Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 3003–3005, May 1989
Statistics

Estimating the superiority of a drug to a placebo when all and only those patients at risk are treated with the drug

(clinical trials/estimation under biased sampling/u, v method)

HERBERT ROBBINS AND C.-H. ZHANG

Department of Statistics, Rutgers University, New Brunswick, NJ 08903

Contributed by Herbert Robbins, December 9, 1988

ABSTRACT It is shown that, under certain assumptions, one can estimate the difference between the effect of a treatment and that of a placebo even when the treatment has been given to all and only those patients who are at risk (as evidenced by a screening examination).

A new drug is to be tested for its effect on, say, hypertension. For a patient randomly chosen from some population let

θ = the patient's "true" (unobservable) blood pressure

x = the patient's blood pressure reading obtained at a screening examination before any treatment is undertaken.

We shall assume that given θ, x is $N(\theta, \sigma^2)$, where σ is a constant, known or unknown. We make no assumption about how θ is distributed in the population.

Suppose that if $x > a$ the patient is regarded as at risk. A standard method for evaluating the new drug is to allocate randomly half of all such patients to the new drug and half to a placebo. Suppose, however, that for ethical or other reasons we have adopted the following allocation protocol:

$$(A) \quad \begin{cases} \text{if } x > a, \text{ the patient is treated with the drug} \\ \text{if } x \le a, \text{ the patient is treated with a placebo.} \end{cases}$$

Let

y = the blood pressure reading of the patient after treatment.

From the observed values $(x_1, y_1), \ldots, (x_n, y_n)$ for n patients, we want to estimate the parameter

τ = mean effect of the drug, as compared to the placebo, over the population at risk $(x > a)$.

It is not clear *a priori* that a consistent estimator of τ can be found under the allocation protocol (A), but in the section below we shall show how to do this under an assumption (3 below) concerning y.

Consistent Estimation of τ. LEMMA. *Assume that* (θ, x) *is a random vector such that for some constant* $\sigma > 0$,

$$\text{given } \theta, \text{ x is } N(\theta, \sigma^2). \quad [1]$$

If $u(\cdot)$ *is of bounded variation* (*b.v.*) *and absolutely continuous* (*a.c.*) *on* $(-\infty, \infty)$, *then*

$$E[u(x)\theta] = E[xu(x)] - \sigma^2 Eu'(x). \quad [2]$$

Proof. $\theta E[u(x)|\theta]$

$$= \frac{\theta}{\sigma}\int u(x)\varphi\left(\frac{x-\theta}{\sigma}\right)dx \qquad (\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2))$$

$$= -\int u(x)\left(\frac{x-\theta}{\sigma}\right)\varphi\left(\frac{x-\theta}{\sigma}\right)dx + \frac{1}{\sigma}\int xu(x)\varphi\left(\frac{x-\theta}{\sigma}\right)dx$$

$$= \sigma\int u(x)\varphi'\left(\frac{x-\theta}{\sigma}\right)dx + E[xu(x)|\theta]$$

$$= -\sigma\int \varphi\left(\frac{x-\theta}{\sigma}\right)u'(x)dx + E[xu(x)|\theta]$$

$$= -\sigma^2 E[u'(x)|\theta] + E[xu(x)|\theta],$$

which, since θ is arbitrary, implies Eq. **2**.

THEOREM 1. *Assume that* (θ, x, y) *is a random vector such that assumption* **1** *holds and also that for some constants* a *and* c,

$$E[y|\theta, x] = \theta + c + \delta_a(x) \cdot t(\theta, x), \quad [3]$$

where by definition

$$\delta_a(x) = \begin{cases} 1 & \text{if } x > a \\ 0 & \text{if } x \le a \end{cases}$$

and $t(\cdot, \cdot)$ *is arbitrary. If* $u(\cdot)$ *is b.v. and a.c. on* $(-\infty, \infty)$, *then*

$$E[u(x)(y - x)] + \sigma^2 Eu'(x) = cEu(x) + E[u(x)\delta_a(x)t(\theta, x)]. \quad [4]$$

Proof. By assumption **3**,

$$E[u(x)y|\theta, x] = u(x)[\theta + c + \delta_a(x)t(\theta, x)],$$

so from formula **2** it follows that

$$E[u(x)y] = E[xu(x)] - \sigma^2 Eu'(x) + cEu(x) + E[u(x)\delta_a(x)t(\theta, x)],$$

which was to be proved.

Setting $u = 1$ in formula **4** gives the following.

COROLLARY 1. *Under assumptions* **1** *and* **3**,

$$E[\delta_a(x)t(\theta, x)] = E(y - x) - c. \quad [5]$$

We shall also need the following.

COROLLARY 2. *Under assumptions* **1** *and* **3**, *if* $u_1(\cdot)$ *and*

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: b.v., bounded variation; a.c., absolutely continuous; a.s., almost surely.

$u_2(\cdot)$ *are b.v. and a.c. on* $(-\infty, \infty)$ *and vanish for* $x > a$, *then*

$$\begin{cases} E[u_1(x)(y - x)] + \sigma^2 Eu_1'(x) = cEu_1(x) \\ E[u_2(x)(y - x)] + \sigma^2 Eu_2'(x) = cEu_2(x) \end{cases} \quad [6]$$

and hence

$$c = \frac{Eu_2'(x) \cdot E[u_1(x)(y - x)] - Eu_1'(x) \cdot E[u_2(x)(y - x)]}{Eu_2'(x) \cdot Eu_1(x) - Eu_1'(x) \cdot Eu_2(x)} \quad [7]$$

provided that the denominator is not 0.

We now define the parameter τ by

$$\tau = E[t(\theta, x)|x > a] = \frac{E[\delta_a(x)t(\theta, x)]}{E\delta_a(x)}$$

$$= \frac{E(y - x) - c}{E\delta_a(x)} \quad \text{(by formula 5)}. \quad [8]$$

(In the case of hypertension we hope that t, and hence τ, is negative.)

We can estimate c and τ by

$$c_n = \frac{\sum_1^n u_2'(x_i) \cdot \sum_1^n u_1(x_i)(y_i - x_i) - \sum_1^n u_1'(x_i) \cdot \sum_1^n u_2(x_i)(y_i - x_i)}{\sum_1^n u_2'(x_i) \cdot \sum_1^n u_1(x_i) - \sum_1^n u_1'(x_i) \cdot \sum_1^n u_2(x_i)} \quad [9]$$

and

$$\tau_n = \frac{\sum_1^n (y_i - x_i) - nc_n}{\sum_1^n \delta_a(x_i)}, \quad [10]$$

where by hypothesis (θ, x, y), (θ_1, x_1, y_1), . . . are independent, identically distributed random vectors such that assumptions **1** and **3** hold. It is clear from formulas **7–10** that the following theorem holds.

THEOREM 2. *As* $n \to \infty$,

$$c_n \to c, \qquad \tau_n \to \tau \qquad a.s.$$

Moreover, $\sqrt{n}(c_n - c)$ *and* $\sqrt{n}(\tau_n - \tau)$ *have limiting normal distributions with* 0 *means.*

Remark 1. The functions $u_1(\cdot)$ and $u_2(\cdot)$ in formulas **6**, **7**, and **9** are assumed to be b.v. and a.c., vanishing for $x > a$, and such that the denominator of formula **7** is not 0. Subject to these restrictions, they are arbitrary. We do not know how to choose them so as to minimize the limiting variance of $\sqrt{n}(c_n - c)$ or $\sqrt{n}(\tau_n - \tau)$.

Remark 2. If σ is *known*, instead of using formulas **9** and **10** we can estimate c by

$$c_n^* = \frac{\sum_1^n u_1(x_i)(y_i - x_i) + \sigma^2 \sum_1^n u_1'(x_i)}{\sum_1^n u_1(x_i)} \quad [11]$$

and τ by

$$\tau_n^* = \frac{\sum_1^n (y_i - x_i) - nc_n^*}{\sum_1^n \delta_a(x_i)}, \quad [12]$$

provided that $u_1(\cdot)$ is b.v. and a.c., vanishes for $x > a$, and $Eu_1(x) \neq 0$.

Remark 3. An inspection of the proofs shows that all the foregoing formulas remain valid even if the b.v. functions $u(\cdot)$ occurring in formulas **2**, **4**, **6**, and **7** are not a.c., provided that we always replace

$$Eu'(x) \quad \text{by} \quad \int f(x)du(x), \quad [13]$$

where $f(\cdot)$ is the probability density function of the random variable x. In particular, choosing

$$u_1(x) = 1 - \delta_a(x), \quad u_2(x) = 1 - \delta_b(x) \text{ for some } b < a, \quad [14]$$

we obtain the formulas

$$c = \frac{E[(1 - \delta_a(x))(y - x)] - \sigma^2 f(a)}{E[1 - \delta_a(x)]}$$

$$= \frac{f(b)E[(1 - \delta_a(x))(y - x)] - f(a)E[(1 - \delta_b(x))(y - x)]}{f(b)E[1 - \delta_a(x)] - f(a)E[1 - \delta_b(x)]}, \quad [15]$$

which can be used when σ is known or unknown to estimate c (and hence τ), provided that we have consistent density estimators $f_n(a)$, $f_n(b)$ of $f(a)$ and $f(b)$. Such estimators are available, and have an $n^{-1/2}$ rate of convergence, if we make the *additional assumption* that for some α and $\beta > 0$,

$$\theta \text{ is } N(\alpha, \beta^2). \quad [16]$$

For then x will (from assumption **1**) be $N(\alpha, \gamma^2)$ with $\gamma^2 = \beta^2 + \sigma^2$, and hence

$$f(x) = \frac{1}{\gamma}\varphi\left(\frac{x - \alpha}{\gamma}\right). \quad [17]$$

But as $n \to \infty$,

$$\bar{x} = \frac{1}{n}\sum_1^n x_i \to \alpha, \qquad s^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2 \to \gamma^2 \qquad \text{a.s.},$$

and hence for any fixed x, as $n \to \infty$

$$f_n(x) = \frac{1}{s}\varphi\left(\frac{x - \bar{x}}{s}\right) \to f(x) \qquad \text{a.s.}$$

We may therefore estimate $f(a)$ and $f(b)$ by

$$f_n(a) = \frac{1}{s}\varphi\left(\frac{a - \bar{x}}{s}\right), \qquad f_n(b) = \frac{1}{s}\varphi\left(\frac{b - \bar{x}}{s}\right) \quad [18]$$

and define (for use when σ is known)

$$c_n^* = \frac{\sum_1^n [1 - \delta_a(x_i)](y_i - x_i) - n\sigma^2 f_n(a)}{\sum_1^n [1 - \delta_a(x_i)]} \quad [19]$$

and (for use when σ is unknown)

$$c_n = \frac{f_n(a)\sum_1^n [1 - \delta_b(x_i)](y_i - x_i) - f_n(b)\sum_1^n [1 - \delta_a(x_i)](y_i - x_i)}{f_n(a)\sum_1^n [1 - \delta_b(x_i)] - f_n(b)\sum_1^n [1 - \delta_a(x_i)]} \quad [20]$$

to obtain consistent estimators of c with $n^{-1/2}$ rates of convergence. It would, however, be safer to use formula **9** or **11** instead of formula **20** or **19** if it is not certain that θ is in fact normally distributed.

Remark 4. From formula **8** and the first part of formula **15** it follows that

$$\begin{aligned}\tau &= \frac{E(y-x)-c}{E\delta_a(x)}\\ &= \frac{E(y-x)}{E\delta_a(x)} - \frac{E[(1-\delta_a(x)(y-x)]-\sigma^2 f(a)}{E[1-\delta_a(x)]E\delta_a(x)}\\ &= \frac{E[1-\delta_a(x)]E(y-x)-E[(1-\delta_a(x))(y-x)]+\sigma^2 f(a)}{E\delta_a(x)\cdot E[1-\delta_a(x)]}\\ &= \frac{E[\delta_a(x)(y-x)]}{E\delta_a(x)} - \frac{E[(1-\delta_a(x))(y-x)]}{E[1-\delta_a(x)]}\\ &\quad + \frac{\sigma^2 f(a)}{E\delta_a(x)E[1-\delta_a(x)]}.\end{aligned}$$

Thus, under (A), the statistic

$$\begin{aligned}&(\text{average of } y_i - x_i \text{ for those treated with drug}) -\\ &\quad(\text{average of } y_i - x_i \text{ for those treated with placebo}) \qquad [21]\end{aligned}$$

converges as $n \to \infty$ to

$$\tau - \frac{\sigma^2 f(a)}{P(x > a)\cdot P(x \le a)},$$

which is *less* than τ, so that even if t and hence τ is 0 the value of the statistic **21** will usually be negative.

Remark 5. If we replace the unknown constant in assumption **3** by any linear combination

$$c_1 g_1(x) + \ldots + c_k g_k(x) \qquad [22]$$

of known functions with unknown coefficients $c_1, \ldots, c_k$, then it is clear how to generalize formulas **6** and **9** to estimate these coefficients by using functions $u_j(x), j = 1, \ldots, k+1$.

Remark 6. When in assumption **3** the function $t(\theta, x)$ is a constant and y, given θ and x, is $N(\theta + c + \delta_a(x)\cdot t, \sigma^2)$, the method of conditional maximum likelihood can be used to estimate t, as by Robbins and Zhang (1). There are some technical difficulties in the present case, and we defer a comparison with the method for consistent estimation of τ described above to a later date.

This research was supported in part by the National Science Foundation.

1. Robbins, H. & Zhang, C.-H. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 3670–3672.