The Annals of Statistics
1974, Vol. 2, No. 6, 1293–1295

A NOTE ON OUTLIER-PRONE FAMILIES OF DISTRIBUTIONS[1]

BY RICHARD F. GREEN

University of California, Riverside

It is shown that (k, n)-outlier-proneness of a family of distributions is equivalent to complete outlier-proneness.

In their paper, "Outlier proneness of phenomena and of related distributions," Neyman and Scott (1971) offer definitions of "outlier" and of "outlier proneness." They show that the family of gamma distributions is outlier-prone completely, as is the family of lognormal distributions. On the other hand, the family of Cauchy distributions is not outlier-prone but is outlier-resistant.

Neyman and Scott define outlier-prone completely in terms of the seemingly weaker condition (k, n)-outlier-prone. In this note it is shown that (k, n)-outlier-proneness is equivalent to complete outlier-proneness.

The following definitions are those given by Neyman and Scott.

Let $S_n = (x_1, x_2, \cdots, x_n)$ be a sample of independent, identically distributed random variables from a distribution F. Let $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ be the ordered values. That is, $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(n)}$.

DEFINITION 1. For a positive number k we shall say that $x_{(n)} \in S_n$ is a k-outlier on the right if its value exceeds that of $x_{(n-1)}$ by more than $k(x_{(n-1)} - x_{(1)})$.

Let $P(k, n \mid F)$ denote the probability that a sample S_n of observations from a distribution F will contain a k-outlier.

Let $\mathscr{F}$ be a family of distributions and let $\pi(k, n \mid \mathscr{F})$ stand for the least upper bound of probabilities $P(k, n \mid F)$ for $F \in \mathscr{F}$.

DEFINITION 2. If $\pi(k, n \mid \mathscr{F}) < 1$ then we shall say that the family $\mathscr{F}$ is (k, n)-outlier-resistant. Otherwise, that is, if $\pi(k, n \mid \mathscr{F}) = 1$, we shall say that the family $\mathscr{F}$ is (k, n)-outlier-prone.

DEFINITION 3. If a family of distributions $\mathscr{F}$ is (k, n)-outlier-prone for all $k > 0$ and all $n > 2$, we shall say that $\mathscr{F}$ is outlier-prone completely.

THEOREM 1. *The family of distributions $\mathscr{F}$ is outlier-prone completely if and only if it is (k, n)-outlier-prone for some $k > 0$, $n > 2$.*

PROOF. That $\mathscr{F}$ is outlier-prone completely means that it is (k, n)-outlier-prone for all $k > 0$, $n > 2$. If $\mathscr{F}$ is (k, n)-outlier-prone for all $k > 0$, $n > 2$, it clearly is (k, n)-outlier-prone for some $k > 0$, $n > 2$.

Received September 1971; revised July 1973.

[1] This paper was prepared with the partial support of NIH Research Grant No. GM-10525, National Institutes of Health, Public Health Service.

AMS 1970 *subject classifications.* Primary 62E15; Secondary 62G30.

Key words and phrases. Outlier, (k, n)-outlier-prone, outlier-prone completely.

Further, if $\mathscr{F}$ is (k_0, n)-outlier-prone for a particular $k_0 > 0$, it will also be (k, n)-outlier-prone for all k such that $0 < k < k_0$.

Therefore, to prove the theorem it suffices to prove three facts for $k > 0$, $n > 2$, namely,

(1) $\mathscr{F}$ being (k, n)-outlier-prone implies that $\mathscr{F}$ is $(k, n + 1)$-outlier-prone.
(2) $\mathscr{F}$ being $(k, 2n)$-outlier-prone implies that $\mathscr{F}$ is (k, n)-outlier-prone.
(3) $\mathscr{F}$ being $(k, 3)$-outlier-prone implies that $\mathscr{F}$ is $(2k, 3)$-outlier-prone.

PROOF OF (1). Assume that $\mathscr{F}$ is (k, n)-outlier-prone. For any $\varepsilon > 0$ there must exist an $F \in \mathscr{F}$, call it F_0, such that $P(k, n \mid F_0) > 1 - \varepsilon/(n + 1)$. Consider a sample S_{n+1} from F_0. The probability that a random subsample of size n from S_{n+1} will have a k-outlier is $> 1 - \varepsilon/(n + 1)$. Therefore the probability of all samples of size n from S_{n+1} having k-outliers is $> 1 - \varepsilon$. But if all samples of size n from S_{n+1} have k-outliers then S_{n+1} itself has a k-outlier. Thus $P(k, n + 1 \mid F_0) > 1 - \varepsilon$ and $\mathscr{F}$ is $(k, n + 1)$-outlier-prone.

PROOF OF (2). Assume that $\mathscr{F}$ is not (k, n)-outlier-prone. Then there exists an $\varepsilon > 0$ such that for any $F \in \mathscr{F}$, $P(k, n \mid F) \leqq 1 - \varepsilon$. Consider two independent samples of size n from F. These can be combined to produce a sample S_{2n}. If both the samples of size n fail to have k-outliers then the combined sample will fail to have a k-outlier as well. Therefore the following inequalities hold for any $F \in \mathscr{F}$:

$$1 - P(k, 2n \mid F) \geqq (1 - P(k, n \mid F))^2 \geqq \varepsilon^2 .$$

Therefore,

$$P(k, 2n \mid F) \leqq 1 - \varepsilon^2 ,$$

and $\mathscr{F}$ is not $(k, 2n)$-outlier-prone.

PROOF OF (3). Assume that $\mathscr{F}$ is $(k, 3)$-outlier-prone. Pick any $\varepsilon > 0$ and show that there exists an $F \in \mathscr{F}$ such that $P(2k, 3 \mid F) > 1 - \varepsilon$. Let $N = [6/\varepsilon] + 1$, $\varepsilon_0 = 3\varepsilon/N^3$. Pick $F \in \mathscr{F}$ such that $P(k, 3 \mid F) > 1 - \varepsilon_0$. Take a sample S_N from F. All subsamples of size 3 from S_N will have k-outliers with probability $\geqq 1 - \binom{N}{3}\varepsilon_0 > 1 - \varepsilon/2$.

Order the points in S_N and consider the probability that a subsample of size 3 will have its largest two values adjacent values from the ordered sample. This probability is $3/N < \varepsilon/2$.

But

$$1 - P(2k, 3 \mid F) < \binom{N}{3}\varepsilon_0 + 3/N < \varepsilon , \qquad \text{or} \qquad P(2k, 3 \mid F) > 1 - \varepsilon ,$$

and $\mathscr{F}$ is $(2k, 3)$-outlier-prone.

This completes the proof of the theorem.

The significance of this result lies in the fact that the strength of complete outlier-proneness of a family of distributions does not come from the requirement that outliers be likely from samples of arbitrary size (arbitrary n), or from the

requirement of arbitrarily wild outliers (arbitrarily large k), but rather from the requirement that for some sample size n and some outlier index k an outlier will occur with arbitrarily high probability less than one.

REFERENCE

NEYMAN, J. and SCOTT, E. L. (1971). Outlier proneness of phenomena and of related distributions. *Optimizing Methods in Statistics*. Academic Press, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
RIVERSIDE, CALIFORNIA 92502