NON-PARAMETRIC ESTIMATION. I. VALIDATION OF ORDER STATISTICS

By H. Scheffé and J. W. Tukey

Syracuse University and Princeton University

1. Summary. Previous work on non-parametric estimation has concerned three problems: (*i*) confidence intervals for an unknown quantile, (*ii*) population tolerance limits, (*iii*) confidence bands for an unknown cumulative distribution function (*cdf*). For problem (*iii*) a solution has been available which is valid for any *cdf* whatever, but for (*i*) and (*ii*) it has heretofore been assumed that the population has a continuous probability density. This paper validates the existing solutions of (*i*) and (*ii*) assuming only a continuous *cdf*. It then modifies these solutions so that they are valid for any *cdf* whatever.

2. Introduction. There are three problems of non-parametric estimation (we exclude point-estimation) for which fairly satisfactory solutions are available; their present status was summarized in a recent paper [4]. The purpose of this series of articles is to extend and complete the theory of non-parametric estimation in directions of both theoretical and practical interest.

In this series we shall employ the following conventions of notation: We distinguish between a random variable and an arbitrary point in the Euclidean space containing its domain by using a capital Roman letter for the former and the corresponding lower case Roman letter for the latter. Thus if X is a (scalar) random variable, and x a real number or $\pm \infty$, we speak of the probability that $X \leq x$ and denote it by $Pr\{X \leq x\}$. Roman capitals will also be used to denote cumulative distribution functions[1] (*cdf*'s): A monotone non-decreasing function $F(x)$ will be called the *cdf* of X if $F(x + 0) = Pr\{X \leq x\}$. The definition of $F(x)$ at its points of discontinuity will be immaterial. Again, $E = (X_1, \cdots, X_n)$ will denote a random sample from a population with *cdf* $F(x)$, whereas $e = (x_1, \cdots, x_n)$ will denote a point in the sample space R_n. If t is a function of e only, $t = \varphi(e)$, then the random variable $T = \varphi(E)$ is a statistic. The *order statistics* of the sample E are defined to be $-\infty, Z_1, \cdots, Z_n, +\infty$, where $z_1 \leq z_2 \leq \cdots \leq z_n$ is a rearrangement of $x_1, x_2, \cdots, x_n$. We shall write $Z_0 = -\infty, Z_{n+1} = +\infty$. The device of including $+\infty$ and $-\infty$ among the order statistics will enable us to avoid special statements to cover the case of one-sided estimation. Confidence coefficients will be denoted by $1 - \alpha$. Finally, it will be convenient to symbolize[2] the following three classes of *cdf*'s: Ω_0 is the class of all univariate *cdf*'s F; Ω_2, the class of all continuous F; Ω_4, the class of all F with continuous derivative $F'(x)$.

[1] One of the authors wishes to point out the need of a clear, concise, and adequate term for this basic and important concept.

[2] The notation follows [3].

We now list the three problems. In each case it is understood that the solution sought is to be valid for all *cdf*'s in some chosen class. The names[3] associated with the problems are (*i*) W. R. Thompson, K. R. Nair, (*ii*) Wilks, (*iii*) Wald, Wolfowitz, Kolmogoroff.

(*i*) To find confidence intervals for an unknown quantile q_p, where q_p is defined by $F(q_p) = p$, $0 < p < 1$; in other words, to find statistics T_1, T_2 such that[4]

$$Pr\{T_1 \leq q_p \leq T_2 \mid F\} = 1 - \alpha. \tag{1.1}$$

(*ii*) To find tolerance limits T_1, T_2 which, with confidence $1 - \alpha$, will cover a proportion b or more of the population, that is,

$$Pr\{F(T_2) - F(T_1) \geq b \mid F\} = 1 - \alpha. \tag{1.2}$$

(*iii*) To find a confidence band for an unknown *cdf* F, that is, a random region $R(E)$ in the x,y-plane such that

$$Pr\{R(E) \text{ covers } g \mid F\} = 1 - \alpha, \tag{1.3}$$

where g is the graph of $y = F(x)$.

The existing solutions of problem (*iii*) are known to be valid for F in Ω_2, but those of problems (*i*) and (*ii*) have been validated only for F in Ω_4. The extension to F in Ω_2 is an immediate consequence of the theorem in section 4; this section also contains a discussion of some other implications of the theorem. In section 5 the appropriate modifications of the solutions of problems (*i*) and (*ii*) are found which extend their validity to the general case F in Ω_0. Whereas Pitman ([1]; also [4], p. 310) has shown how non-parametric tests may be extended to the possibly discontinuous case, the only solution of the three estimation problems previously extended to this case is that of Kolmogoroff for problem (*iii*). Extension from Ω_2 to Ω_0 is of considerable practical interest, not only in the case of populations ordinarily considered discrete, but also as affecting the problem of the finiteness of the number of significant figures in measurements and the resulting occurrence of "ties" in ranked measurements. Before making these extensions we discuss in the next section the transformations on which they are based.

3. Two useful transformations of random variables. We shall reserve the symbol X^* for a random variable having a uniform distribution on the interval from 0 to 1. Its *cdf* is

$$U(x^*) = Pr\{X^* \leq x^*\} = \begin{cases} 0 \text{ if } x^* < 0, \\ x^* \text{ if } 0 \leq x^* \leq 1, \\ 1 \text{ if } x^* > 1. \end{cases} \tag{1.4}$$

[3] For bibliography see [4].

[4] The notation $Pr\{R \mid F_0\}$ denotes the probability of the relation R being true, calculated under the assumption that the *cdf* of the population is $F_0(x)$.

The device of transforming from any random variable X with *cdf* F in Ω_2 to one with *cdf* U was early used by Karl Pearson and more recently by many others; it is known in the literature as the "probability integral transformation." We define the transformation $x^* = h_F(x)$ as follows: For $-\infty < x < +\infty$, $h_F(x) = F(x)$, $h_F(+\infty) = +\infty$, $h_F(-\infty) = -\infty$. If F is in Ω_2, the following statements are evident for the transform $X^* = h_F(X)$: X^* has $U(x^*)$ as its *cdf*. With $X_i^* = h_F(X_i)$, a random sample $E = (X_1, \cdots, X_n)$ from F transforms into a random sample $E^* = (X_1^*, \cdots, X_n^*)$ from U. The order statistics $\{Z_i\}$ of E transform into the order statistics $\{Z_i^*\}$ of E^* with $Z_i^* = h_F(Z_i)$, $i = 0, 1, \cdots, n + 1$.

It is easily seen that if F is not in Ω_2, the above transformation $Y = h_f(X)$ does not give Y the *cdf* U; indeed, if F is not in Ω_2, the *cdf* of any single-valued function Y of X is also not in Ω_2, for there will be at least one point $x = x_0$ with positive probability, and likewise for its transform y_0. Nevertheless our arguments in section 4 depend on relating a random variable with arbitrary *cdf* F in Ω_0 to the uniformly distributed X^*. While it is not possible to transform from X to X^*, without introducing a further random process, *it is possible to transform directly from* X^* *to* X. This suffices for our needs. We shall always denote this transformation by $X = g_F(X^*)$. The following definition of the function $x = g_F(x^*)$ makes it independent of the normalization of F at its discontinuities:

$$F(x - 0) \leq U(x^*) \leq F(x + 0). \tag{1.5}$$

A sketched diagram may aid the reader in following the argument: To every x^* $(-\infty \leq x^* \leq +\infty)$ there corresponds at least one x, and this x is unique unless it lies in an interval to which F assigns zero probability. In the latter case we shall assume that some x in the interval is designated to be $g_F(x^*)$. It will be seen that it is immaterial which x is thus chosen. However if $x = -\infty$ or $+\infty$ is in an interval of constancy of F we specify $g_F(-\infty) = -\infty$, $g_F(+\infty) = +\infty$.

To prove that $g_F(X^*)$ has the *cdf* $F(x)$ and thus can be identified with X, it is sufficient to prove that $Pr\{g_F(X^*) \leq x\} = F(x + 0)$. Now $g_F(X^*) \leq x$ if and only if $X^* \leq x_+^*$, where

$$x_+^* = \sup_{x=g_F(x^*)} x^*.$$

Hence $Pr\{g_F(X^*) \leq x\} = Pr\{X^* \leq x_+^*\} = U(x_+^*) = F(x + 0)$. It follows that a random sample E^* from U transforms into a random sample E from F. The transformation preserves the relation "$\leq$," that is, if $x_a = g_F(x_a^*)$, $x_b = g_F(x_b^*)$, then $x_a^* \leq x_b^*$ implies $x_a \leq x_b$. This means that the order statistics $\{Z_i^*\}$ of E^* transform into the order statistics $\{Z_i\}$ of E. We remark that $x_a^* < x_b^*$ does not imply $x_a < x_b$; there is trouble when $x_b^* \leq 0$ or $x_a^* \geq 1$, and more serious trouble if x_a^* and x_b^* both go into the same discontinuity of F. However, we shall need to utilize the fact that $x_a < x_b$ implies $x_a^* \leq x_b^*$.

4. Extension to continuous cdf's. A sufficient condition on T_1 and T_2 for a solution (1.2) of problem (*ii*) to be valid for all F in Ω_2 is clearly that the joint distribution of $F(T_1)$ and $F(T_2)$ be independent of F in Ω_2. If $Pr\{F(T_i) = p \mid F\} = 0$ $(i = 1, 2)$, then (1.1) is equivalent to

$$Pr\{F(T_1) \leq p \leq F(T_2) \mid F\} = 1 - \alpha, \tag{1.6}$$

and so a sufficient condition that a solution (1.1) of problem (*i*) be valid for all F in Ω_2 is again that the joint distribution of $F(T_1)$ and $F(T_2)$ be independent of F in Ω_2. We are thus led to consider sufficient conditions on a set $T_1, T_2, \cdots, T_r$ of statistics, which will insure that the joint distribution of $F(T_1)$, $F(T_2)$, $\cdots$, $F(T_r)$ be independent of F in Ω_2.

THEOREM: *A sufficient condition for the joint distribution of* $F(T_1), F(T_2), \cdots, F(T_r)$ *to be independent of* F *in* Ω_2 *is that the* $\{T_j\}$ *be a subset of the order statistics* $\{Z_i\}$ *of the sample.*

To prove the theorem it will suffice to show that the joint distribution of the set of n random variables $F(Z_1), F(Z_2), \cdots, F(Z_n)$ is independent of F in Ω_2. Let the *cdf* of the joint distribution be

$$G_F(\lambda_1, \lambda_2, \cdots, \lambda_n) = Pr\{F(Z_1) \leq \lambda_1, \cdots, F(Z_n) \leq \lambda_n \mid F\}. \tag{1.7}$$

Employing the transformation $x^* = h_F(x)$ discussed in section 3, we see that the above probability equals

$$Pr\{Z_1^* \leq \lambda_1, \cdots, Z_n^* \leq \lambda_n\}, \tag{1.8}$$

where $Z_0^*, Z_1^*, \cdots, Z_{n+1}^*$ are the order statistics of a random sample E^* from the uniform *cdf* U. But this probability does not depend on F.

Since the existing solutions of problems (*i*) and (*ii*) are obtained by taking T_1 and T_2 to be order statistics, we have validated these solutions for all F in Ω_2. That the existing solutions of problem (*iii*) are valid for F in Ω_2 has been demonstrated by their authors; this is however also an easy consequence of the above theorem. The sufficiency condition expressed by this theorem together with a necessity condition of Robbins' [2] may indicate a natural path to the formulation and solution of further problems of non-parametric estimation.

From a theoretical point of view it is of interest to note that even in those pathological cases where no probability density function exists for the *cdf* F in Ω_2 (F is non-absolutely continuous), the joint distribution (1.7) of $F(Z_1)$, $F(Z_2), \cdots, F(Z_n)$ always possesses a density. That this density is $n!$ for $0 \leq F(Z_1) \leq F(Z_2) \leq \cdots \leq F(Z_n) \leq 1$, and zero elsewhere, is evident if we consider (1.8). By "integrating out" the other variables we are led to the following practically useful result (it is well known for F in Ω_4): Choose any set $\{r_j\}$ of s integers $(1 \leq r_1 < r_2 < \cdots < r_s \leq n)$, and consider the joint distribution of $F(Z_{r_1}), F(Z_{r_2}), \cdots, F(Z_{r_s})$. This has a probability density function $f(t_1, t_2, \cdots, t_s)$, providing F is in Ω_2, given by the formula

$$(1.9) \qquad f(t_1, t_2, \cdots, t_s) = \frac{n! t_1^{r_1-1}(1-t_s)^{n-r_s}}{(r_1-1)!(n-r_s)!} \prod_{i=1}^{s-1} \frac{(t_{i+1}-t_i)^{r_{i+1}-r_i-1}}{(r_{i+1}-r_i-1)!}$$

for $0 \leq t_1 \leq t_2 \leq \cdots \leq t_s \leq 1$, and $f = 0$ elsewhere. As is conventional, the result of applying $\prod_{i=1}^{0}$ is to be interpreted as unity, and the meaning of f is given by

$$Pr\{F(Z_{r_i}) \leq a_i (i = 1, 2, \cdots, s) | F\} = \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \cdots \int_{-\infty}^{a_s} f(t_1, t_2, \cdots, t_s)\, dt_s \cdots dt_2\, dt_1.$$

5. Extension to discontinuous cdf's. Suppose we have a solution of problem (i) based on order statistics and hence valid for F in Ω_2, say,

$$(1.10) \qquad Pr\{Z_k \leq q_p \leq Z_t \mid F\} = 1 - \alpha,$$

where $0 \leq k < t \leq n + 1$. In particular this is valid for the uniform case,

$$(1.11) \qquad Pr\{Z_k^* \leq p \leq Z_t^*\} = 1 - \alpha.$$

We now transform from the uniform *cdf* U to an arbitrary F in Ω_0 by means of the transformation $x = g_F(x^*)$ described in section 3. Suppose q_p is defined by $q_p = g_F(p)$. This means the quantile q_p of the distribution with *cdf* F is determined from the relation

$$F(q_p - 0) \leq p \leq F(q_p + 0),$$

which assigns to the quantile its usual meaning if $F(x)$ is continuous and non-constant at $x = q_p$, and a sensible definition if F is discontinuous or constant at q_p. From the discussion in section 3 we have

$$(Z_k < q_p < Z_t) \text{ implies } (Z_k^* \leq p \leq Z_t^*) \text{ implies } (Z_k \leq q_p \leq Z_t),$$

and hence the probability relations

$$Pr\{Z_k < q_p < Z_t \mid F\} \leq Pr\{Z_k^* \leq p \leq Z_t^*\} \leq Pr\{Z_k \leq q_p \leq Z_t \mid F\}.$$

Substituting (1.11), we have

$$(1.12) \quad Pr\{Z_k < q_p < Z_t \mid F\} \leq 1 - \alpha \leq Pr\{Z_k \leq q_p \leq Z_t \mid F\}.$$

The statistical interpretation of (1.12) is the following: Consider any solution (1.10) of problem (i), giving a confidence interval for the quantile q_p, valid for F in Ω_2. Then with the same values of n, k, t, and α, the probability of the random interval from Z_k to Z_t covering the unknown quantile q_p is $\leq 1 - \alpha$ for the open interval, $\geq 1 - \alpha$ for the closed interval, no matter what the unknown *cdf* F. If F is continuous, the two probabilities are of course equal.

To extend the solution of problem (*ii*) to the general case F in Ω_0, suppose we have a solution (1.2) using order statistics, say $T_1 = Z_k$, $T_2 = Z_t$ $(0 \leq k < t \leq n + 1)$. Such a solution will be valid for all F in Ω_2, in particular for $F = U$,

$$Pr\{U(Z_t^*) - U(Z_k^*) \geq b\} = 1 - \alpha.$$

Given now any arbitrary distribution F, we again use the transformation $x = g_F(x^*)$. From (1.5),

$$F(Z_i - 0) \leq U(Z_i^*) \leq F(Z_i + 0) \qquad (i = k, t).$$

Hence

$$B_- \leq B^* \leq B_+,$$

where

$$B_- = F(Z_t - 0) - F(Z_k + 0),$$
$$B^* = U(Z_t^*) - U(Z_k^*),$$
$$B_+ = F(Z_t + 0) - F(Z_k - 0).$$

The implications

$$(B_- \geq b) \text{ implies } (B^* \geq b) \text{ implies } (B_+ \geq b)$$

yield the relations

$$Pr\{B_- \geq b\} \leq Pr\{B^* \geq b\} \leq Pr\{B_+ \geq b\}.$$

These may be written

$$(1.13) \quad Pr\{F(Z_t - 0) - F(Z_k + 0) \geq b \mid F\} \leq 1 - \alpha \leq Pr\{F(Z_t + 0) - F(Z_k - 0) \geq b \mid F\}$$

To interpret (1.13), let us say that a Borel set S covers a proportion π of a population with *cdf* $F(x)$ if $\int_S dF(x) = \pi$. If S is an interval from x' to x'', then the proportion covered by S is $F(x'' + 0) - F(x' - 0)$ if S is closed, and $F(x'' - 0) - F(x' + 0)$ if S is open. The proportion covered by a point x_0 is the jump $F(x_0 + 0) - F(x_0 - 0)$ of the *cdf* F at x_0. The statistical meaning of (1.13) is now clear: For the random interval from Z_k to Z_t, the probability that the open interval cover a proportion $\geq b$ of the population is $\leq 1 - \alpha$, the probability that the closed interval cover a proportion $\geq b$ of the population is $\geq 1 - \alpha$, regardless of the population. Again, for a continuous F the two probabilities are equal.

REFERENCES

[1] E. J. G. PITMAN, *Suppl. J. Roy. Stat. Soc.*, Vol. 4 (1937), pp. 117–130.
[2] H. ROBBINS, *Annals of Math. Stat.*, Vol. 15 (1944), pp. 214–216.
[3] H. SCHEFFÉ, *Annals of Math. Stat.*, Vol. 14 (1943) pp. 227–233.
[4] H. SCHEFFÉ, *Annals of Math. Stat.*, Vol. 14 (1943), pp. 305–332.