ESTIMATION OF THE MEDIANS FOR DEPENDENT VARIABLES

By Olive Jean Dunn[1]

Statistical Laboratory, Iowa State College

1. Summary. Joint intervals of bounded confidence are suggested for the medians of a bivariate population with continuous marginal distributions. The two intervals are of the classic type based on sample order statistics.

2. Introduction. The problem considered in this paper is that of using a nonparametric method to estimate by a confidence set the unknown medians of two dependent variables. In various types of research, it is convenient to consider a sample of n individuals and to take measurements on the same n individuals at two different times or at two different levels of treatment. The two measurements on the same individual cannot be assumed to be independent, so that it is appropriate to consider the $2n$ measurements as a sample of size n from a bivariate distribution.

Let the two variables y_1, y_2 with medians ν_1, ν_2 have the c.d.f. (cumulative distribution function) $F(y_1, y_2)$. By a set of simultaneous confidence intervals of bounded confidence level $1 - \alpha$ for ν_1, ν_2 is meant a set of four functions of the sample values, say g_1, g_2, h_1, h_2, such that

$$P(g_1 < \nu_1 < h_1, g_2 < \nu_2 < h_2) \geqq 1 - \alpha.$$

The probability relationship must hold for all underlying distributions in a specified set of distributions. In this paper the specified set will consist of all bivariate distributions whose marginals have continuous c.d.f.'s.

The method used in this paper to obtain confidence intervals uses order statistics and requires only the assumption that the marginal distributions be continuous.

3. Confidence intervals for the medians of a bivariate distribution. Let the c.d.f. of the variables y_1, y_2 be $F(y_1, y_2)$ and let the two marginal distributions be $F_1(y_1)$ and $F_2(y_2)$, both of which are continuous.

A random sample of n observations will be denoted by (y_{11}, y_{21}), $\cdots$, (y_{1n}, y_{2n}). For $i = 1$ and 2, the set $y_{i1}, \cdots, y_{in}$ will be reordered from smallest to largest and renamed $z_{i1}, \cdots, z_{in}$. Thus $z_{i1} \leqq z_{i2} \leqq \cdots \leqq z_{in}$ for $i = 1$ or 2. The z_{1j} and z_{2j} need not belong to the same observation.

Two positive integers, r and s, such that $2r + s = n$, are selected. Let E_i be the event that $z_{ir} < \nu_i < z_{i,r+s+1}$ for $i = 1, 2$. Then, for $i = 1, 2$,

$$(1) \qquad P(E_i) = \sum_{j=r}^{r+s} \binom{n}{j} (\tfrac{1}{2})^n = (1 - \alpha)^{1/2}, \qquad \text{say.}$$

Received February 25, 1957; revised June 20, 1958.

[1] Now at the University of California at Los Angeles.

If the variables were independent, probabilities could be multiplied, so that $P(E_1E_2) = 1 - \alpha$. This would give a set of intervals of exact confidence level for ν_1 and ν_2, namely z_{1r} to $z_{1,r+s+1}$ for ν_1, z_{2r} to $z_{2,r+s+1}$ for ν_2.

For dependent variables, the same set of intervals may be used as a set with bounded confidence level for it can be shown that $P(E_1E_2) \geqq 1 - \alpha$. It should be noted that symmetric order statistics have been used, and indeed the result would not hold otherwise.

The following proof establishes the necessary inequality.

THEOREM. $P(E_1E_2) \geqq P(E_1)P(E_2)$.

PROOF. Since $P(E_1E_2) = P(E_2 \mid E_1)P(E_1)$, it will be sufficient to prove that $P(E_2 \mid E_1) \geqq P(E_2)$.

If for a certain observation, $y_{1j} > \nu_1$, then let the conditional probability that $y_{2j} > \nu_2$ (this will be referred to as the probability of a "success") be denoted by p. Then using the fact that $F_1(\nu_1) = F_2(\nu_2) = \frac{1}{2}$:

$$p = P(y_{2j} > \nu_2 \mid y_{1j} > \nu_1) = 2F(\nu_1, \nu_2). \tag{2}$$

Similarly, if it is known that $y_{1j} < \nu_1$, then let q be the conditional probability that $y_{2j} > \nu_2$ (probability of a "success"). Then

$$q = P(y_{2j} > \nu_2 \mid y_{1j} < \nu_1) = 1 - 2F(\nu_1, \nu_2). \qquad \text{Thus, } p + q = 1. \tag{3}$$

If it is known that E_1 has occurred, then $r + i$ observations have $y_{1j} < \nu_1$, and so have a conditional probability of success of q, where i may be $0, 1, \cdots s$; $r + s - i$ observations have $y_{1j} > \nu_1$, and so have a conditional probability of success of p.

To obtain a generating function for the probabilities of various numbers of successes, conditioned by the fact that $y_{1r} < \nu_1 < y_{1,r+s+1}$, one may proceed as follows. Let $E_1(i)$ be the event that E_1 occurs with $r + i$ observations having $y_{1j} < \nu_1$. Then $E_1 = \bigcup_0^s E_1(i)$.

Let $Y_<$ be the number of successes in the $r + i$ observations which have $y_{1j} < \nu_1$; let $Y_>$ be the number of successes in the $r + s - i$ observations which have $y_{1j} > \nu_2$; let $Z = Y_< + Y_>$ be the total number of successes. Then

$$\begin{aligned} E\left(t^Z \mid \bigcup_0^s E_1(i)\right) &= \sum_{i=0}^s E(t^Z \mid E_1(i)) \cdot P(E_1(i) \mid E_1) \\ &= \sum_{i=0}^s E(t^Z \mid E_1(i)) \cdot P(E_1(i))/P(E_1). \end{aligned} \tag{4}$$

Since under the condition $E_1(i)$, $Y_<$ and $Y_>$ are independent,

$$E\left(t^Z \mid \bigcup_0^s E_1(i)\right) = \sum_{i=1}^s E(t^{Y<} \mid E_1(i))E(t^{Y>} \mid E_1(i))P(E_1(i))/P(E_1). \tag{5}$$

The generating function G for the conditional probabilities of various numbers of successes, given that $z_{1r} < \nu_1 < z_{1,r+s+1}$, is, finally,

$$G = C \sum_{i=0}^{s} \frac{1}{(r+i)!(r+s-i)!} (p+qt)^{r+i}(q+pt)^{r+s-i}, \tag{6}$$

where

$$C = 1 \Big/ \sum_{i=0}^{s} \frac{1}{(r+i)!(r+s-i)!}.$$

If a_j, $j = 0, 1, \cdots, n$, is defined as the coefficient of t^j in G, then

$$G = \sum_{j=0}^{n} a_j t^j, \tag{7}$$

and

$$P(E_2 \mid E_1) = \sum_{j=r}^{r+s} a_j. \tag{8}$$

The task now is to show that $P(E_2 \mid E_1)$ has a minimum at $p = \frac{1}{2}$. To do this, the derivative of G is obtained indirectly by differentiating equation (6) with respect to p, and then manipulating the derivative, G_p, as follows:

$$\begin{aligned} G_p = C \sum_{i=0}^{s} \frac{(1-t)}{(r+i)!(r+s-i)!} [(r+i)(p+qt)^{r+i-1}(q+pt)^{r+s-i} \\ - (r+s-i)(p+qt)^{r+i}(q+pt)^{r+s-i-1}] \\ = \frac{C(1-t)}{(r-1)!(r+s)!} (p+qt)^{r-1}(q+pt)^{r-1}[(q+pt)^{s+1} - (p+qt)^{s+1}]. \end{aligned} \tag{9}$$

Let $C^* = (C/(r-1)!(r+s)!)$. Then

$$\begin{aligned} G_p &= C^*(1-t)(p+qt)^{r-1}(q+pt)^{r-1}[(q^{s+1} - p^{s+1}) + \tbinom{s+1}{1} pq(q^{s-1} - p^{s-1})t \\ &+ \tbinom{s+1}{2} p^2q^2(q^{s-3} - p^{s-3})t^2 + \cdots + \tbinom{s+1}{2} p^2q^2(p^{s-3} - q^{s-3})t^{s-1} \\ &+ \tbinom{s+1}{1} pq(p^{s-1} - q^{s-1})t^s + (p^{s+1} - q^{s+1})t^{s+1}] \\ &= C^*(1-t)(q-p)(p+qt)^{r-1}(q+pt)^{r-1}[b_0(1 - t^{s+1}) \\ &+ b_1 t(1 - t^{s-1}) + b_2 t^2(1 - t^{s-3}) + \cdots] \\ &= C^*(1-t)^2(q-p)(p+qt)^{r-1}(q+pt)^{r-1}[b_0 + (b_0 + b_1)t \\ &+ (b_0 + b_1 + b_2)t^2 + \cdots + (b_0 + b_1 + b_2)t^{s-2} + (b_0 + b_1)t^{s-1} + b_0 t^s]. \end{aligned} \tag{10}$$

Here $b_j = \binom{s+1}{j} p^j q^j (q^{s-2j} + q^{s-2j-1}p + \cdots + p^{s-2j})$; hence the partial sums of the b_j's appearing as coefficients within the preceding square brackets are positive and (excluding the trivial cases $p = 0$ or 1) they increase toward the center coefficient(s).

Multiplication of the polynomial within the square brackets by

$$(p+qt)^{r-1}(q+pt)^{r-1}$$

can be accomplished by successive multiplications by $(p + qt)(q + pt)$, and it is easily verified that each multiplication yields a polynomial whose coefficients increase toward the center coefficient(s). Thus

$$(11)\quad G_p = C^*(q - p)(1 - 2t + t^2)$$
$$(c_0 + c_1t + c_2t^2 + \cdots + c_2t^{n-4} + c_1t^{n-3} + c_0t^{n-2})$$
$$= C^*(q - p)(d_0 + d_1t + d_2t^2 + \cdots + d_2t^{n-2} + d_1t^{n-1} + d_0t^n).$$

Here $d_j = c_j - 2c_{j-1} + c_{j-2}$, for $j = 0, 1, \cdots$, and it is understood that $c_{-1} = c_{-2} = 0$.

The derivative with respect to p of the conditional probability $P(E_2 \mid E_1)$ is then $\sum_{j=r}^{r+s} C^*(q - p)\, d_j = 2C^*(q - p)(c_{r-2} - c_{r-1})$. Since $c_{r-2} - c_{r-1}$ is always negative, the derivative is positive, zero, or negative according to whether $p > \frac{1}{2}$, $p = \frac{1}{2}$ or $p < \frac{1}{2}$. Thus $P(E_2 \mid E_1)$ has a minimum at $p = \frac{1}{2}$, and $P(E_1E_2) \geqq P(E_1)P(E_2)$.

Since $E_1E_2 \subset E_1$, one may further write $P(E_1) \geqq P(E_1E_2) \geqq P(E_1)P(E_2)$. The confidence level for the intervals z_{1r} to $z_{1,r+s+1}$, z_{2r} to z_{2r} to $z_{2,r+s+1}$ can actually be as high as $(1 - \alpha)^{1/2}$ (for a distribution function such that $p = 0$ or $p = 1$) or as low as $(1 - \alpha)$ (when $p = \frac{1}{2}$).

4. Evaluation. One way to compare sets of confidence intervals is on the basis of their lengths, or the expected values of their lengths. I shall exhibit some length comparisons when y_1, y_2 are jointly normally distributed with means (or medians) ν_1, ν_2, variances σ_1^2, σ_2^2, and arbitrary co-variance. In Table I, the intervals for the medians obtained by the method of this paper (Method I) are compared with three sets of intervals for the means of a bivariate normal distribution (Methods II and III and IV) obtained in another paper [1]. It should be mentioned that all four methods lead to intervals of bounded confidence.

The figures given in the body of the table are values of $(n^{1/2}/\sigma_i)E(\frac{1}{2}l_i)$, where l_i is the length of the confidence interval for ν_i, $i = 1$ or 2. Values of $1 - \alpha$ (which for Method I cannot be chosen arbitrarily) have been selected as close as possible to .95.

Method II (section 4.2 in [1]), which uses Hotelling's T distribution, is similar to Method I in that no assumptions need be made concerning the variances. When n is small, the intervals are seen to be slightly longer on the average for II than for I.

Method III, based on the Student t distribution, requires that the variances be equal, though they may be unknown. These intervals are seen to be somewhat shorter than those from Method I. Method III is found in section 7.2 of [1].

In Method IV, the variances are assumed to be known and the intervals obtained (the method of section 7.1 in [1]) are the shortest possible intervals for means of the bivariate normal distribution when nothing is known about the

TABLE I

Comparison of Expected Values of Lengths of Confidence Intervals for the Means of a Bivariate Normal Distribution

$\frac{\sqrt{n}}{\sigma_i} E\,(\tfrac{1}{2}l_i)$ for:

n	$1 - \alpha$	$(1 - \alpha)^{1/2}$	Method I: Order Statistics		Method II: Hotelling's T	Method III: Variances Unknown but Equal	Method IV: Variances Known
6	.939	.969	z_{i1} to z_{i6}	3.10	3.72	2.83	2.16
8	.984	.992	z_{i1} to z_{i8}	4.03	4.40	3.54	2.65
10	.958	.979	z_{i2} to z_{i9}	3.17	3.21	2.72	2.31
20	.976	.988	z_{i5} to z_{i16}	3.33	3.12	2.77	2.51
100	.958	.979	z_{i39} to z_{i62}	2.93	2.57	2.33	2.31

l_i = length of the confidence interval for μ_i (or ν_i), $i = 1, 2$.

Method I: $E(\frac{1}{2}l_i)$ computed using the expected value of order statistics in [2].

Method II: The intervals are $\bar{y}_i \pm (\hat{\sigma}_i/\sqrt{n})c_\alpha$, $i = 1, 2$, where c_α is the $1 - \alpha$ point in the distribution of Hotelling's T, and $\hat{\sigma}_i^2 = \sum_{j-1}^{n} (y_{ij} - \bar{y}_i)^2/(n - 1)$, $i = 1, 2$.

Method III: The intervals are $\bar{y}_i \pm (\hat{\sigma}_1/\sqrt{n})c_\alpha$, $i = 1, 2$, where c_α is the

$$[1 + (1 - \alpha)^{1/2}]/2$$

point of the Student t distribution with $n - 1$ degrees of freedom, and

$$\hat{\sigma}_1^2 = \sum_{j-1}^{n} (y_{1j} - \bar{y}_1)^2/(n - 1)$$

Method IV: The intervals are $\bar{y}_1 \pm (\hat{\sigma}_i/\sqrt{n})\cdot c_\alpha$, $i = 1, 2$, where c_α is the

$$[1 + (1 - \alpha)^{1/2}]/2$$

point of the standard normal distribution.

covariance. They thus make a useful standard for purposes of comparison. These "best" intervals are considerably shorter than those from Method I, but this must be balanced against the fact that for the latter method no assumptions are necessary concerning the variance or concerning distributional form (except for continuity).

The fact that Method I, assuming nothing about the form of the distribution, gives shorter intervals for small n than Method II, which demands a normal distribution may seem somewhat surprising. The explanation lies in the fact that, for a given value of α, the actual probability of coverage is higher for Method II than for Method I. For $\rho = 0$ and $\rho = 1$, the actual probabilities of coverage for all four methods are as follows:

n	Methods I, III, IV		Method II	
	$\rho = 0$	$\rho = 1$	$\rho = 0$	$\rho = 1$
6	.939	.969	.989	.994
8	.984	.992	.997	.999
10	.958	.979	.991	.995
20	.976	.988	.994	.997
100	.958	.979	.988	.994

Throughout the preparation of this paper, it was conjectured that the same set of intervals developed here might be used for a k-variate distribution. Mr. Ernest V. Scheuer has, however, recently drawn the author's attention to the following counterexample, which shows that it is possible for $P(E_1E_2E_3)$ to be less than $P^3(E_1)$, where E_i is the event that $z_{ir} < \nu_i < z_{i,r+s+1}$, $i = 1, 2, 3$.

Let $r = 1$, and, for simplicity, let $\nu_1 = \nu_2 = \nu_3 = 0$.

Let $p_{111} = P(y_1 > 0, y_2 > 0, y_3 > 0)$, $p_{110} = P(y_1 > 0, y_2 > 0, y_3 < 0)$, and similarly for p_{101}, p_{011}, p_{100}, p_{010}, p_{001}, and p_{000}.

It can be readily verified that $P(z_{11} < 0 < z_{1n}, z_{21} < 0 < z_{2n}, z_{31} < 0 < z_{3n})$ is smaller for $p_{111} = p_{100} = p_{010} = p_{001} = 1/4$, $p_{110} = p_{101} = p_{011} = p_{000} = 0$ than it is under independence ($p_{111} = p_{110} = \cdots = p_{000} = 1/8$).

REFERENCES

[1] Olive Jean Dunn, "Estimation of the means of dependent variables," *Ann. Math. Stat.*, Vol. 29 (December 1958), pp. 1095–1111.

[2] D. Teichroew, "Tables of expected values of order statistics and products of order statistics for samples of size twenty or less from the normal distribution," *Ann. Math. Stat.* Vol. 27 (June 1956), pp. 410–426.