The Annals of Mathematical Statistics
1969, Vol. 40, No. 2, 644–647

NOTES

THE EXPECTED COVERAGE TO THE LEFT OF THE ith ORDER STATISTIC FOR ARBITRARY DISTRIBUTIONS

By Barry H. Margolin[1] and Frederick Mosteller

Yale University and Harvard University

1. Introduction. The *coverage* of the ith order statistic $X_{(i)}$, $i = 1, 2, \cdots, n$, in a sample of size n drawn from the continuous distribution F is $F(X_{(i)})$. The distribution of $F(X_{(i)})$ is well known ([2], p. 236) to be a beta distribution with parameters i and $n - i + 1$, and the expected coverage

$$E(F(X_{(i)})) = i/(n+1). \tag{1}$$

We want a definition of coverage of the ith order statistic that has expectation $i/(n + 1)$ in the general case where the parent distribution may have atoms.

A natural way to define coverage in the general case involves the Scheffé-Tukey transformation [1], described below, plus a special randomization when the ith ordered observation falls at an atom. This approach generates coverages distributed according to the same beta distribution as the usual coverages generated by samples from a continuous distribution. Instead of using this approach for the general case, we introduce below a *modified* definition of coverage that avoids randomization and nevertheless has expected coverage equal to $i/(n + 1)$. For a continuous parent distribution F, the modified definition agrees with the usual one; if the parent has at least one atom, the distribution of the modified coverage is not beta-distributed, but also has at least one atom.

2. The modified definition of coverage and its expectation. Let X be a random variable (whose distribution is continuous, discrete, or mixed), and let

$$\begin{aligned} F^-(x) &= \Pr\{X < x\}, \qquad F(x) = \Pr\{X \leqq x\}, \\ p(x) &= F(x) - F^-(x) = \Pr\{X = x\}, \quad V = \{x \mid p(x) > 0\}. \end{aligned} \tag{2}$$

In a random sample of size n from F, let $X_{(i)}$ be the ith ranked observation in ascending order of magnitude, so that $X_{(1)} \leqq X_{(2)} \leqq \cdots \leqq X_{(i)} \leqq \cdots \leqq X_{(n)}$. If there are ties in the sample, we may not be able to say which of the tied observations is the ith, only that it lies in a particular clump.

Received 30 October 1967; revised 29 July 1968.

[1] The first author's research was supported in part by the Army, Navy, Air Force and NASA under a contract administered by the Office of Naval Research. The second author's research was facilitated by a National Science Foundation grant (GS-341). Reproduction in whole or in part is permitted for any purpose of the United States Government.

For given values of i and n, suppose that in the sample

$$\begin{aligned} &T_i \text{ observations have values less than } X_{(i)}, \\ (3)\qquad &W_i \text{ observations have values equal to } X_{(i)}, \\ &n - T_i - W_i \text{ observations have values greater than } X_{(i)}. \end{aligned}$$

T_i may take on the values $0, 1, \cdots, i-1$, and consequently W_i can take on the values $i - T_i, \cdots, n - T_i$.

DEFINITION. The modified coverage of $X_{(i)}$ for a sample of size n is defined as

$$C_i(X_{(i)}, T_i, W_i) = F^-(X_{(i)}) + (i - T_i)(W_i + 1)^{-1}p(X_{(i)}) \tag{4}$$

where T_i and W_i are described in (3).

We use $X_{(i)} \varepsilon V$ to mean that the value of $X_{(i)}$ is an atom. If $X_{(i)} \varepsilon V$, then since $1 \leqq i - T_i \leqq W_i$

$$F^-(X_{(i)}) < C_i(X_{(i)}, T_i, W_i) < F(X_{(i)}). \tag{5}$$

Note that if in a sample $X_{(i)} \not\varepsilon V$, then $p(X_{(i)}) = 0$, $F^-(X_{(i)}) = F(X_{(i)})$, $W_i = 1$ and $T_i = i - 1$ with probability 1, and the modified coverage $C_i(X_{(i)}, T_i, W_i) = F(X_{(i)})$, the usual coverage for the continuous case. In a random sample of size n, we have the

THEOREM.

$$E(C_i(X_{(i)}, T_i, W_i)) = i/(n + 1). \tag{6}$$

PROOF. Our proof uses the Scheffé-Tukey ([1], p. 189) transformation which we now describe. Let X^* be a random variable having a uniform distribution on the interval from 0 to 1. Let U denote the cumulative distribution function of X^*, i.e., if x^* is a value of X^*

$$\begin{aligned} U(x^*) &= 0 && \text{if } x^* < 0 \\ &= x^* && \text{if } 0 \leqq x^* \leqq 1 \\ &= 1 && \text{if } 1 < x^*. \end{aligned}$$

Recall that F is the cdf of the random variable X. Consider the transformation $X^* \to g_F(X^*)$ such that

$$F(g_F(X^*) - 0) \leqq U(X^*) \leqq F(g_F(X^*) + 0).$$

Observe that if $F^-(x) < x^* \leqq F(x)$, then $g_F(x^*) = x$, where $x \varepsilon V$. Scheffé and Tukey observed that to every x^*, $-\infty \leqq x^* \leqq \infty$, there corresponds at least one $g_F(x^*)$ and that this $g_F(x^*)$ is unique unless it lies in an interval to which F assigns zero probability. In this case they (and we) assume that some value in the interval is designated to be $g_F(x^*)$; which value is immaterial for our purposes. Scheffé and Tukey proved that $g_F(X^*)$ has the cdf F and can thus be identified with the random variable X.

A random sample $X_1^*, \cdots, X_n^*$ from U transforms into a random sample $X_1, \cdots, X_n$ from F. For fixed i, consider those samples from U in which:

T_i observations are less than or equal to $F^-(X_{(i)})$,

W_i observations have values in the interval $(F^-(X_{(i)}), F(X_{(i)})]$,

$n - T_i - W_i$ observations are greater than $F(X_{(i)})$,

$$T_i = 0, \cdots, i-1, \qquad W_i = i - T_i, \cdots, n - T_i,$$

i.e., for these samples the ith order statistic from the uniform sample, $X_{(i)}^*$, falls in the half-open interval $(F^-(X_{(i)}), F(X_{(i)})]$. The conditional distribution of $X_{(i)}^*$, given $X_{(i)}$, T_i, W_i for $X_{(i)} \varepsilon V$ is that of the $(i - T_i)$th order statistic of a sample of size W_i from a uniform distribution on the interval $(F^-(X_{(i)}), F(X_{(i)})] = (F^-(X_{(i)}), F^-(X_{(i)}) + p(X_{(i)})]$. Thus from the well-known theorem on order statistics from the uniform distribution, the expected value of $X_{(i)}^*$, given $X_{(i)}$, T_i, W_i, for $X_{(i)} \varepsilon V$, is

$$(7) \quad E(X_{(i)}^* \mid X_{(i)}, T_i, W_i) = F^-(X_{(i)}) + (i - T_i)(W_i + 1)^{-1} p(X_{(i)}) = C_i(X_{(i)}, T_i, W_i).$$

This is obviously true as well for $X_{(i)} \not\varepsilon V$. We conclude that

$$(8) \quad E(C_i(X_{(i)}, T_i, W_i)) = E(E(X_{(i)}^* \mid X_{(i)}, T_i, W_i)) = E(X_{(i)}^*) = i/(n+1).$$

COROLLARY 1. *If the distribution F has at least one atom then*

$$(9) \quad E(F^-(X_{(i)})) < i/(n+1) < E(F(X_{(i)})).$$

PROOF. This follows from the strict inequalities of (5).

COROLLARY 2.

$$(10) \quad (i+1)^{-1} p_* P_i(V) \leqq E(F(X_{(i)})) - i(n+1)^{-1} \leqq (n-i+1)(n-i+2)^{-1} p^* P_i(V)$$

where $P_i(V) = \Pr\{X_{(i)} \varepsilon V\}$, $p_* = \inf_{x \varepsilon V} p(x)$, *and* $p^* = \sup_{x \varepsilon V} p(x)$.

PROOF. For $X_{(i)} \varepsilon V$,

$$F(X_{(i)}) - C_i(X_{(i)}, T_i, W_i) = [1 - ((i - T_i)/(W_i + 1))] p(X_{(i)}),$$
$$T_i = 0, \cdots, i-1, \quad W_i = i - T_i, \cdots, n - T_i,$$

and for $X_{(i)} \not\varepsilon V$

$$F(X_{(i)}) - C_i(X_{(i)}, T_i, W_i) = 0.$$

Hence for all $X_{(i)}$

$$F(X_{(i)}) - C_i(X_{(i)}, T_i, W_i) = [1 - ((i - T_i)/(W_i + 1))] p(X_{(i)}) I_V(X_{(i)})$$

where $I_V(x)$ is the indicator for the set V.

Let $R_i = W_i + 1 - (i - T_i)$. Then as W_i goes from $i - T_i$ to $n - T_i$, R_i goes from 1 to $n - i + 1$. Now

$$1 - ((i - T_i)/(W_i + 1)) = 1 - ((i - T_i)/(i - T_i + R_i)),$$

and is monotonically increasing in both R_i and T_i. Hence

$$1/(i + 1) \leqq 1 - ((i - T_i)/(W_i + 1)) \leqq (n - i + 1)/(n - i + 2).$$

Therefore,

$$(i + 1)^{-1}p_*I_V(X_{(i)}) \leqq F(X_{(i)}) - C_i(X_{(i)}, T_i, W_i)$$
$$\leqq (n - i + 1)(n - i + 2)^{-1}p^*I_V(X_{(i)}).$$

Taking expectations gives

$$(i + 1)^{-1}p_*P_i(V) \leqq E(F(X_{(i)})) - i/(n + 1)$$
$$\leqq (n - i + 1)(n - i + 2)^{-1}p^*P_i(V).$$

COROLLARY 3.

$$(11) \quad (n - i + 2)^{-1}p_*P_i(V) \leqq i/(n + 1) - E(F^-(X_{(i)}))$$
$$\leqq i(i + 1)^{-1}p^*P_i(V).$$

PROOF. Similar to that for Corollary 2.

REMARK. Corollaries 2 and 3 are probably more useful for the special case of a discrete distribution F, for which $P_i(V) = 1$, than for the mixed case.

3. Acknowledgment. The authors wish to express appreciation for discussions with Paul Holland and I. R. Savage and for suggestions from the referees.

REFERENCES

[1] SCHEFFÉ, H. and TUKEY, J. W. (1945). Non-parametric estimation. I. Validation of order statistics. *Ann. Math. Statis.* **16** 187–192.

[2] WILKS, SAMUEL S. (1962). *Mathematical Statistics*. Wiley, New York.