NONPARAMETRIC ESTIMATION IV

BY D. A. S. FRASER AND R. WORMLEIGHTON[1]

University of Toronto

1. Summary. In the three papers, [1], [2], [3], entitled "Nonparametric estimation", Scheffé and Tukey generalized previous results on tolerance regions and extended them to cover all continuous and discontinuous distribution functions. This note contains four comments arising from these papers: first, on a method for giving bounds to the confidence level in the discontinuous case which can lower the probability that the end points need to have part, a random variable, of their probability neglected to maintain the given confidence level; second, on a correction of a statement of results in [2]; third, on a proof in [2] requiring a further statement; fourth, a necessary restatement of theorems in [3].

2. Bounds for the confidence level. In paper [1] Scheffé and Tukey extend the theory of tolerance regions to the one dimensional discontinuous case, and obtain the following statement:

$$Pr\{C_{(p)-0,(q)+0} \geqq \beta\} \geqq 1 - \alpha_{q-p} \geqq Pr\{C_{(p)+0,(q)-0} \geqq \beta\}, \tag{2.1}$$

where $C_{(p)\pm 0,(q)\mp 0}$ are respectively the coverages of the open and closed intervals with end points the pth and qth order statistics z_p and $z_q (q > p)$ of a sample of n from a distribution and α_{q-p} is the incomplete Beta function $I_\beta(q - p, n - p + q + 1)$.

This statement implies the following statements:

$$1 - \alpha_{q-p+2} \geqq Pr\{C_{(p-1)+0,(q+1)-0} \geqq \beta\},$$

$$Pr\{C_{(p+1)-0,(q-1)+0} \geqq \beta\} \geqq 1 - \alpha_{q-p-2}\,.$$

This suggests giving bounds for the confidence levels of the tolerance regions of statement (2.1).

Let us consider the one dimensional representation theorem with its "inverse probability integral transformation". This transformation labelled $g_F(x^*)$ mapped x^* with a uniform distribution into x with the given distribution represented by $F(x)$. z^* and z refer to the corresponding order statistics. Take any interval on the range of the uniform distribution whose end points lie respectively in the closed intervals (z^*_{p-1}, z^*_p) and (z^*_q, z^*_{q+1}). The confidence level, that the coverage of this interval is at least β, lies between $1 - \alpha_{q-p}$ and $1 - \alpha_{q-p+2}$. Apply the mapping $g_F(x^*)$. The confidence level lies between $1 - \alpha_{q-p}$ and $1 - \alpha_{q-p+2}$ that the following coverage is greater than or equal to β: $C_{(p)-0,(q)+0}$, if z_p is distinct from z_{p-1} and z_q is distinct from z_{q+1}; $C_{(p)-0,(q)-0}$ + "fraction" of the coverage of z_q if z_p is distinct from z_{p-1} and z_q identical to z_{q+1}; and similarly for the other two possible cases. The "fraction" (a number between 0 and 1) can

[1] The authors wish to thank Professor John Tukey for suggesting Definitions 5.2 and 5.3.

be considered as a random variable as determined by the above mapping or as a fixed value since the relation must be true for at least one fixed value for any given distribution and integers p and q. In either case it is unknown to the practitioner and the interpretation would be unimportant.

Similarly we obtain the following result: The confidence level lies between $1 - \alpha_{q-p}$ and $1 - \alpha_{q-p+r+s}$ that the following coverage is greater than or equal to β: $C_{(p)-0,(q)+0}$, if z_p is distinct from z_{p-r} and z_q is distinct from z_{q+s}; $C_{(p)-0,(q)-0}$ + "fraction" of the coverage of z_q, if z_p is distinct from z_{p-r} and z_q is identical to z_{q+s}; and similarly for the other two cases.

The open interval can be treated in a similar manner.

The application of these results would be for the practitioner who was familiar with the type of data he was to receive and realized that perhaps two or three order statistics would be tied on one or perhaps both tails. He would then choose r and s to give as tight control of the confidence level consistent with a reasonable determinacy in the tolerance interval (the probability being small that the coverage should be considered as including only part instead of all of the coverage of the end points).

These results also generalize to the multivariate case with little alteration. For example consider the following result which would correspond to the closed interval case above. The confidence level lies between $1 - \alpha_{q-p}$ and $1 - \alpha_{q-p+r}$ that the following coverage is greater than or equal to β: cov $\{\bar{B}_\lambda\}$, if $\bar{B}_\lambda$ is contained in $B_{\lambda+\mu}$ where μ consists of r of the integers $(1, 2, \cdots, n + 1)$ which are not contained in λ; cov $\{B_\lambda\}$ + "fraction" of (cov $\{\bar{B}_\lambda\}$ − cov $\{B_\lambda\}$), if $\bar{B}_\lambda$ is not contained in $B_{\lambda+\mu}$. Here the "fraction" can be considered a random variable or fixed, in either case unknown to the practitioner (λ containing $q - p$ integers).

In formulating the above generalization, attention was drawn to the fact that the block groups did not form a proper sequence as λ was increased. By the following counter example the theorems in [3] are seen to be incorrect using the given definition of block groups. Rectifying definitions are presented in Section 5.

Following the notation of [3], let

$$\begin{aligned}\varphi_1(x, y) &= y,\\ \varphi_2(x, y) &= x,\\ \varphi_3(x, y) &= -y,\\ \varphi_4(x, y) &= -x,\end{aligned}$$

and $p_i = i (i = 1, 2, 3, 4)$.

Consider the distribution $F(x, y) = \epsilon(x)\epsilon(y)$ where $\epsilon(x)$ is defined by

$$\begin{aligned}\epsilon(x) &= 0, \quad x < 0,\\ &= 1, \quad x \geqq 0.\end{aligned}$$

Take a sample of $n \geqq 6$ from this distribution; the sample values will all be (0, 0) with probability one.

$$\begin{aligned}
S_1 &= \{(x, y) \mid y > 0 \text{ or } y = 0,\ x > 0\},\\
T_1 &= \{(0, 0)\},\\
S_2 &= \{(x, y) \mid y < 0,\ x \geqq 0\},\\
T_2 &= \text{Null set},\\
S_3 &= \{(x, y) \mid x < 0,\ y \leqq 0\},\\
T_3 &= \text{Null set},\\
S_4 &= \text{Null set}.
\end{aligned}$$

The corresponding coverages are respectively 0, 1, 0, 0, 0, 0, 0.

Taking $\lambda = \{3\}$ we find by the definition of block groups that

$$\bar{B}_\lambda = T_2 \cup S_3 \cup T_3, \qquad \bar{C}_\lambda = 0$$

with probability 1. Thus $Pr\{\bar{C}_\lambda < a\} = 1 \nleqq I_a(1, n)$.

Taking $\lambda = \{1, 2\}$ we have $B_\lambda = S_1 \cup T_1 \cup S_2$, and $C_\lambda = 1$ with probability 1. Thus $Pr\{C_\lambda < a\} = 0 \ngeqq I_a(2, n - 1)$.

The proof in [3] is in error on page 39, the seventh line from the bottom.

3. Correction of a statement of results. In paper [2] on page 536 a "Statement of Results for Measure Theorists" is given. The theorem B_{n+1} should read: Hold the n functions $\varphi_1, \varphi_2, \cdots, \varphi_n$ and the probability measure fixed, then T^n is mapped on B_n and the power measure μ^n is carried by that mapping into a measure of B_n. This measure is always $n!/\sqrt{n+1}$ times Lebesgue measure.

4. A proof; a further statement. In the proof on page 537 of paper [3], the problem is to show that the distribution of $n - m$ variates is the same when obtained by two methods of calculation; more particularly, to show that, given that in a sample of n, one value falls in each of the sets $A_1, A_2, \cdots, A_m$ and the remaining $n - m$ fall in B, then the distribution of the $n - m$ in B is that of a sample of $n - m$ restricted to B. The statement is made that the probability of the above, and in addition that the $n - m$ falling in B, fall in $R \subset B$, is

$$\frac{n!}{(n-m)!}\, \mu(A_1)\mu(A_2) \cdots \mu(A_m)\mu^{n-m}(B)$$

times the probability that a sample of $n - m$ restricted to B falls in R. To show that the distributions are identical, a further statement is needed: that for one variate in each A_i, for p falling in R, and $n - m - p$ in $B - R$, then the probability is equal to

$$\frac{n!}{(n-m)!}\, \mu(A_1) \cdots \mu(A_m)\mu^{n-m}(B)$$

times the probability that a sample of $n - m$ restricted to B divides p into R and $n - m - p$ into $B - R$.

5. Restatement of theorems in [3]. As has been noted in Section 2 above, Theorems $A^*_{m|n+1}$ and B^*_{n+1} fail when actual ties (coincident points) occur. The following redefinition of the block groups overcomes this difficulty and the proof follows as given in [3].

Define S_i and T_i as in (4.1) in [3].

DEFINITION 5.1. *Let S_i be given by the definition for S_i where $<$ is replaced by $\leqq$ and $>$ is replaced by $\geqq$.*

DEFINITION 5.2. *The block group B_λ consists of the union of all S_i with i in λ and all T_j not contained in any $\bar{S}_i$ with i not in λ.*

DEFINITION 5.3. *The closed block group $\bar{B}_\lambda$ consists of the union of all S_i with i in λ and all T_j contained in any $\bar{S}_i$ with i in λ.*

Using the above definitions, Theorems $A^*_{m|n+1}$ and B^*_{n+1} follow provided the "m-system of functions" is chosen so that all T_i are reduced to points.

A more general definition of block groups which will cover cases where the "m-system of functions" does not reduce all cuts to points and which is identical to that of (5.2) and (5.3) when all cuts are necessarily points is given by (5.4) and (5.5).

DEFINITION 5.4. *The closed block group $\bar{B}_\lambda$ consists of the union of all $\bar{S}_i$ with i in λ.*

DEFINITION 5.5. *The block group B_λ consists of the complement of $\bar{B}_{C(\lambda)}$ where $C(\lambda)$ is the complement of λ with respect to the integers $(1, 2, \cdots, n + 1)$.*

According to the representation theorem in [3], we have a continuous joint distribution of variates $U_1, U_2, \cdots, U_m$. By means of monotone functions $g_1(U_1), \cdots, g_m(U_m)$ this continuous distribution is mapped into a discontinuous distribution identical to the distribution of $\psi_1(w_1), \cdots, \psi_m(w_m)$.

Let $S_1' = \{(U_1, \cdots, U_m) \mid U_1 > u_1(i_{(1)})\}$,

$$S_2' = \{(U_1, \cdots, U_m) \mid U_1 < u_1(i_{(1)}),\ U_2 > u_2(i_{(2)})\},$$

$$\cdot$$
$$\cdot$$
$$\cdot$$

$$S_m' = \{(U_1, \cdots, U_m) \mid U_1 < u_1(i_{(1)}), \cdots, U_m > u_m(i_{(m)})\},$$

$$S_{m|n+1}' = \{(U_1, \cdots, U_m) \mid U_1 < u_1(i_{(1)}), \cdots, U_m < u_m(i_{(m)})\}.$$

Also we have:

$$S_1^* = \{g_1(U_1), \cdots, g_m(U_m) \mid g_1(U_1) > g_1(u_1(i_{(1)}))\},$$

$$S_2^* = \{g_1(U_1), \cdots, g_m(U_m) \mid g_1(U_1) < g_1(u_1(i_{(1)})),\ g_2(U_2) > g_2(u_2(i_{(2)}))\},$$

$$\cdot$$
$$\cdot$$
$$\cdot$$

$$S_{m|n+1}^* = \{g_1(U_1), \cdots, g_m(U_m) \mid g_1(U_1) < g_1(u_1(i_{(1)})), \cdots, g_m(U_m) < g_m(u_m(i_{(m)}))\},$$

and $\bar{S}_1^*$, $\bar{S}_2^*$, etc., are defined as S_1^*, S_2^*, etc., where $<$ is replaced by $\leqq$ and $>$ is replaced by $\geqq$.

Consider now the inverse mapping of the sets $S_1^*, S_2^*, \cdots$ and $\bar{S}_1^*, \bar{S}_2^*, \cdots$ into the space of $(U_1, U_2, \cdots, U_m)$. We shall have

$$g^{-1}(S_i^*) \subset S_i' \subset g^{-1}(\bar{S}_i^*)$$

because

$$g_i(U_i) > g_i(a) \rightarrow U_i > a \rightarrow g_i(U_i) \geqq g_i(a).$$

Thus we have the following inequality for the corresponding coverages:

$$\text{cov } (S_i^*) \leqq \text{cov } (S_i') \leqq \text{cov } (\bar{S}_i^*).$$

The proof follows directly from this relation as in section (9) of [3].

REFERENCES

[1] H. Scheffé and J. W. Tukey, "Nonparametric estimation I," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 187–192.

[2] J. W. Tukey, "Nonparametric estimation II," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 529–539.

[3] J. W. Tukey, "Nonparametric estimation III," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 30–39.