*Proc. Natl. Acad. Sci. USA*
Vol. 85, pp. 3670–3672, June 1988
Statistics

# Estimating a treatment effect under biased sampling

**(conditional maximum likelihood/u,v method/maximum-likelihood-type estimators/clinical trials)**

HERBERT ROBBINS AND CUN-HUI ZHANG

Institute of Biostatistics and Department of Statistics, Rutgers University, New Brunswick, NJ 08903

*Contributed by Herbert Robbins, February 8, 1988*

**ABSTRACT Methods are presented for the estimation of a treatment effect based on before- and after-treatment values, where for ethical reasons all and only those patients are treated whose before-treatment values exceed a given constant.**

Can one reliably estimate the effect of a treatment when there is no comparable untreated control group? For example, assume that the before-treatment $X_i$ and the after-treatment $Z_i$ of the $i$th person in a group of $n$ are independent normal random variables with respective means $\theta_i$ and $\lambda + \theta_i$ and common variance 1, the $\theta_i$ and $\lambda$ being unknown. If all $n$ persons were treated we could estimate the treatment effect $\lambda$ by the average of the $n$ differences $Z_i - X_i$. Assume, however, that for ethical reasons *all and only* those persons with $X_i > a$ are treated, where $a$ is a preassigned constant, so that the value $Z_i$ is available if and only if $X_i > a$. The average value of $Z_i - X_i$ over all persons with $X_i > a$ will clearly be a downwardly biased estimator of $\lambda$, and something better is needed.

A more general problem of this nature can be modeled as follows. Let $f(x; \lambda, \theta)$ be a family of probability density functions with respect to a $\sigma$-finite measure $\mu(dx)$. Let $X_i, Z_i, 1 \leq i \leq n$, be independent random variables such that

$$X_i \sim f(x; c, \theta_i),\ Z_i \sim f(z; \lambda, \theta_i),$$

where $c$ is a known constant and $\lambda$ and the $\theta_i$ are unknown parameters. Assume that $Z_i$ is observed if and only if $X_i$ is in a set $A$, but that $X_i$ is observed for each $i$. We are interested in estimating $\lambda$ on the basis of observed values of $X_i, Y_i, 1 \leq i \leq n$, where

$$Y_i = \delta(X_i)Z_i,\ \delta(x) = I\{x \in A\} = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ otherwise.} \end{cases}$$

(In the preceding paragraph, $f(x; \lambda, \theta) = \varphi(x - \lambda - \theta)$ and $c = 0$, where $\varphi$ is the standard normal density function.)

This estimation problem was considered by Robbins (1), who proposed a "u,v" method for estimating $\lambda$. In this note we consider a conditional maximum likelihood method for estimating $\lambda$ and compare it with the u,v estimator in two special cases.

**The Conditional Maximum Likelihood Estimator (CMLE).** The random vector $(X_i, Y_i)$ has the joint density function

$$f(x, y; \lambda, \theta_i) = f(x; c, \theta_i)[f(y; \lambda, \theta_i)]^{\delta(x)}$$

with respect to the measure $\nu(dx, dy) = \mu(dx)[\mu(dy)]^{\delta(x)}$. To save notation let $(X, Z)$ be a random vector such that $(X, Z) \sim f(x; c, \theta)f(z; \lambda, \theta)$ and let $Y = \delta(X)Z$. Let $\psi(x, y, \lambda)$ be a function such that

$$\int \psi(x, y, \lambda)f(x, y; \lambda, \theta)\nu(dx, dy) = E\psi(X, Y, \lambda) = 0,\ \forall\lambda, \theta. \quad [1]$$

A solution $\hat{\lambda}_n(\psi)$ of the equation $\sum_{i=1}^n \psi(X_i, Y_i, \lambda) = 0$ is called a maximum-likelihood-type estimator (M-estimator) corresponding to the influence function $\psi$. Under certain regularity conditions on $\psi$ and the sequence $\theta_i$, the Taylor expansion

$$0 = \sum_{i=1}^n \psi(X_i, Y_i, \hat{\lambda}_n(\psi)) = \sum_{i=1}^n \psi(X_i, Y_i, \lambda) + (1 + o(1))(\hat{\lambda}_n(\psi) - \lambda) \sum_{i=1}^n \psi'(X_i, Y_i, \lambda)$$

and the law of large numbers and central limit theorem will hold, so that as $n \to \infty$

$$(\hat{\lambda}_n(\psi) - \lambda)/\sigma_n(\psi) \to N(0, 1),$$

where

$$\sigma_n^2(\psi) = \frac{\sum_{i=1}^n E[\psi(X_i, Y_i, \lambda)]^2}{[\sum_{i=1}^n E\psi'(X_i, Y_i, \lambda)]^2}. \quad [2]$$

See, for example, Huber (2) for details.

Andersen (3) proposed a method of finding suitable influence functions (and therefore M-estimators) for mixture models, and his method can be applied to our case as follows. Suppose there are functions $s = s(x, y, \lambda)$, $t(s, \theta)$, $g = g(x, y, \lambda)$ and $h(s, \lambda)$ such that

$$f = f(x, y; \lambda, \theta) = \exp[t(s, \theta) + g],\ s' = \partial s/\partial\lambda = h(s, \lambda). \quad [3]$$

Define

$$\rho(X, Y, \lambda) = \partial \log f/\partial\lambda - E[\partial \log f/\partial\lambda | s] = g'(X, Y, \lambda) - E[g'(X, Y, \lambda)|s(X, Y, \lambda)], \quad [4]$$

where $g' = \partial g/\partial\lambda$. The function $\rho$ does not depend on the value of $\theta$, and

$$\int \psi(x, y, \lambda)f(x, y; \lambda, \theta)\nu(dx, dy) = E\rho(X, Y, \lambda) = 0,\ \forall\lambda, \theta. \quad [5]$$

Andersen (3) called the M-estimator $\hat{\lambda}_n(\rho)$ the CMLE. If we can exchange the order of differentiation $\partial/\partial\lambda$ and integration on the left-hand side of Eq. 5, then $E\rho' = -E\rho^2$ and

$$\sigma_n^2(\rho) = \left[\sum_{i=1}^n I(\theta_i)\right]^{-1}, \qquad I(\theta) = E\rho^2(X, Y, \lambda),$$

which implies by the Schwarz inequality that

$$E[\psi(X, Y, \lambda)|s(X, Y, \lambda)] = 0 \text{ almost surely} \Rightarrow \sigma_n^2(\rho) \leq \sigma_n^2(\psi), \quad [6]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: CMLE, conditional maximum likelihood estimator; M-estimator, maximum-likelihood-type estimator.

provided that the differentiation $\partial/\partial\lambda$ of the left-hand side of Eq. **1** can be performed under the integral sign. Lindsay (4) proved that the CMLE is in a sense asymptotically efficient when the left-hand side equation in expression **6** is equivalent to Eq. **1** [the "completeness" of the "sufficient statistic" $s(X, Y, \lambda)$] under a mixture setting.

**The u,v Method.** Let $u(x)$ and $v(x)$ be two functions such that $u(x) = 0$ on $A$ and

$$\theta \int u(x)f(x; c, \theta)\mu(dx) = \int v(x)f(x; c, \theta)\mu(dx), \ \forall\theta. \quad [7]$$

Robbins (1) used $u$ and $v$ to derive estimating equations for $\lambda$ in the following two cases:

(*i*) Suppose that $EZ = \lambda + \theta$. Then $Eu(X)Y = Eu(X)Z = \lambda Eu(X) + Ev(X)$. The M-estimator

$$\hat{\lambda}_n(\psi_{u,v}) = \left[\sum_{i=1}^n u(X_i)\right]^{-1} \left[\sum_{i=1}^n (u(X_i)Y_i - v(X_i))\right] \quad [8]$$

corresponding to the influence function $\psi_{u,v}(x, y, \lambda) = u(x) \cdot (y - \lambda) - v(x)$ is called the u,v estimator, and

$$\sigma_n^2(\psi_{u,v}) = \frac{\sum_{i=1}^n E[u(X_i)(Y_i - \lambda) - v(X_i)]^2}{[\sum_{i=1}^n Eu(X_i)]^2}. \quad [9]$$

(*ii*) Suppose that $EZ = \lambda\theta$. Then $Eu(X)Y = \lambda Ev(X)$. The u,v influence function is $\psi_{u,v}(x, y, \lambda) = u(x)y - \lambda v(x)$, the u,v estimator is

$$\hat{\lambda}_n(\psi_{u,v}) = \left[\sum_{i=1}^n v(X_i)\right]^{-1} \left[\sum_{i=1}^n u(X_i)Y_i\right], \quad [10]$$

and

$$\sigma_n^2(\psi_{u,v}) = \frac{\sum_{i=1}^n E[u(X_i)Y_i - \lambda v(X_i)]^2}{[\sum_{i=1}^n Ev(X_i)]^2}. \quad [11]$$

**The Normal Case.** Assume as in the first paragraph of the introduction that

$$f(x; \lambda, \theta) = \varphi(x - \lambda - \theta), \ c = 0, \ A = [0, \infty), \quad [12]$$

so that $\delta(x) = I\{x \geq 0\}$ and

$$f(x, y; \lambda, \theta) = (2\pi)^{-(1+\delta(x))/2}\exp[-(x - \theta)^2/2 - \delta(x)(y - \lambda - \theta)^2/2].$$

The condition **3** holds with $s = (s_1, s_2) = (\delta(x), x + \delta(x) \cdot (y - \lambda))$, $s' = (0, -s_1) = h(s)$, and $g'(x, y, \lambda) = \delta(x)(y - \lambda)$. Since $(X, Z - \lambda)|X + Z - \lambda = s_2 \sim (W, s_2 - W)$ with $W \sim N(s_2/2, 1/2)$,

$$\begin{aligned} E[g'|s] &= \delta(X)E[Z - \lambda|\delta(X) = 1, X + Z - \lambda = s_2] \\ &= \delta(X)E(s_2 - W)I\{W \geq 0\}/P\{W \geq 0\} \\ &= \delta(X)s_2/2 - \delta(X)\varphi(s_2/\sqrt{2})/[\sqrt{2}\,\Phi(s_2/\sqrt{2})]. \end{aligned}$$

Therefore, by Eq. **4**

$$\rho(x, y, \lambda) = \delta(x)\left[(y - \lambda - x)/2 + \frac{\varphi((x + y - \lambda)/\sqrt{2})}{\sqrt{2}\,\Phi((x + y - \lambda)/\sqrt{2})}\right].$$

Since $\rho(x, y, \lambda)$ is nonincreasing in $\lambda$, the CMLE $\hat{\lambda}_n(\rho)$ is uniquely defined by

$$\sum_{i=1}^n \rho(X_i, Y_i, \hat{\lambda}_n(\rho)) = 0. \quad [13]$$

Let $u(x)$ be such that $u(x) = 0$ for $x < 0$. Then Eq. **7** holds if

$$v(x) = xu(x) - u'(x). \quad [14]$$

The u,v influence function is $\psi_{u,v} = u(x)(y - \lambda) - v(x)$ and the u,v estimator is

$$\hat{\lambda}_n(\psi_{u,v}) = \left[\sum_{i=1}^n u(X_i)\right]^{-1} \left[\sum_{i=1}^n (u(X_i)(Y_i - X_i) + u'(X_i))\right]. \quad [15]$$

Since $(X, Z - \lambda)|X + Z - \lambda = s_2 \sim (W, s_2 - W)$,

$$\begin{aligned} E[\psi_{u,v}|s] &= \delta(X)E[u(X)(Z - \lambda - X) + u'(X)|\delta(X) = 1, \\ &\qquad X + Z - \lambda = s_2] \\ &= \delta(X)E[u(W)(s_2 - 2W) + u'(W)]/P\{W \geq 0\} \\ &= \delta(X) \int [u(w)(s_2 - 2w) + u'(w)] \\ &\qquad \times \varphi((w - s_2/2)\sqrt{2})dw\sqrt{2}/P\{W \geq 0\} \\ &= 0 \text{ (integrating by parts),} \end{aligned}$$

which implies the left-hand side equation of expression **6**. Hence, $\sigma_n^2(\rho) \leq \sigma_n^2(\psi_{u,v})$.

On the other hand, the performance of the u,v estimator **15** is not so bad either. If we choose

$$u(x) = \delta(x)[\varphi(0) - \varphi(kx)]$$

for some $k \neq 0$, then it can be shown that there exists an $M < \infty$ such that

$$\sigma_n^2(\psi_{u,v}) < M\sigma_n^2(\rho), \ \forall\theta_1, \ldots, \theta_n.$$

**The Poisson Case.** Consider the Poisson family

$$f(x; \lambda, \theta) = e^{-\lambda\theta}(\lambda\xi)^x/x!, \ x = 0, 1, \ldots$$

and take $c = 1$. The condition **3** is satisfied with $s = (s_1, s_2) = (\lambda\delta(x), x + \delta(x)y)$ and $g' = \delta(x)y/\lambda$. Since $(X, Z)|X + Z = s_2 \sim (W, s_2 - W)$ with $W \sim$ binomial $b(s_2, 1/(1 + \lambda))$,

$$\begin{aligned} \lambda E[g'|s] &= \delta(X)E[Z|\delta(X) = 1, X + Z = s_2] \\ &= \delta(X)E(s_2 - W)I\{W \in A\}/P\{W \in A\} \\ &= \delta(X)[s_2 - \textstyle\sum_{w\in A}wb(w; s_2, 1/(1 + \lambda))/ \\ &\qquad \textstyle\sum_{w\in A}b(w; s_2, 1/(1 + \lambda))], \end{aligned}$$

where $b(k; n, p) = \binom{n}{k}p^k(1 - p)^{n-k}$. Therefore,

$$\rho(x, y, \lambda) = \lambda^{-1}\delta(x)\left[\frac{\sum_{w\in A}wb(w; x + y, 1/(1 + \lambda))}{\sum_{w\in A}b(w; x + y, 1/(1 + \lambda))} - x\right]. \quad [16]$$

The u,v relationship **7** holds if

$$v(x) = xu(x - 1). \quad [17]$$

The u,v influence function is $\psi_{u,v} = u(x)y - \lambda xu(x - 1)$ and the u,v estimator is

$$\hat{\lambda}_n(\psi_{u,v}) = \left[\sum_{i=1}^n u(X_i - 1)X_i\right]^{-1} \left[\sum_{i=1}^n u(X_i)Y_i\right]. \quad [18]$$

Consider the following two cases:

(*i*) Let $A = \{x : x \geq a\}$ for some nonnegative integer $a$. Then $s(X, Y, \lambda)$ is a complete statistic for every fixed $\lambda$ and the left-hand side of expression **6** is equivalent to **1**, which implies that the CMLE is asymptotically efficient by Lindsay (4). Hence $\sigma_n^2(\rho) \leq \sigma_n^2(\psi)$, $\forall \psi$.

(*ii*) Let $A = \{x : x = a\}$ for some nonnegative integer $a$. Then by Eq. **16** $\rho = 0$ and the CMLE is not defined, while the u,v estimator is given by

$$\hat{\lambda}_n(\psi_{u,v}) = \frac{\sum_{i \leq n, X_i = a} Y_i}{C + (a+1)[\text{number of } i \leq n \text{ such that } X_i = a + 1]} \quad [19]$$

with $u(x) = \delta(x) = I\{x = a\}$, where $C$ is a positive constant. (The original u,v estimator **18** is modified here to avoid dividing by 0.) Incidentally, the asymptotic normality **2** holds for the u,v estimator **19** whenever $\Sigma_{i=1}^{\infty} P\{X_i = a\} = \infty$.

**Remarks.** In this section we consider the mixture model in which the nuisance parameters $\theta_i$ are treated as independent identically distributed (iid) random variables with some unknown distribution function $G$, so that $(X, Y), (X_1, Y_1), \ldots$ are iid random vectors with common density function

$$f(x, y) = f(x, y; \lambda, G) = \int f(x, y; \lambda, \theta) dG(\theta).$$

We shall study the mixture model in detail elsewhere. Here, to simplify the discussion, we consider the normal case **12** and use the notations of that section.

(*i*) The CMLE $\hat{\lambda}_n(\rho)$ is not efficient for this example. Define

$$\xi_w(s) = (1 - s_1) I\{w < s_2 < 0\} + s_1[1 - \Phi(-\sqrt{2}w + s_2/\sqrt{2})/\Phi(s_2/\sqrt{2})],$$

where $s = (s_1, s_2) = (\delta(x), x + \delta(x)(y - \lambda))$ and $w < 0$ is a constant. The function $\psi = \xi_w$ satisfies **1** and the influence function $\rho + \alpha\xi$ is more efficient than $\rho$, where $\alpha$ depends on $G$ but can be consistently estimated. In general, the efficient influence function can be written as

$$\psi^* = \rho + \xi^*, \ \xi^* = \xi^*(s_1, s_2), \ \xi^*(1, t) = -\Phi^{-1}(t/\sqrt{2})\sqrt{2} \int_{-\infty}^{0} \xi^*(0, x)\varphi(\sqrt{2}(x - t/2))dx, \quad [20]$$

where the choice of $\xi^*(0, x)$ depends on $G$.

(*ii*) We can also consider a doubly parametric model in which the distribution $G$ of the $\theta_i$ is assumed to be normal with unknown mean $\mu$ and variance $\sigma^2$. In this case $f(x, y; \lambda, G)$ is a smooth parametric family with parameters $\lambda$, $\mu$, and $\sigma^2$, and we can use the maximum likelihood estimator of $\lambda$. The efficient score function here has the form

$$\partial \log f/\partial\lambda + \alpha \partial \log f/\partial\mu + \beta \partial \log f/\partial\sigma, \quad [21]$$

where $\alpha$ and $\beta$ are constants, $\partial \log f/\partial\lambda = \delta(x)\{y - \lambda - E[\theta|s]\}$, $\partial \log f/\partial\mu = E[(\theta - \mu)/\sigma^2|s]$, and $\partial \log f/\partial\sigma = E[(\theta - \mu)^2/\sigma^3 - 1/\sigma|s]$. Noting that the conditional distribution of $\theta|s$ is normal with mean $[\mu + s_2\sigma^2]/[1 + (1 + s_1)\sigma^2]$ and variance $\sigma^2/[1 + (1 + s_1)\sigma^2]$, we find that the score function **21** does not have the form **20**. Therefore, the maximum likelihood (or any other efficient) estimator for this doubly parametric model does not have a stable performance when $G$ is not assumed to be normal or when the $\theta_i$ are unknown constants.

(*iii*) Both the CMLE and the u,v estimator remain the same under the "double truncation" case where $(\delta(X_i)X_i, Y_i)$ are observed instead of $(X_i, Y_i)$. For this case, the efficient influence function is

$$\psi^o = \rho + \alpha^o\xi^o, \ \xi^o = (1 - s_1) - s_1[\Phi^{-1}(s_2/\sqrt{2}) - 1].$$

Since $\int [\xi^o(s(x, y, \lambda))]^2 f(x, y; \lambda, \theta)\nu(dx, dy) < \infty$ if and only if $\theta > 0$, the CMLE is fully efficient in the double truncation case when $G(0) > 0$.

(*iv*) Let us replace the assumption $Z|\theta \sim N(\lambda + \theta, 1)$ by $E[Z|\theta] = \lambda + \theta$ and $\mathrm{Var}(Z|\theta) = 1$ (or $\leq C < \infty$). The u,v estimator still works but the CMLE does not. The naive estimator

$$\bar{\lambda}_n = \sum_{i=1}^{n} \delta(X_i)(Y_i - X_i)/\sum_{i=1}^{n} \delta(X_i)$$

is inconsistent, since

$$\bar{\lambda}_n \to \lambda - f(0)/[1 - F(0)],$$

where $f(x)$ and $F(x)$ are the marginal density and distribution functions of $X$, respectively. To correct $\bar{\lambda}_n$ for bias requires either a knowledge of $G$ or the use of a density estimator for $f(0)$ based on $X_1, \ldots, X_n$ under the singly parametric assumption. In the doubly parametric case $X$ is marginally a normal random variable, so that $\lambda$ can be estimated by the $n^{-1/2}$ consistent estimator

$$\tilde{\lambda}_n = \bar{\lambda}_n + \varphi(\bar{X}_n/S_n)/[S_n\Phi(\bar{X}_n/S_n)],$$

where $\bar{X}_n = n^{-1}\Sigma_{i=1}^{n} X_i$ and $S_n^2 = n^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X}_n)^2$.

(*v*) We may also consider the case where the treatment effect for the $i$th person is $\lambda_i$ instead of a common value $\lambda$, and we are interested in estimating the average treatment effect for treated persons

$$\left[\sum_{i=1}^{n} \delta(X_i)\right]^{-1} \left[\sum_{i=1}^{n} \delta(X_i)\lambda_i\right], \quad [22]$$

which is an unobservable random variable. The bias-correction method discussed in *iv* can be used without changing the estimator. The CMLE method does not work here. The u,v estimator can be used for the case where $E[Z_i|\theta_i] = \lambda_i + \theta_i$ and Eq. **7** holds for $u(x) = \delta(x)$ or for the case where $E[Z_i|\theta_i] = \lambda_i\theta_i$ and Eq. **7** holds for $v(x) = \delta(x)$. The u,v estimator remains unchanged in both cases. For the normal case, the u,v method does not apply since $u(x)$ has to be differentiable for Eq. **7** to hold.

This research was supported by the National Science Foundation and the Air Force Office of Scientific Research.


1. Robbins, H. (1988) in *Statistical Decision Theory and Related Topics IV*, eds. Gupta, S. S. & Berger, J. O. (Springer, New York), Vol. 1, pp. 265–270.
2. Huber, P. J. (1977) *Robust Statistical Procedures* (SIAM, Philadelphia).
3. Andersen, E. B. (1970) *J. R. Stat. Soc. B* **32**, 283–301.
4. Lindsay, B. G. (1983) *Ann. Stat.* **11**, 486–497.