

OUTLIER-PRONE AND OUTLIER-RESISTANT DISTRIBUTIONS
(OUTLIER-PRONE DISTRIBUTIONS)

BY RICHARD F. GREEN

Technical Report No. 22

Department of Statistics
University of California
Riverside, California 92502

January, 1975

OUTLIER-PRONE AND OUTLIER-RESISTANT DISTRIBUTIONS

(OUTLIER-PRONE DISTRIBUTIONS)

BY RICHARD F. GREEN

University of California, Riverside

1. Introduction. A number of tests have been suggested for detecting outliers. Usually these tests are based on the assumption that the observed data is normally distributed, with or without normal contamination. Neyman and Scott (1971) have raised the possibility that certain distributions might be likely to produce "outliers" as they are implicitly defined by the tests to measure them.

Let S_n be a sample of size n of independent observations of random variables with common distribution F . Let the variable values be denoted X_1, X_2, \dots, X_n and let the ordered values be $X_{n1}, X_{n2}, \dots, X_{nn}$. Then Neyman and Scott call X_{nn} a k -outlier if $X_{nn} - X_{nn-1} > k(X_{nn-1} - X_{n1})$. Following Neyman and Scott we have the following definition.

DEFINITION 1. A family of distributions \mathcal{F} will be said to be outlier-prone completely if for each $\epsilon > 0$, $k > 0$, $n > 2$ there exists a distribution $F_{\epsilon, k, n} \in \mathcal{F}$ such that for a sample S_n of size n from F

(1) $P(X_{nn} - X_{nn-1} > k(X_{nn-1} - X_{n1})) > 1 - \epsilon.$

If a family of distributions is not outlier-prone it will be said to be outlier-resistant.

Neyman and Scott show that the family of gamma distributions is outlier-prone as is the family of

lognormal distributions.

Unfortunately, the idea of outlier-proneness in this form does not apply to individual distributions. This is shown in the following theorem.

THEOREM 1. No single distribution (that is, no one-member family of distributions) can be outlier-prone completely.

PROOF. Let F be any distribution. Assume F is not unitary since if it were (1) would not hold for any $\epsilon \leq 1$.

If F is not unitary then there exist disjoint closed intervals $A = [a_1, a_2]$, $B = [b_1, b_2]$, where $a_1 \leq a_2 < b_1 \leq b_2$, such that $P(A) > 0$, $P(B) > 0$.

Let $\delta = \min(P(A), P(B))$.

Then, for $\epsilon = 3\delta^3 > 0$, $k = (b_2 - b_1)/(b_1 - a_2) > 0$, $n = 3$, we have

$$P(X_{33} - X_{32} > k(X_{32} - X_{31})) < 1 - \epsilon$$

and thus F does not itself comprise a completely outlier-prone family. \square

It might be noted here that no family consisting of a finite number of distributions can be outlier-prone completely.

The concern, however, is with individual distributions since the observations of interest will be considered to come from some distribution and not from some family of distributions. In the next section definitions are given which apply the ideas of outlier-proneness and outlier-resistance to individual distributions. Gnedenko's (1943)

definitions, his law of large numbers for maxima and his theorem on relative stability of maxima are quoted.

In the third section of this paper theorems are given connecting the definitions of outlier-proneness and outlier-resistance to the classical laws of large numbers for maxima given by Gnedenko. In the fourth section conditions on the density sufficient for outlier-proneness or outlier-resistance are given. Finally, a classification of distributions according to their outlier properties is suggested.

2. Preliminaries. Throughout this paper the following two assumptions will be made about the distribution function $F(x)$:

ASSUMPTION 1. $F(\infty) = 1$.

ASSUMPTION 2. $F(x) < 1$ for all finite x .

We now define outlier-prone and outlier-resistant for distributions.

DEFINITION 2. A distribution F will be said to be absolutely outlier-resistant if for all $\epsilon > 0$ we have

$$P(X_{nn} - X_{nn-1} > \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

DEFINITION 3. A distribution F will be said to be relatively outlier-resistant if for all $k > 1$ we have

$$P(X_{nn}/X_{nn-1} > k) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

DEFINITION 4. A distribution F will be said to be absolutely outlier-prone if there exist constants $\epsilon > 0$ and $\delta > 0$ and an integer n_0 such that

$$P(X_{nn} - X_{nn-1} > \epsilon) \geq \delta$$

for any integer $n \geq n_0$.

DEFINITION 5. A distribution F will be said to be relatively outlier-prone if there exist constants $k > 1$, $\delta > 0$ and an integer n_0 such that

$$P(X_{nn}/X_{nn-1} > k) \geq \delta$$

for any integer $n \geq n_0$.

In his 1943 paper Gnedenko defined a law of large numbers for maxima and relative stability for maxima.

DEFINITION 6. The sequence of successive maxima (2)

$$X_{11}, X_{22}, \dots, X_{nn}, \dots$$

of a sequence of independent random variables

$$X_1, X_2, \dots, X_n, \dots$$

having distribution F is said to obey the law of large numbers if there exists a sequence of constants $\{A_n\}$ such that for any $\epsilon > 0$

$$P(|X_{nn} - A_n| < \epsilon) \rightarrow 1 \text{ as } n \rightarrow \infty.$$

DEFINITION 7. The sequence of successive maxima (2) will be called relatively stable if there exists a sequence of constants $\{B_n\}$ such that for all $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_{nn}}{B_n} - 1\right| < \epsilon\right) = 1.$$

Gnedenko proved the following two theorems.

THEOREM 2. For the sequence (2) to obey the law of large numbers when Assumption 2 holds it is necessary and sufficient that for all $\epsilon > 0$

$$\lim_{x \rightarrow \infty} \frac{1 - F(x+\epsilon)}{1 - F(x)} \rightarrow 0.$$

THEOREM 3. For the sequence (2) to be relatively stable when Assumption 2 holds it is necessary and sufficient that for all $k > 1$

$$\lim_{x \rightarrow \infty} \frac{1 - F(kx)}{1 - F(x)} \rightarrow 0.$$

3. Theorems on outlier-proneness and outlier-resistance. Geffroy (1958) showed that if the sequence (2) satisfies the law of large numbers, so does the sequence of $k+1$ st largest observations with the same constants. He also showed that if the sequence (2) is relatively stable then so is the sequence of $k+1$ st largest observations with the same constants.

From the results of Gnedenko and Geffroy the following two new theorems about outlier-resistance follow as corollaries.

THEOREM 4. Under Assumptions 1 and 2 the distribution F will be absolutely outlier-resistant if and only if for all $\epsilon > 0$

$$\lim_{x \rightarrow \infty} \frac{1 - F(x+\epsilon)}{1 - F(x)} = 0.$$

THEOREM 5. Under Assumptions 1 and 2 the distribution F will be relatively outlier-resistant if and only if for all $k > 1$

$$\lim_{x \rightarrow \infty} \frac{1 - F(kx)}{1 - F(x)} = 0.$$

COMMENT. If the members of a location family are absolutely outlier-resistant then the location parameter may be consistently estimated by the use of the maxima.

If the members of a scale family are relatively outlier-resistant then the scale parameter may be consistently estimated by the use of the maxima.

Absolute and relative outlier-proneness are equivalent to conditions on the distribution function similar to those of Gnedenko. This is shown in the following two theorems.

THEOREM 6. Under Assumptions 1 and 2 the distribution F will be absolutely outlier-prone if and only if there exist constants $\epsilon > 0$ and $\delta > 0$ such that

$$\frac{1 - F(x+\epsilon)}{1 - F(x)} \geq \delta \text{ for all finite } x.$$

PROOF. The proof consists of two parts.

(1) Assume that there exist constants $\alpha > 0$, $\beta > 0$, say α_0 , β_0 such that

$$\frac{1 - F(x+\beta_0)}{1 - F(x)} \geq \alpha_0 \text{ for all finite } x.$$

Now prove that there exist constants $\epsilon > 0$, $\delta > 0$ and an integer n_0 such that

$$P(X_{nn} - X_{nn-1} > \epsilon) \geq \delta \text{ for all } n \geq n_0.$$

Let $\epsilon = \beta_0$, $\delta = \frac{1}{2}e^{-1}\alpha_0^2$ and $n_0 = 2$.

Pick any integer $n \geq n_0$ and define

$$u_n = \sup(x: F(x) \leq 1 - 1/n). \text{ We have}$$

$$(3) \quad P(X_{nn} - X_{nn-1} > \beta_0) \geq nF(u_n)^{n-1}(1 - F(u_n + \beta_0)).$$

Consider the right-hand side of (3) factor by factor.

$$F(u_n) \geq 1 - 1/n \text{ by definition of } u_n.$$

$$F(u_n)^{n-1} \geq F(u_n)^n \geq (1 - 1/n)^n \geq \frac{1}{2}e^{-1}.$$

While

$$\frac{1 - F(u_n + \beta_0)}{1 - F(u_n)} \geq \alpha_0 \text{ and}$$

$$\frac{1 - F(u_n)}{1 - F(u_n - \beta_0)} \geq \alpha_0 \text{ by assumption.}$$

Therefore

$$\frac{1 - F(u_n + \beta_0)}{1 - F(u_n - \beta_0)} \geq \alpha_0^2.$$

But

$$1 - F(u_n - \beta_0) \geq 1/n, \text{ so}$$

$$n(1 - F(u_n + \beta_0)) \geq \alpha_0^2.$$

Therefore

$$P(X_{nn} - X_{nn-1} > \beta_0) \geq \frac{1}{2}e^{-1}\alpha_0^2 \text{ for all } n \geq n_0.$$

This completes the proof of part (1).

(2) Assume that for all $\alpha > 0$, $\beta > 0$ there exists an $x(\alpha, \beta)$, call it x_0 , such that

$$\frac{1 - F(x_0 + \beta)}{1 - F(x_0)} < \alpha.$$

Now prove that for any constants $\epsilon > 0$, $\delta > 0$ and any integer n_0 there exists an integer $n \geq n_0$ such that

$$P(X_{nn} - X_{nn-1} > \epsilon) < \delta.$$

Let $\beta = \epsilon$, $\alpha = \frac{\delta}{2} \min(1/24, \delta/4, 1/n_0)$, and let x_0 satisfy

$$\frac{1 - F(x_0 + \beta)}{1 - F(x_0)} < \alpha.$$

Let

$$n = \left\lceil \frac{\delta}{2(1 - F(x_0 + \beta))} \right\rceil$$

and show that $n \geq n_0$ and that

$$P(X_{nn} - X_{nn-1} > \varepsilon) < \delta.$$

First, $n \geq n_0$, since

$$1 - F(x_0 + \beta) < \frac{1 - F(x_0 + \beta)}{1 - F(x_0)} < \frac{\delta}{2n_0}, \text{ or}$$

$$n_0 < \frac{\delta}{2(1 - F(x_0 + \beta))}, \text{ while}$$

$$n = \left\lceil \frac{\delta}{2(1 - F(x_0 + \beta))} \right\rceil.$$

Now show

$$P(X_{nn} - X_{nn-1} > \varepsilon) < \delta.$$

We have

$$(4) \quad P(X_{nn} - X_{nn-1} > \varepsilon) \leq P(X_{nn} > x_0 + \beta) \\ + P(X_{nn} \leq x_0) + P(X_{nn-1} \leq x_0, X_{nn} > x_0).$$

Consider the right-hand side of (4) term by term.

$$P(X_{nn} > x_0 + \beta) = 1 - P(X_{nn} \leq x_0 + \beta) \\ = 1 - (F(x_0 + \beta))^n \\ = 1 - (1 - (1 - F(x_0 + \beta)))^n \\ < n(1 - F(x_0 + \beta)).$$

Therefore,

$$P(X_{nn} > x_0 + \beta) < n(1 - F(x_0 + \beta)) < \delta/2$$

by definition of n .

Next

$$P(X_{nn} > x_0) = (F(x_0))^n \\ = (1 - (1 - F(x_0)))^n < \exp(-n(1 - F(x_0))).$$

Finally

$$\begin{aligned}
 P(X_{nn-1} \leq x_0, X_{nn} > x_0) & \\
 &= n(F(x_0))^{n-1}(1 - F(x_0)) \\
 &= n(1 - F(x_0))(1 - (1 - F(x_0)))^{n-1} \\
 &< n(1 - F(x_0))\exp(-(n-1)(1 - F(x_0))) \\
 &\leq n(1 - F(x_0))\exp(-(n/2)(1 - F(x_0))).
 \end{aligned}$$

But

$$\begin{aligned}
 1 - F(x_0) &> \frac{1 - F(x_0 + \beta)}{\alpha} \text{ by assumption, and} \\
 n &> \frac{\delta}{4(1 - F(x_0 + \beta))} \text{ by definition of } n, \text{ so} \\
 n(1 - F(x_0)) &> \delta/4\alpha.
 \end{aligned}$$

But

$$\alpha = \frac{\delta}{2} \min(1/24, \delta/4, 1/n_0).$$

Therefore

$$n(1 - F(x_0)) > 2/\delta \text{ and } n(1 - F(x_0)) > 12.$$

We have, therefore,

$$\begin{aligned}
 \exp(-n(1 - F(x_0))) + n(1 - F(x_0))\exp(-(n/2)(1 - F(x_0))) \\
 < \delta/2.
 \end{aligned}$$

Thus

$$P(X_{nn} - X_{nn-1} > \varepsilon) < \delta. \quad \square$$

This completes the proof of part (2) and thus of the theorem.

There is an analogous theorem for relative outlier-proneness.

THEOREM 7. Under Assumptions 1 and 2 the distribution F will be relatively outlier-prone if and only if there exist constants $k > 1$, $\delta > 0$ such that

$$\frac{1 - F(kx)}{1 - F(x)} \geq \delta \text{ for all finite } x.$$

PROOF. Theorem 7 follows directly from Theorem 6 if the following transformation is made:

$$\begin{aligned} Y &= \log X \text{ for } X > 1, \\ &= 0 \text{ for } X \leq 1. \end{aligned}$$

Under this transformation F_X will be relatively outlier-prone if and only if F_Y is absolutely outlier-prone.

4. Conditions based on densities. The conditions for outlier-proneness and outlier-resistance used in Theorems 4 through 7 are conditions on the distribution function. For certain well-known distributions the density exists and is more likely to be used to represent the distribution than is the distribution function. In such cases it would be useful to have conditions for outlier-proneness and outlier-resistance based explicitly on the densities. Such conditions are given in the following two theorems.

THEOREM 8. If Assumptions 1 and 2 hold and if the density $f(x)$ exists then the conditions

- a) $f(x+\epsilon)/f(x) \rightarrow 0$ as $x \rightarrow \infty$ for all $\epsilon > 0$, and
- b) $f(kx)/f(x) \rightarrow 0$ as $x \rightarrow \infty$ for all $k > 1$

are sufficient for absolute and relative outlier-resistance respectively. These conditions are also necessary if we add the condition that the density has a monotone

right tail.

PROOF. The proof is given for absolute outlier-resistance. The argument for relative outlier-resistance is similar.

(1) Sufficiency.

Assume that

$$f(x+\epsilon)/f(x) \rightarrow 0 \text{ as } x \rightarrow \infty \text{ for all } \epsilon > 0.$$

Simply choose x_0 such that

$$f(x+\epsilon)/f(x) < \delta \text{ for } x \geq x_0.$$

Then we have

$$\begin{aligned} \frac{1 - F(x+\epsilon)}{1 - F(x)} &= \frac{\int_x^\infty f(z) dz}{\int_x^\infty f(z) dz} \\ &= \frac{\int_x^\infty f(y+\epsilon) dy}{\int_x^\infty f(y) dy} < \delta \text{ for } x \geq x_0. \end{aligned}$$

(2) Necessity.

For any $\epsilon_0 > 0$, let x lie in the monotone tail and define

$$a = F(x+\epsilon_0/2) - F(x),$$

$$b = F(x+\epsilon_0) - F(x+\epsilon_0/2),$$

$$c = 1 - F(x+\epsilon_0). \text{ Note: } b \leq a.$$

Now

$$\begin{aligned} \frac{f(x+\epsilon_0)}{f(x)} &\leq \frac{b}{a} \leq \frac{b+c}{a+c} \leq \frac{2(b+c)}{a+b+c} \\ &= 2 \frac{1 - F(x+\epsilon_0/2)}{1 - F(x)} \rightarrow 0 \text{ as } x \rightarrow \infty. \quad \square \end{aligned}$$

THEOREM 9. If Assumptions 1 and 2 hold and if the density $f(x)$ exists then for absolute and relative outlier-proneness the following conditions are sufficient respectively:

c) There exist constants $\epsilon > 0$, $\delta > 0$, x_0 such that

$$\frac{f(x+\epsilon)}{f(x)} \geq \delta \text{ for all } x \geq x_0, \text{ and}$$

d) there exist constants $k > 1$, $\delta > 0$, x_0 such that

$$\frac{f(kx)}{f(x)} \geq \delta \text{ for all } x \geq x_0.$$

These conditions are not necessary even if we add the condition that the density has a monotone right tail.

PROOF. The proof is similar to that for Theorem 8.

5. Classification of distributions according to their outlier properties. The ideas of absolute and relative outlier-resistance and outlier-proneness refer to the right tail of a distribution. (We could, of course, apply the same ideas to the left tail.) It is possible to classify distributions according to properties of their tails (right tails).

Any normal distribution is absolutely outlier-resistant and therefore also relatively outlier-resistant. Such a distribution (call it Class I) cannot be absolutely outlier-prone. A Cauchy distribution will be relatively outlier-prone (Class V) and therefore also absolutely outlier-prone. Such a distribution cannot be relatively outlier-resistant. An exponential distribution is absolutely outlier-prone and relatively outlier-resistant (Class III). There are six possible classes of distributions according to properties of the right tail. These classes are shown in Figure 1.

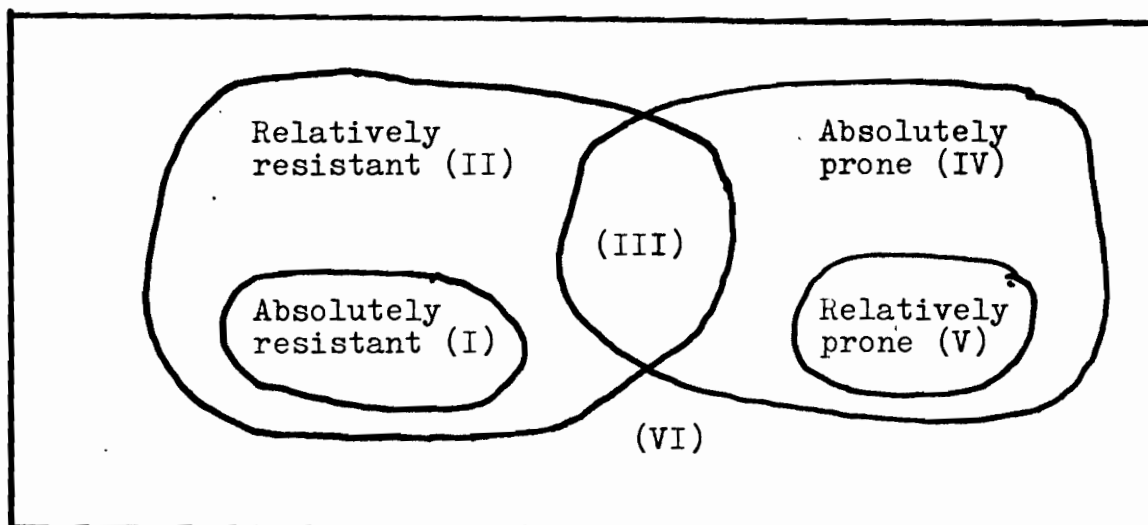


Figure 1 The six outlier classes.

The Poisson distribution belongs in Class II, being relatively resistant but neither absolutely resistant nor absolutely prone. It was necessary to construct examples of distributions belonging to Classes IV and VI.

EXAMPLE OF CLASS IV. Let X have the discrete distribution indicated by the following:

$$P(X = f(k)) = 2^{-k}, \quad k = 1, 2, 3, \dots \text{ where}$$

$$f(k) = 10^{\lfloor \log_2 k \rfloor} + k - 2^{\lfloor \log_2 k \rfloor}.$$

Then X is in Class IV.

EXAMPLE OF CLASS VI. Let Y have the discrete distribution indicated by the following:

$$Y = 2^k, \quad k = 0, 1, 2, \dots$$

$$P(Y = 2^k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad \lambda > 0.$$

Then Y is in Class VI.

REFERENCES

- GEFFROY, J. (1958 and 1959). Contributions à la théorie des valeurs extrêmes. Publ. Inst. Statist. Univ. Paris 7 and 8 37-185.
- GNEDENKO, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. Ann. of Math. 44 423-453.
- NEYMAN, J. and SCOTT, E. L. (1971). Outlier proneness of phenomena and of related distributions. Optimizing Methods in Statistics. Academic Press, New York.