
B	$u \sim \mathrm{Un}(0, 1]$, $y_1 = g(u)e^{s_B(u)}$	LN(7, 1·5)	100	1000
C	$u \sim \mathrm{Un}(0, 1]$, $y_1 = g(u)e^{s_C(u)}$	LN(7, 1·5)	100	1000
D	$\hat{F}_1$	$y_2 \sim \hat{F}_2$	100	1000
E	$\mathrm{Ga}(2{\cdot}5, 2{\cdot}5/\bar{y}_1)$	$\mathrm{Ga}(2{\cdot}5, 2{\cdot}5/\bar{y}_2)$	100	1000

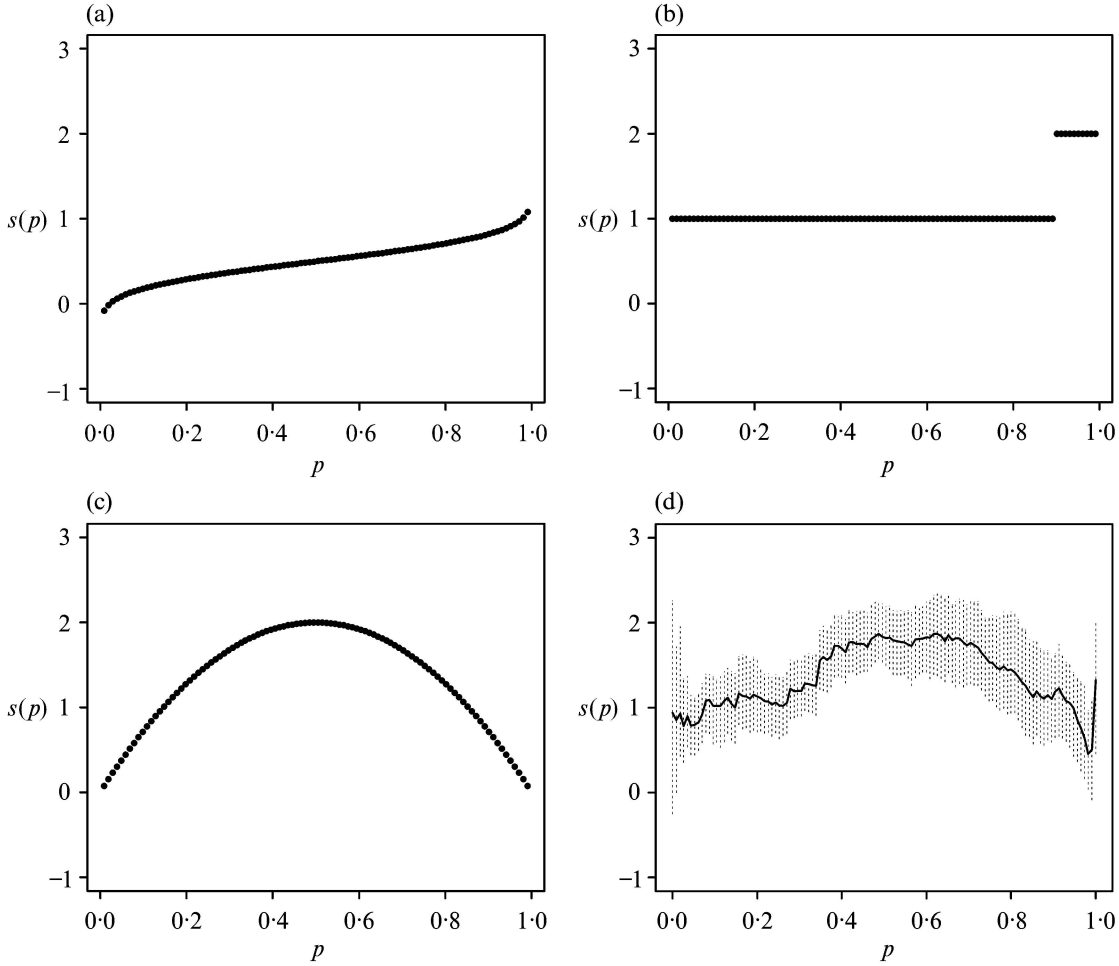

Fig. 3. Theoretical, (a)–(c), and empirical (d), $s(p)$ curves. In (d), the solid curve is $\log(y_{1(i)}/q_{2(i)})$ plotted at the percentiles $p_{1i} = i/(n_1 + 1)$, for $i = 1, \ldots, n_1$, where $q_{2(1)}, \ldots, q_{2(n_1)}$ are the order statistics of the $y_{21}, \ldots, y_{2n_2}$ interpolated at percentiles p_{1i}. The vertical segments represent 95% pointwise bootstrap confidence intervals.

We generated 1000 datasets for each scenario, and we compared estimators with equal sample sizes $n_1 = n_2 = 100$ and unequal samples with $n_1 = 100$, $n_2 = 1000$. The results were qualitatively similar and hence we report only the unequal case. For each dataset, we implement our method with $\lambda = 2$, $\lambda = 4$ and $\lambda = \hat{\lambda}$ estimated by crossvalidation. In all cases natural cubic splines are used as basis functions. These results show that $\hat{\Delta}_{SQ}$ has a smaller mean squared error than either $\bar{y}_1 - \bar{y}_2$ or the log-normal estimators. Table 2 presents the relative mean squared error, $\{\text{MSE}(\bar{y}_1 - \bar{y}_2) - \text{MSE}(\hat{\Delta})\}/\text{MSE}(\bar{y}_1 - \bar{y}_2)$, and the relative bias, $\{E(\hat{\Delta}) - \Delta\}/\Delta$, as percentages. Negative values for the relative mean squared error imply that $\bar{y}_1 - \bar{y}_2$ is preferred, and positive ones favour $\hat{\Delta}_{SQ}$.

Table 2: *Simulation study. Mean squared error relative to* $\bar{y}_1 - \bar{y}_2$ *defined by* $[\{\text{MSE}(\bar{y}_1 - \bar{y}_2) - \text{MSE}(\hat{\Delta})\}/\text{MSE}(\bar{y}_1 - \bar{y}_2)] \times 100$, RMSE, *and percentage bias relative to* $\bar{y}_1 - \bar{y}_2$ *defined by* $[\{E(\hat{\Delta}) - \Delta\}/\Delta] \times 100$, RB, *under the data generation mechanisms described in* § 3. *The degrees of freedom* λ *are estimated by the crossvalidation approach illustrated in equation* (9) *for* $B = 10$

	Scenario A		Scenario B		Scenario C		Scenario D		Scenario E	
$\hat{\Delta}$	RMSE	RB	RMSE	RB	RMSE	RB	RMSE	RB	RMSE	RB
$\hat{\Delta}_{SQ}(\hat{\lambda})$	58	−3	17	−6	8	11	14	4	0	0
$\hat{\Delta}_{SQ}(2)$	66	−9	28	−17	32	10	24	6	0	0
$\hat{\Delta}_{SQ}(4)$	59	−3	18	−5	21	10	20	2	0	1
$\hat{\Delta}_{SQ}(\text{LN}, 1)$	66	−3	29	−23	−436	47	−201	45	−1	1
LN	65	7	22	−27	−3725	131	−1393	112	−28	4
$\text{MSE}(\bar{y}_1 - \bar{y}_2)$	17827		51555		1390		6090		500	
Δ	4982		15225		5244		7144		7144	

In scenario A, when both populations are log-normal, $\hat{\Delta}_{SQ}(\hat{\lambda})$ and $\hat{\Delta}_{SQ}(\lambda)$ for $\lambda = 2$ and $\lambda = 4$ are approximately 60% better than $\bar{y}_1 - \bar{y}_2$. Note that the smooth quantile ratio estimates perform better even than the log-normal maximum likelihood estimate, which in this case is asymptotically efficient.

In scenario B, the five estimators have comparable performance and they are all superior to the sample mean difference. In scenarios C and D, both $\hat{\Delta}_{SQ}(\text{LN}, 1)$ and LN perform very poorly because of the substantial nonlinearity of $s(p)$ whereas the smooth quantile ratio estimators are 20% to 30% better than the sample mean difference. Finally, for the empirical scenario D, the smooth quantile ratio estimators are 10% to 20% better than the sample mean difference, and again the log-normal estimator performs very poorly. In scenario E, the maximum likelihood estimator of Δ is the sample mean difference, the smooth quantile ratio estimator's performance is similar to that of the maximum likelihood estimator, and much better than that of the log-normal maximum likelihood estimator.

Table 2 also displays relative biases, showing that $\hat{\Delta}_{SQ}(\lambda)$ has small biases in the cases considered. As expected, the bias of $\hat{\Delta}_{SQ}(\text{LN}, 1)$ is small only when $s(p)$ is almost constant. Finally, except in scenario A when the two populations are log-normal, the LN estimator is badly biased.

We also varied the choice of the basis functions. For each of the datasets, and for each scenario, we estimated $s(p)$ using natural cubic splines, smoothing splines and polynomials. These estimates are all quite close to each other and to the true $s(p)$, results not shown.

We compared the asymptotic variance in equation (8) with the sample variance under scenario A for several sample-size specifications. Results shown in Table 3 indicate good

Table 3. *Estimates of the asymptotic variance* σ^2 *in equation* (8) *and of the sample variance,* $\hat{\sigma}^2$, *when* $y_1 \sim \text{LN}(7{\cdot}5, 1{\cdot}75)$ *and* $y_2 \sim \text{LN}(7, 1{\cdot}5)$ *corresponding to scenario* A *with* $n_2/n_1 = 10$

	$n_1 + n_2$						
	10	50	100	1000	2000	2500	10000
$\hat{\sigma}^2$	44826·40	15189·20	10127·80	2811·50	1945·10	1745·70	856·60
σ^2	19294·00	12203·20	8628·90	2728·70	1929·50	1725·80	862·90

agreement between the two variance estimators. L. Cope's Ph.D thesis shows that the results for the uniform, log-normal and Pareto samples also display good agreement between the asymptotic and the sample variance for moderate to large samples.

Finally, we estimate the mean difference between annual Medicare expenditures for cases and controls in the real data. Of course not all of those with lung cancer and chronic obstructive pulmonary disease are smokers. Since medical expenditures may be different for smokers and nonsmokers, we therefore partition the subjects according to smoking status and analyse them separately, as well as estimating the overall mean difference. Figure 4 shows boxplots of 500 bootstrap estimates of Δ for all subjects and for the smokers alone. Estimators used are those compared in the simulation study. As in the simulation results, the smooth quantile ratio estimator is far more efficient than the selected competitors. In addition, the nonlinearity of the estimated $s(p_i, \beta)$, see Fig. 3(d), also suggests that the maximum likelihood estimator, LN, is likely to be biased. Estimates for the smokers are slightly larger than for everyone.

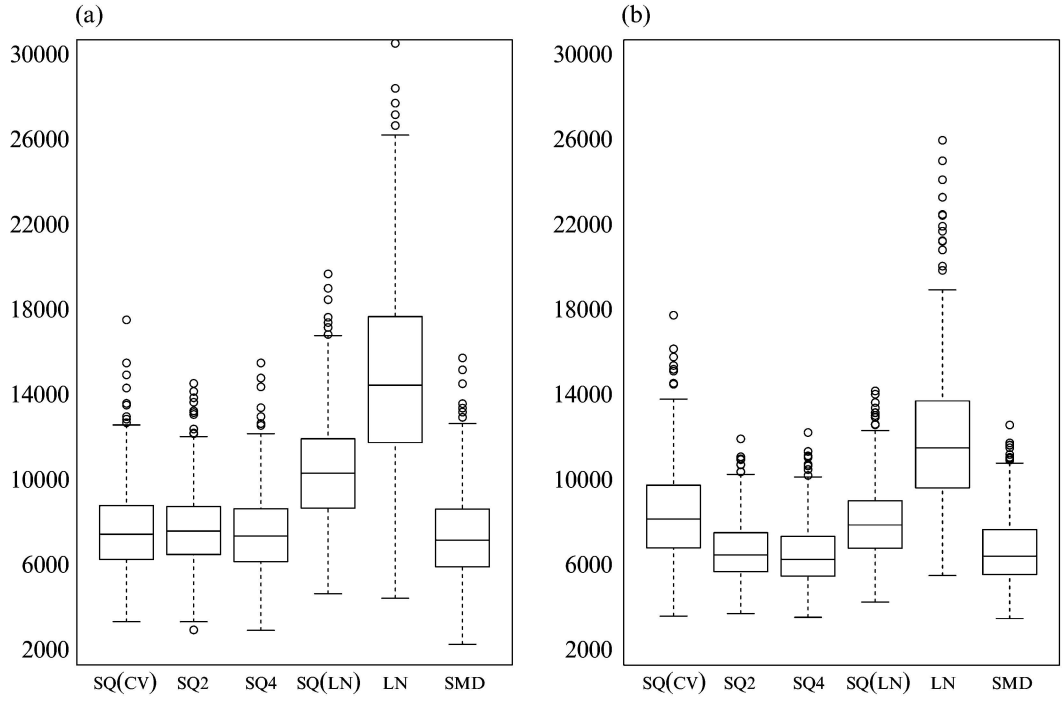

Fig. 4. Boxplots of 500 bootstrap samples of the estimated mean differences $\hat{\Delta}$ of Medicare expenditures for people with and without smoking-attributable diseases. Results are reported (a) for everyone in the sample ($n_1 = 118$, $n_2 = 2262$) and (b) for smokers only ($n_1 = 112$, $n_2 = 980$). The labels denote the smooth quantile ratio estimates defined in Table 2 and SMD denotes the sample mean difference.

5. Discussion

The software for implementing the method, the data for reproducing all the analyses reported in this paper, details of the asymptotic properties and extensions to the regression case are all available at http://biostat.jhsph.edu/~fdominic/square.html.

The idea of linking two samples in a semiparametric model is obviously not new. Perhaps the most famous and influential example is the Cox proportional hazards model (Cox, 1972) where the target is the hazard ratio. A second example is the density ratio model of Qin & Zhang (1997). Here, the ratio of densities $f(x)/g(x)$ is assumed to be a smooth function of x. This model would lead to an estimator of the mean difference that is analogous to ours but where a smooth function of the unordered data is used in equation (7) rather than a smooth function of the order statistics.

We have investigated how to generalise our theoretical results to the case where λ is unknown and needs to be estimated using crossvalidation. It can be shown that, with probability converging to 1, generalised crossvalidation selects a large enough number of basis functions and consistency is retained. However, in this same situation we are not able to prove asymptotic normality and simulation studies suggest that asymptotic normally fails to hold.

As an alternative to our method, we could assume that $Q_{\log Y_1}(p) = s\{Q_{\log Y_2}(p), \lambda\}$, and estimate Δ by using the fitted values of the Q–Q plot. We included this estimator in our simulation study, but found it not to be as efficient as smooth quantile ratio estimation.

Our analysis of medical expenditures allows smoking status to modify the effect of disease on expenditures. We examine this modification effect by first stratifying the cases and the controls with respect to their smoking status, and then applying our methods separately to smokers and nonsmokers, within each group. A more desirable goal would be to compare medical expenditures for cases and controls taking into account individual-level characteristics x. In this case smooth quantile ratio estimation can be extended to the regression case by assuming that

$$\log Q_1(p; x) = \log Q_2(p; x) + s(p; x). \tag{10}$$

To control for systematic differences in covariates between two populations, a common strategy is to group units into subclasses based on covariate values, for example using propensity score matching (Cochran & Rubin, 1973; Rubin, 1973), and then to apply our method within strata of propensity scores. The extension of smooth quantile regression estimation to the regression case and a comparison with common econometric models such as two-part log-linear regression models (Duan, 1983) are described in a technical report by F. Dominici and S. Zeger, available at http://www.bepress.com/jhubiostat/paper16/.

In clinical trials our approach can be used to estimate treatment effects that vary smoothly with respect to the percentiles of the health outcome. If Y has a more nearly symmetric distribution, rather than smoothing the log ratio of the quantiles, we can smooth their difference; that is, we can assume that $Q_1(p) - Q_2(p) = s(p)$. Under this model, we estimate the treatment effect, Δ, by $\int s(p)dp$. The plot of the estimated $s(p)$ versus p is also informative for identifying the outcome percentiles where the treatment is mostly effective.

Acknowledgement

Funding for Scott L. Zeger, Francesca Dominici, and David Q. Naiman was provided, respectively by grants from the National Institute of Mental Health, the National Institute of Environmental Health Science and the National Science Foundation. We thank

Timothy Wyant for providing data on the National Medical Expenditures Survey, Mark van der Laan, Giovanni Parmigiani, Michael Griswold, Nathaniel D. Mercaldo and Tom Louis for comments and suggestions on the paper, and Elizabeth Johnson for assistance in database development and software.

Appendix

Sketch of proof of Theorem 2

To show that $\hat{\Delta} - \Delta$ asymptotically has a normal distribution we use the von Mises functional δ-method (Serfling, 1980, Ch. 6). To implement this method, we first establish the asymptotic equivalence between the smooth quantile ratio estimator and the functional $T(F_1, F_2)$ defined below. The proof of asymptotic equivalence is detailed in L. Cope's thesis. Secondly, we expand the functional in a one-term Taylor series. The first derivative of the functional at the point (F_1, F_2) converges to a Gaussian distribution. If the Taylor remainder term converges in probability to zero, then the estimator, like the derivative, has a Gaussian limiting distribution. The necessary assumptions, bounding conditions on components of the functional, are very similar to those required to prove that L-estimators are asymptotically normal.

Our functional takes the form

$$T(F_1, F_2) = \frac{1}{2}\int_0^1 F_1^{-1}(p)[1 - \exp\{-s(p, \beta)\}]dp + \frac{1}{2}\int_0^1 F_2^{-1}(p)[\exp\{s(p, \beta) - 1\}]dp,$$

where

$$s(p, \beta) = \sum_{j=0}^{\lambda} \beta_j B_j(p), \quad \beta_j = \int_0^1 B_j(p)[\log\{F_1^{-1}(p)\} - \log\{F_2^{-1}(p)\}]dp,$$

so that the functional version of the estimator is given by $T(\hat{F}_1, \hat{F}_2)$.

In using the differentiable statistical function approach, the largest task is to demonstrate that the remainder,

$$R_1 = \sqrt{n}[T(\hat{F}) - T(F) - d_1\{T, F; \sqrt{n}(\hat{F} - F)\}],$$

converges to zero in distribution, because then $\sqrt{n}\{T(\hat{F}) - T(F)\}$ is equivalent to $d_1\{T, F; \sqrt{n}(\hat{F} - F)\}$ and the asymptotic properties of the former can be derived from the latter. In the case of smooth quantile ratio estimation,

$$d_1\{T, F; \sqrt{n}(\hat{F} - F)\} = \frac{\sqrt{n}}{2}\int_0^1 [F_1\{\hat{F}_1^{-1}(p)\} - p]\eta_1(p)dp + \frac{\sqrt{n}}{2}\int_0^1 [F_2\{\hat{F}_2^{-1}(p)\} - p]\eta_2(p)dp.$$

Both $\sqrt{n}[F_1\{\hat{F}_1^{-1}(p)\} - p]$ and $\sqrt{n}[F_2\{\hat{F}_2^{-1}(p)\} - p]$ converge to Brownian bridges, so the derivative asymptotically has a normal distribution with variance σ^2 as defined above.

Sketch of proof that the remainder converges to zero. At points in this proof it is necessary to evaluate expressions like $\int_{p=0}^1 \hat{F}^{-1}(p)J_n(p)dp$, where $\hat{F}^{-1}$ is an empirical quantile function and $J_n(p)$ may also be data-dependent. In order to simplify treatment of these expressions, the following lemma establishes conditions under which the range of integration can be truncated.

Lemma A1. *Let $x_1, x_2, \ldots, x_n$ be a random sample. Suppose that $\hat{F}^{-1}$ is the empirical quantile function corresponding to these data, and let $J_n:(0, 1) \to \Re$ be a possibly random function. Assume that there exist positive constants M, b and δ such that*

(i) *the quantile function $F^{-1}(p) \leqslant M\{p(1-p)\}^{-b+\delta}$, and*

(ii) *the random function $|J_n(x)| \leqslant [M\{p(1-p)\}^{-1/2+b}]^{1+\varepsilon_n}$, where $\varepsilon_n \to 0$ in probability.*

Then, in probability,

$$T_n = \sqrt{n}\left\{\int_0^{k/n} \hat{F}^{-1}(p)J_n(p)dp + \int_{(n-k)/n}^{1} \hat{F}^{-1}J_n(p)dp\right\} \to 0.$$

The proof is not included here.

We break the remainder up into several pieces and prove convergence separately for each piece. Here R_1 can be written as

$$R_1 = R_{11} + R_{12} + R_{13} + R_{21} + R_{22} + R_{23},$$

where

$$R_{11} = \frac{\sqrt{n}}{2}\int_0^1 \left(\hat{F}_1^{-1}(p) - F_1^{-1}(p) - \frac{[\hat{F}_1\{F_1^{-1}(p)\} - p]}{f_1\{F_1^{-1}(p)\}}\right)[1 - \exp\{-s(p,\beta)\}]dp,$$

$$R_{12} = -\int_0^1 \sqrt{n}F_1^{-1}(q)\left\{\frac{\exp\{-s(q,\hat{\beta})\} - \exp\{-s(q,\beta)\}}{2} - \exp\{-s(q,\beta)\}\right.$$
$$\left.\times \sum_{i=0}^{\lambda} B_i(q)\int_0^1 B_i(p)\left(\frac{[\hat{F}_1\{F_1^{-1}(p)\} - p]}{2F_1^{-1}(p)f_1\{F_1^{-1}(p)\}} - \frac{[\hat{F}_2\{F_2^{-1}(p)\} - p]}{2F_2^{-1}(p)f_2\{F_2^{-1}(p)\}}\right)\right\} dp\,dq,$$

$$R_{13} = \frac{\sqrt{n}}{2}\int_0^1 [\exp\{s(p,\hat{\beta})\} - \exp\{s(p,\beta)\}]\{\hat{F}_2^{-1}(p) - F_2^{-1}(p)\}dp.$$

As a result of the symmetry of the functional, the other remainder terms, R_{21}, R_{22} and R_{23}, are identical in form to these expressions and are treated in the same fashion.

The first term, R_{11}, is the remainder from a function δ-method approach to L-statistics. As such, this is easily shown to converge in probability to zero. After a little bit of algebra, R_{12} can likewise largely be expressed in terms of L-statistic remainders and demonstrated to converge in probability to zero. With some manipulation, R_{12} can be written as

$$R_{12} = \frac{\sqrt{n}}{2}\int_0^1 \left(\log \hat{F}_1^{-1} - \log F_1^{-1} - \frac{[\hat{F}_1\{F_1^{-1}(p)\} - p]}{F_1^{-1}(p)f_1\{F_1^{-1}(p)\}}\right)\sum_{j=1}^{\lambda} B_j(p)\int_0^1 F_2^{-1}B_j(q)dq \qquad \text{(A1)}$$

$$+ \frac{\sqrt{n}}{2}\int_0^1 \left(\log \hat{F}_2^{-1} - \log F_2^{-1} - \frac{[\hat{F}_2\{F_2^{-1}(p)\} - p]}{F_2^{-1}(p)f_2\{F_2^{-1}(p)\}}\right)\sum_{j=1}^{\lambda} B_j(p)\int_0^1 F_2^{-1}B_j(q)dq \qquad \text{(A2)}$$

$$+ \frac{\sqrt{n}}{2}\int_0^1 F_2^{-1}(p)\exp\left\{-\sum_j \xi_j B_j(p)\right\}\sum_j \{(\hat{\beta}_j - \beta_j)B_j(p)\}^2 dp. \qquad \text{(A3)}$$

Each of the first two terms, (A1) and (A2), is the remainder from the differentiable statistical functional form of an L-statistic with functional, and so converges in probability to zero.

To deal with (A3), we have that

$$\exp\left\{-\sum_j \xi_j B_j(p)\right\} \leqslant \exp\left\{\left|\sum_j \beta_j B_j(p)\right|\right\}^{1+\Gamma_n},$$

and thus

$$\frac{\sqrt{n}}{2}\int_0^1 F_2^{-1}(p)\exp\left\{-\sum_j \xi_j B_j(p)\right\}\left\{\sum_j (\hat{\beta}_j - \beta_j)B_j(p)\right\}^2 dp$$
$$\leqslant n^{1/2}M_2^{1+\Gamma_n}\max_i (\hat{\beta}_i - \beta_i)^2 \int_0^1 [\{p(1-p)\}^{-1/2+\delta^2/(\delta+2)}]^{1+\Gamma_n}dp. \qquad \text{(A4)}$$

If we set $\hat{F}^{-1}(p) \equiv 1$, Lemma A1 can be applied so that (A4) is bounded from above by

$$M_3 \max_i (\hat{\beta}_i - \beta_i)^2 n^{1/2 + \{1/2 - \delta^2/(\delta+2)\}(1+\Gamma_n)}.$$

When n is sufficiently large, and therefore Γ_n is sufficiently small, the exponent on n is less than 1, ensuring convergence in probability. This completes the treatment of R_{12}, and the final term R_{13} is handled in a similar fashion.

References

Aitchison, J. & Shen, S. M. (1980). Logistic normal distributions: Some properties and uses. *Biometrika* **67**, 261–71.

Cochran, W. G. & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā* A **35**, 17–46.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Doksum, K. A. & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63**, 421–34.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *J. Am. Statist. Assoc.* **78**, 605–10.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Fenn, P., McGuire, A., Backhouse, M. & Jones, D. (1996). Modelling programme costs in economic evaluation. *J. Health Econ.* **15**, 115–25.

Hlatky, M. A., Rogers, W. J. & Johnstone, I. et al. (1997). Medical care costs and quality of life after randomization to coronary angioplasty and coronary bypass surgery. *New Engl. J. Med.* **336**, 92–9.

Huber, P. J. (1996). *Robust Statistical Procedures*, 2nd ed., CBMS-NSF Regional Conference Series in Applied Mathematics, Number 68. Philadelphia, PA: Soc. Industr. Appl. Math.

Lin, D. (2000). Linear regression analysis of censored medical costs. *Biostatistics* **1**, 35–47.

Lin, D. Y., Feuer, E. J., Etzioni, R. & Wax, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53**, 419–34.

Lipscomb, J., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Samsa, G. P. & Matchar, D. B. (1999). Predicting the cost of illness: A comparison of alternative models applied to stroke. *Med. Decis. Making* **18**, S39–S56.

National Center for Health Services Research (1987). *National Medical Expenditure Survey. Methods II. Questionnaires and data collection methods for the household survey and the Survey of American Indians and Alaska Natives.* Rockville, MD: National Center for Health Services Research and Health Technology Assessment.

O'Brien, P. C. (1988). Comparing two samples: Extensions of the *t*, rank-sum, and log-rank tests. *J. Am. Statist. Assoc.* **83**, 52–61.

Qin, J. & Zhang, A. (1997). A goodness of fit test for the logistic regression model based on case-control data. *Biometrika* **84**, 609–18.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29**, 185–203.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Shorack, G. (1972). Functions of order statistics. *Ann. Math. Statist.* **43**, 412–27.

Tu, W. & Zhou, X.-H. (1999). A Wald test comparing medical cost based on log-normal distributions with zero valued costs. *Statist. Med.* **18**, 2749–61.

Wellner, J. (1977). A Glivenko-Cantelli theorem and strong laws of large numbers for functions of order statistics. *Ann. Statist.* **5**, 473–80.

Zellner, A. (1971). Bayesian and non-Bayesian analysis of the log-normal distribution and log-normal regression. *J. Am. Statist. Assoc.* **66**, 327–30.

Zhou, X.-H. & Gao, S. (1997). Confidence intervals for the log-normal mean. *Statist. Med.* **16**, 783–90.

Zhou, X.-H., Gao, S. & Hui, S. L. (1997). Methods for comparing the means of two independent log-normal samples. *Biometrics* **53**, 1129–35.

[*Received November* 2003. *Revised January* 2005]