

Outlier-prone Distributions

By

Richard Franklin Green

A.B. (University of Minnesota) 1963
M.A. (University of California) 1968

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

..... J. Neyman
..... E. L. Scott
..... Gordon F. Newell

Committee in Charge

..... June 16, 1973

OUTLIER-PRONE DISTRIBUTIONS *
ABSTRACT
Richard Franklin Green

The problem of what to do about outliers is one of the oldest in statistics. An outlier may be thought of as a value in a sample which seems too large or too small. There are many papers on outliers, some of which are mentioned in the articles by Anscombe (1960) and Grubbs (1969). There are two main strategies used in dealing with outliers:

(1) Test for them, usually assuming a single normal distribution which may or may not have normal contamination and throw out the outliers if they are detected.

(2) Use methods which are not sensitive to outliers.

These strategies have certain disadvantages. The assumption of normality may not be justified and the seeming outliers may be important to keep.

A generalization of this situation has been considered by Neyman and Scott (1971). The basic idea is that, with certain phenomena, "outliers" are a rule rather than an exception. In such cases, the attempt to "discover" the outliers and to eliminate them would mutilate the observational data.

Families of distributions that customarily fit

the observations on such phenomena have properties

*This investigation was partially supported by NIH Research Grant GM-10525, National Institutes of Health, Public Health Service.

described as outlier-proneness.

In the first section of this thesis Neyman and Scott's definitions of (k,n) -outlier-prone and completely outlier-prone families are given and their equivalence is proven.

In the next section a definition of outlier-prone family is given similar to that of Neyman and Scott. A theorem gives conditions on the distribution functions that are necessary and sufficient for the family of distributions to be outlier-prone.

The Neyman-Scott definition of outlier-proneness applies to families of distributions, not to individual members of such families. The third section of this thesis gives definitions of absolute and relative outlier-proneness and outlier-resistance which apply to individual distributions. Theorems are proved giving conditions on a distribution function necessary and sufficient for absolute and relative outlier-proneness and also for outlier-resistance. The conditions for absolute outlier-resistance are the same as those used by Gnedenko (1943) in proving a law of large numbers for maxima. Gnedenko's theorem on relative stability of maxima and the theorem here giving conditions for relative outlier-resistance are similar.

A classification of distributions according to their outlier properties is given in the next section and in the final section some propositions are stated which may help in the classification of continuous distributions when the densities are known. Examples of members of the six possible classes of distributions are given.

Acknowledgment. I want to thank Professor Jerzy Neyman for his support, stimulation and encouragement during the work on this thesis and during the better part of my time in Berkeley.

CONTENTS

0. Introduction.	p. 1
1. The equivalence of (k,n) -outlier-proneness and complete outlier-proneness.	p. 6
2. A theorem on outlier-prone families of distributions.	p. 10
3. Outlier-prone and outlier-resistant distributions.	p. 13
3.1. A theorem on absolute outlier-resistance.	p. 14
3.2. A theorem on absolute outlier-proneness.	p. 20
3.3. A theorem on relative outlier-resistance.	p. 26
3.4. A theorem on relative outlier-proneness.	p. 27
4. Classification of distributions according to their outlier properties.	p. 28
5. Conditions on densities.	p. 30
Bibliography	p. 35

Outlier-prone Distributions*

Ph.D. Dissertation, June 1973

by

Richard Franklin Green

Department of Statistics

University of California, Berkeley

*This investigation was partially supported by NIH
Research Grant GM-10525, National Institutes of
Health, Public Health Service.

0. Introduction. The problem of how to handle observations which seem too large or too small is a very old one. It is often impossible to recheck the data to see if mistakes were made. A number of methods have been suggested for dealing with outliers, as these seemingly anomalous observations are called. Many papers have been written on the subject of outliers, some of which are mentioned in articles by Anscombe (1960) and Grubbs (1969). Anscombe includes a brief history of the subject.

There are two main strategies used in dealing with outliers:

(1) Test for them, usually assuming a single normal distribution which may or may not have normal contamination, and throw out the outliers if they are detected.

(2) Use methods which are not sensitive to outliers.

The large amount of current research on "non-parametric" methods uses this second strategy but without being explicitly concerned with outliers.

Most work on outliers uses the first strategy. Various test statistics have been proposed to test for outliers. Dixon (1950) discusses a number of them. Usually the distributions of the test statistics have been calculated only for normal

distributions. Among the statistics mentioned by Dixon are the normed extreme deviation from the sample mean, the normed range and various ratios of order statistics.

Using this first strategy has certain disadvantages. The assumption of normality may not be justified and the seeming outliers may be important to keep.

A generalization of this situation has been considered by Neyman and Scott (1971). The basic idea is that, with certain phenomena, "outliers" are a rule rather than an exception. In such cases, the attempt to "discover" the outliers and to eliminate them would mutilate the observational data. Families of distributions that customarily are used to fit the observations on such phenomena have properties described as outlier-proneness. The idea is that in some cases it is likely that a point or points will occur that will be rejected as "outliers" by the usual test criteria.

The definition of outlier used by Neyman and Scott is similar to the ratio criterion developed by Dixon (1950, 1951, 1953) in a series of papers.

Neyman and Scott define when a family of distributions is (k,n) -outlier-prone. They define a family to be outlier-prone completely if it is outlier-prone for all $k > 0$ and all $n > 2$.

In the first part of this thesis it is shown that if a family is (k,n) -outlier-prone for some $k > 0$ and some $n > 2$, then it is outlier-prone completely.

In their paper Neyman and Scott show that the family of Gamma distributions is outlier-prone as is the family of lognormals. On the other hand, they show that location families and scale families are not outlier-prone. Thus, the family of Cauchy distributions is not outlier-prone but is outlier-resistant.

In the second part of this thesis a definition of outlier-prone family of distributions is given similar to that of Neyman and Scott. A theorem gives conditions on the distribution functions that are necessary and sufficient for the family of distributions to be outlier-prone. It says that outlier-proneness of a family is equivalent to saying that some member of the family will be arbitrarily spread out and that some member must be likely to produce samples where all observations are far apart.

The Neyman-Scott definition of outlier-proneness applies to families of distributions, not to individual members of such families. It is possible to consider one-member families but such families will not be outlier-prone. The third section of this thesis gives definitions of absolute and relative outlier-proneness and outlier-

resistance which apply to individual distributions. Theorems are proved giving conditions on a distribution function necessary and sufficient for absolute and relative outlier-proneness and also for outlier-resistance.

The conditions for absolute outlier-resistance are the same as those used by Gnedenko (1943) in proving a law of large numbers for maxima. The conditions for relative outlier-resistance are the same as Gnedenko used in proving the relative stability of maxima.

The condition Gnedenko uses for his law of large numbers for maxima was used earlier by von Mises (1923) to prove a similar theorem. Geffroy (1958, 1959) proved, among other things, that if the maximum satisfies the law of large numbers or is relatively stable, then the k^{th} largest order statistics satisfy the law of large numbers or are relatively stable respectively, with the same norming constants. These results, combined with those of Gnedenko (proved again by Geffroy and several other authors) make the present theorems on outlier-resistance corollaries, although the proofs here are direct. Many ideas useful in this general area are contained in Gumbel's book (1958).

A classification of distributions according to their outlier properties is given in the fourth section. In the final section some propositions

are stated which may help in the classification of continuous distributions when the densities are known. There are six possible classes of distributions according to their outlier properties. Examples of the six classes are given but in two cases it was necessary to invent the distributions.

1. The equivalence of (k,n)-outlier-proneness and complete outlier proneness. Let $S_n = (x_1, x_2, \dots, x_n)$ be a sample of independent, identically distributed random variables from a distribution F . Let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ be the ordered values. That is, $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Following Neyman and Scott we have:

DEFINITION 1.1. 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 $k(x_{(n-1)} - x_{(1)})$.

It is possible to define k -outlier on the left analogously. The work of Neyman and Scott and also of this thesis is concerned only with outliers on the right. Similar results, of course, hold for outliers on the left.

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

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

DEFINITION 1.2. If $\pi(k,n | \mathcal{F}) < 1$ then we shall say that the family \mathcal{F} is (k,n) -outlier-resistant. Otherwise, that is, if $\pi(k,n | \mathcal{F}) = 1$, we shall say that the family \mathcal{F} is (k,n) -outlier-prone.

DEFINITION 1.3. If a family of distributions \mathcal{F} is (k,n) -outlier-prone for all $k > 0$ and all $n > 2$, we shall say that \mathcal{F} is outlier-prone completely.

Since the definition of k -outlier depends on three order statistics it is meaningless to talk about (k,n) -outlier-proneness unless $n > 2$. The case that $k = 0$ is trivial since $P(0,n | F) = 1$ for all F when $n > 2$. The only meaningful cases are those where $k > 0$, $n > 2$.

The following theorem states the equivalence of (k,n) -outlier-proneness and complete outlier-proneness.

THEOREM 1.1. The family of distributions \mathcal{F} is outlier-prone completely if and only if it is (k,n) -outlier-prone for some $k > 0$, $n > 2$.

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

Further, if \mathcal{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) \mathcal{F} is (k,n) -outlier-prone implies \mathcal{F} is $(k,n+1)$ -outlier-prone.

PROOF OF (1). Assume \mathcal{F} is (k,n) -outlier-prone.

For any $\epsilon > 0$ there must exist an $F \in \mathcal{F}$, call it F_0 , such that $P(k,n | F_0) > 1 - \epsilon/(n+1)$. 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 $> 1 - \epsilon/(n+1)$. Therefore the probability of all samples of size n from S_{n+1} having a k -outlier is $> 1 - \epsilon$. But if all samples of size n from S_{n+1} have k -outliers then S_{n+1} itself has a k -outlier. Thus $P(k,n+1 | F_0) > 1 - \epsilon$.

(2) \mathcal{F} is $(k,2n)$ -outlier-prone implies \mathcal{F} is (k,n) -outlier-prone.

PROOF OF (2). Assume \mathcal{F} is not (k,n) -outlier-prone. Then there exists an $\epsilon > 0$ such that for any $F \in \mathcal{F}$, $P(k,n | F) \leq 1 - \epsilon$. 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:

$$1 - P(k,2n | F) \geq (1 - P(k,n | F))^2 \geq \epsilon^2.$$

Therefore,

$$P(k,2n | F) \leq 1 - \epsilon^2.$$

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

PROOF OF (3). Assume \mathcal{F} is $(k,3)$ -outlier-prone. Pick any $\epsilon > 0$ and show that there exists

an $F \in \mathcal{F}$ such that $P(2k, 3 | F) > 1 - \epsilon$. Let $N = \lceil 6/\epsilon \rceil + 1$, $\epsilon_0 = 3\epsilon/N^3$. (The square brackets indicate greatest integer function.) Pick $F \in \mathcal{F}$ such that $P(k, 3 | F) > 1 - \epsilon_0$. Take a sample S_N from F . All subsamples of size 3 from S_N will have k -outliers with probability $\geq 1 - \binom{N}{3}\epsilon_0 > 1 - \epsilon/2$.

Order the points 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 $3/N < \epsilon/2$.

But

$$1 - P(2k, 3 | F) < \binom{N}{3}\epsilon_0 + 3/N < \epsilon, \text{ or}$$

$$P(2k, 3 | F) > 1 - \epsilon.$$

This completes the proof of Theorem 1.1.

2. A theorem on outlier-prone families of distributions. The reason that families of distributions are considered is that experimental data are often fitted by a member of some family of distributions, such as the normal or lognormal or Gamma.

In a number of cases, for instance, in the case of nonzero amounts of daily rainfall considered by Neyman and Scott, the data are fitted by a family of distributions which are almost surely positive. In these cases it is simpler to use the ratio $x_{(n)}/x_{(n-1)}$ for a measure of outliers than $(x_{(n)} - x_{(n-1)})/(x_{(n-1)} - x_{(1)})$. The values of interest for this new measure of outliers will be $k > 1$ rather than $k > 0$.

The following theorem indicates how strong a condition outlier-proneness of a family is. It implies that outlier-proneness of a family is equivalent to stating that at least one member of the family will be arbitrarily spread out and that at least one member will be likely to produce samples in which all observations are relatively far apart.

DEFINITION 2.1. Let \mathcal{F} be a family of almost surely positive distributions. \mathcal{F} will be called positively outlier-prone if:

CONDITION 2.1. For any $n \geq 2$ and any constants

$k > 1$, $\alpha > 0$ there exists a distribution $F \in \mathcal{F}$ such that

$$(2.1) \quad P\left(\frac{X_{(n)}}{X_{(n-1)}} \geq k\right) > 1 - \alpha,$$

where $X_{(n)}$ is the largest and $X_{(n-1)}$ the next largest random variables in an independent sample of size n from distribution F .

THEOREM 2.1. Condition 2.1 is equivalent to each of the following two conditions:

CONDITION 2.2. For any constants $\epsilon > 0$, $c > 1$, there exists a distribution $G \in \mathcal{F}$ such that

$$(2.2) \quad G(cx) - G(x) < \epsilon \text{ for all } x > 0.$$

CONDITION 2.3. For any integer $m \geq 2$ and any constants $b > 1$ and $\beta > 0$ there exists a distribution $H \in \mathcal{F}$ such that for X_1, X_2, \dots, X_m independent random variables with distribution H ,

$$(2.3) \quad P\left(\min_{i=1, \dots, m-1} \frac{X_{(i+1)}}{X_{(i)}} \geq b\right) > 1 - \beta.$$

PROOF. (1) Condition 2.1 implies Condition 2.2. Assume that Condition 2.2 is false and show that Condition 2.1 is then false. Assume there exist constants $\epsilon > 0$, $c > 1$, such that for any $G \in \mathcal{F}$ there exists an x_0 such that

$$(2.4) \quad G(cx_0) - G(x_0) \geq \epsilon.$$

In Condition 2.1 let $n = 2$, $k = c$, $\alpha = \epsilon^2$. Then

$$(2.5) \quad P\left(\frac{X_{(2)}}{X_{(1)}} < k\right) \geq (G(cx_0) - G(x_0))^2 \geq \epsilon^2, \text{ or}$$

$$(2.6) \quad P\left(\frac{X(2)}{X(1)} \geq k\right) \leq 1 - \epsilon^2 = 1 - \alpha.$$

(2) Condition 2.3 implies Condition 2.1.

This is obvious.

(3) Condition 2.2 implies Condition 2.3.

Pick m , b and β as in Condition 2.3. (Pick $\beta < 1$.)

Use $c = b$ and $\epsilon = \beta/2(m-1)^2$ for Condition 2.2.

$$(2.7) \quad \begin{aligned} & P\left(\min_{i=1, \dots, m-1} \frac{X(i+1)}{X(i)} \geq b\right) \\ & > (1-2\epsilon)(1-4\epsilon)(1-6\epsilon)\dots(1-2(m-1)\epsilon) \\ & \geq (1-2(m-1)\epsilon)^{m-1} \geq 1 - 2(m-1)^2\epsilon \\ & = 1 - \beta. \end{aligned}$$

The first inequality in (2.7) can be seen by imagining X_1 given and looking at the conditional probability of X_2 being far away from X_1 : ($X_2/X_1 \geq b$ or $X_1/X_2 \geq b$). Then, given X_1 and X_2 far apart, look at the conditional probability of X_3 being far from X_1 and X_2 . This first conditional probability will be $> 1 - 2\epsilon$, the second will be $> 1 - 4\epsilon$, and so on.

This completes the proof of Theorem 2.1.

3. Outlier-prone and outlier-resistant distributions. In defining outlier-proneness for families, three factors are considered:

- k: outlier index,
- n: sample size, and
- $1-\alpha$: the probability of observing the desired outlier.

If a family of distributions is outlier-prone, then, for any $\alpha > 0$, $n \geq 2$ and $k > 1$, there must be a member of the family producing a k-outlier in S_n with probability greater than $1-\alpha$.

To obtain a definition of outlier-proneness that applies to individual distributions it is necessary to relax these conditions. The condition used here for outlier-proneness of individual distributions is that for some $\alpha < 1$ and some $k > 1$ and some integer N a sample S_n from the distribution will have a k-outlier with probability at least $1-\alpha$ for all $n \geq N$. The index used here for outliers is the relative difference between the two largest values, namely, $x_{(n)}/x_{(n-1)}$.

A distribution will be called outlier-resistant if for any $\alpha < 1$ and $k > 1$ there exists an integer N such that a sample S_n from the distribution will have a k-outlier with probability less than $1-\alpha$ for all $n \geq N$.

It is possible to consider the absolute difference $x_{(n)} - x_{(n-1)}$ as well as the relative

difference $x_{(n)}/x_{(n-1)}$. Absolute outlier-proneness and absolute outlier-resistance will be defined analogously to the relative outlier-proneness and relative outlier-resistance described above.

It should be noted that while Neyman and Scott define outlier-proneness of families as the complement of outlier-resistance (if a family isn't outlier-resistant it is outlier-prone) these definitions are different. A relatively outlier-resistant distribution cannot be relatively outlier-prone but a distribution need not be either one.

There are four theorems, each giving necessary and sufficient conditions for a given type of outlier-proneness or outlier-resistance of a distribution in terms of properties of its distribution function. The proofs of the theorems may be simplified by confining attention to continuous distributions. However, certain discrete distributions, particularly the Poisson, have interesting outlier properties and should be included in the theory.

3.1. A theorem on absolute outlier-resistance.

DEFINITION 3.1. A distribution F will be said to be absolutely outlier-resistant if for all $\epsilon > 0$ we have

$$(3.1.1) \quad P(X_{(n)} - X_{(n-1)} > \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty,$$

where $X_{(n)}$ and $X_{(n-1)}$ are the largest and next largest observations respectively from a collection of n independent random variables from distribution F .

In what follows two assumptions will be made.

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

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

CONDITION 3.1. $(1-F(x+\epsilon))/(1-F(x)) \rightarrow 0$ as $x \rightarrow \infty$ for all $\epsilon > 0$.

THEOREM 3.1. Under Assumptions 1 and 2 about distribution F , Definition 3.1 will hold if and only if Condition 3.1 holds.

PROOF. The idea of the proof is to first assume Condition 3.1 holds and show that for any non-zero distance it will be possible to find an integer such that for any sample size larger than that integer it will be possible to choose an interval such that the two largest observations will both lie in the interval. This is expressed mathematically in (3.1.5).

Then assume Condition 3.1 fails to hold and show that there exists a non-zero width and a non-zero probability such that for any integer there is some sample size larger than that integer and an interval of the given non-zero width such that the largest observation from the sample will lie above the interval

and the next largest observation will lie below the interval with the non-zero probability. This is expressed mathematically in (3.1.18).

(1) Condition 3.1 implies Definition 3.1.

Pick any $\delta > 0$ (assume $\delta \leq 1/6$), $\epsilon > 0$. It is necessary to show that there exists an integer n_0 (require that $n_0 \geq 2$) such that

$$(3.1.2) \quad P(X_{(n)} - X_{(n-1)} > \epsilon) < \delta \text{ for } n \geq n_0.$$

If condition 3.1 holds then for any $\alpha > 0$, $\beta > 0$, there exists a number x_0 such that

$$(3.1.3) \quad \frac{1 - F(x+\beta)}{1 - F(x)} < \alpha \text{ for } x \geq x_0.$$

Set $\beta = \epsilon/3$, $\alpha = \delta/2$.

Now let x_0 be a number such that

$$(3.1.4) \quad \frac{1 - F(x+\epsilon/3)}{1 - F(x)} < \delta/2 \text{ for } x \geq x_0.$$

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

Find some n_0 such that

$$u_{n_0} \geq x_0 + 2\beta.$$

Now consider any $n \geq n_0$. (Notice that $u_n \geq u_{n_0}$.)

Define. $p_{n,\epsilon} = P(X_{(n)} - X_{(n-1)} > \epsilon)$.

It is necessary to show that $p_{n,\epsilon} < \delta$.

Let $A = (u_n + \beta, \infty)$

$$B = [-\infty, u_n - 2\beta].$$

Define a and b as follows:

$$a/n = 1 - F(u_n + \beta) = P(A)$$

$$b/n = 1 - F(u_n - 2\beta) = 1 - P(B).$$

We have

$$(3.1.5) \quad p_{n,\xi} \leq P(X_{(n)} \in A) + P(X_{(n)} \in B) \\ + P(X_{(n-1)} \in B, X_{(n)} \notin B).$$

Consider the terms on the right side of

(3.1.5) one at a time. First,

$$(3.1.6) \quad P(X_{(n)} \in A) = 1 - P(X_{(n)} \notin A).$$

But

$$(3.1.7) \quad P(X_{(n)} \notin A) = (1 - a/n)^n > 1 - a,$$

so

$$(3.1.8) \quad P(X_{(n)} \in A) < a.$$

Next,

$$(3.1.9) \quad P(X_{(n)} \in B) = (1 - b/n)^n \leq e^{-b}.$$

Finally,

$$(3.1.10) \quad P(X_{(n-1)} \in B, X_{(n)} \notin B) = n(1 - b/n)^{n-1}b/n \\ = b(1 - b/n)^{n-1} \\ < be^{-b(n-1)/n} \\ < be^{-b/2}, \text{ since } n \geq n_0 \geq 2.$$

Therefore, (3.1.5) becomes

$$(3.1.11) \quad p_{n,\xi} < a + e^{-b} + be^{-b/2} \\ < a + 2be^{-b/2}.$$

But

$$(3.1.12) \quad 2be^{-b/2} < 1/b \text{ for } b \geq 12.$$

$$(3.1.13) \quad a = \frac{1 - F(u_n + \beta)}{1/n} < \delta/2, \text{ by Condition 3.1,}$$

(3.1.4) and choice of β , since $n \geq n_0$ and since $1 - F(u_n) \leq 1/n$.

By definition

$$b = \frac{1 - F(u_n - 2\beta)}{1/n}, \text{ whence}$$

$$(3.1.14) \quad 1/b = \frac{1/n}{1 - F(u_n - 2\beta)} \leq \frac{1 - F(u_n - \beta)}{1 - F(u_n - 2\beta)} < \delta/2,$$

but $\delta \leq 1/6$, so, from (3.1.11), (3.1.12), (3.1.13) and (3.1.14),

$$p_{n,\epsilon} < \delta/2 + \delta/2 = \delta \text{ for } n \geq n_0.$$

This completes the proof of (1).

Now prove

(2) If Condition 3.1 fails then Definition 3.1 also fails to hold.

Assume Condition 3.1 fails to hold. Then there exist constants $\alpha > 0$ (assume $\alpha < 1/2$), $\beta > 0$ such that for any x_0 there exists an $x \geq x_0$ such that

$$\frac{1 - F(x+\beta)}{1 - F(x)} \geq \alpha.$$

It is necessary to show that if Condition 3.1 fails then there exist constants $\epsilon > 0$, $\delta > 0$ such that for any integer n_0 there exists an integer $n \geq n_0$ such that

$$(3.1.15) \quad P(X_{(n)} - X_{(n-1)} > \epsilon) \geq \delta.$$

$$\text{Identify: } \epsilon = \beta, \delta = \frac{1}{2}e^{-1/\alpha}.$$

Since for $a > 0$

$$(1 - a/n)^n \rightarrow e^{-a} \text{ as } n \rightarrow \infty$$

we can consider n_0 large enough so that $n_0 \geq 2$

and

$$(3.1.16) \quad \left(1 - \frac{1/\alpha}{n}\right)^n > \frac{1}{2}e^{-1/\alpha} \text{ for all } n \geq n_0.$$

Pick $x_0 = u_{n_0}$. There exists an $x \geq x_0$

such that

$$(3.1.17) \quad \frac{1 - F(x+\beta)}{1 - F(x)} \geq \alpha.$$

Pick one such x , say x_1 .

Let n_1 be the largest integer n such that

$$u_n \leq x_1 + \beta.$$

Explicitly, $n_1 = \left[\frac{1}{1 - F(x_1 + \beta)} \right]$, where the heavy square brackets indicate the greatest integer function.

Clearly $n_1 \geq n_0$.

Define a and b as follows:

$$a/n_1 = 1 - F(x_1 + \beta), \text{ thus, } a \leq 1,$$

$$b/n_1 = 1 - F(x_1); \quad 1 - b/n_1 = F(x_1).$$

Thus we have that

$$(3.1.18) \quad P(X_{(n_1)} - X_{(n_1-1)} > \varepsilon) \geq n_1 (1 - b/n_1)^{n_1-1} a/n_1 \\ = a(1 - b/n_1)^{n_1-1}.$$

Factor the last expression and consider a and $(1 - b/n_1)^{n_1-1}$ separately. First

$$(3.1.19) \quad a/n_1 = 1 - F(x_1 + \beta) > 1/(n_1 + 1) \text{ by definition of } n_1.$$

Therefore $a > 1/2$. Next

$$(3.1.20) \quad (1 - b/n_1)^{n_1-1} > (1 - b/n_1)^{n_1}.$$

Now, (3.1.16) says

$$\frac{1 - F(x_1 + \beta)}{1 - F(x_1)} \geq \alpha, \text{ or, equivalently,}$$

$$(3.1.21) \quad \frac{1 - F(x_1)}{1 - F(x_1 + \beta)} \leq 1/\alpha.$$

But, by definition of a and b ,

$$(3.1.22) \quad \frac{b/n_1}{a/n_1} \leq 1/\alpha, \text{ so}$$

$$b \leq 1/\alpha \text{ since } a \leq 1.$$

Therefore

$$(3.1.23) \quad (1 - b/n_1)^{n_1} \geq (1 - \frac{1/\alpha}{n_1})^{n_1} > \frac{1}{2}e^{-1/\alpha} \text{ by}$$

choice of n_0 and n_1 . Finally,

$$P(X_{(n_1)} - X_{(n_1-1)} > \epsilon) \geq \frac{1}{2}e^{-1/\alpha} = \delta.$$

This completes the proof of (2) and also of Theorem 3.1.

3.2. A theorem on absolute outlier-proneness.

DEFINITION 3.2. 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

$$(3.2.1) \quad P(X_{(n)} - X_{(n-1)} > \epsilon) \geq \delta$$

for any integer $n \geq n_0$.

CONDITION 3.2. There exist constants $\epsilon > 0$ and $\delta > 0$ such that

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

Note. Condition 3.2 is equivalent to

CONDITION 3.2a. There exist constants ξ , $\delta > 0$ and x_0 such that

$$(3.2.2) \quad \frac{1 - F(x+\xi)}{1 - F(x)} \geq \delta \text{ for all } x \geq x_0.$$

Proof of note. Clearly, Condition 3.2 implies Condition 3.2a.

Assume Condition 3.2a holds for $\xi = \xi_1$, $\delta = \delta_1$, $x_0 = x_1$. Then,

$$\frac{1 - F(x_1 + \xi_1)}{1 - F(x_1)} \geq \delta_1, \text{ or}$$

$$1 - F(x_1 + \xi_1) \geq (1 - F(x_1))\delta_1 > 0.$$

Therefore,

$$(3.2.3) \quad \frac{1 - F(x + \xi_1)}{1 - F(x)} \geq (1 - F(x_1))\delta_1 \text{ for all } x < x_1$$

by monotonicity of F .

Thus, for all finite x :

$$(3.2.4) \quad \frac{1 - F(x + \xi_1)}{1 - F(x)} \geq (1 - F(x_1))\delta_1 > 0.$$

This completes the proof of the note.

THEOREM 3.2. Under Assumptions 1 and 2 about distribution F , Definition 3.2 holds if and only if Condition 3.2 holds.

PROOF. The proof of Theorem 3.2 is similar to that of Theorem 3.1. It consists of two parts. Part (1) here is similar to part (2) for Theorem 3.1 and part (2) here is similar to part (1) for Theorem 3.1. Part (2) for Theorem 3.2 is a bit more difficult than part (1) for Theorem 3.1

because the assumption that Condition 3.2 fails to hold is weaker than the assumption that Condition 3.1 holds.

Expression (3.2.7) here corresponds to key expression (3.1.18) in the proof of Theorem 3.1 and (3.2.24) here corresponds to (3.1.5).

(1) Condition 3.2 implies Definition 3.2.

Condition 3.2 says that for some $\alpha > 0$, $\beta > 0$, say α_0 , β_0 we have

$$(3.2.5) \quad \frac{1 - F(x + \beta_0)}{1 - F(x)} \geq \alpha_0 \text{ for all finite } x.$$

Assume Condition 3.2 holds and prove that there exist $\epsilon > 0$, $\delta > 0$ and an integer n_0 such that

$$(3.2.6) \quad P(X_{(n)} - X_{(n-1)} > \epsilon) \geq \delta \text{ for all } n \geq n_0.$$

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 consider the random variable X having distribution F . Define the events A and B :

$$A = (X > u_n + \beta_0),$$

$$B = (X \leq u_n).$$

Here again, $u_n = \sup\{x: F(x) \leq 1 - 1/n\}$.

We have

$$(3.2.7) \quad \begin{aligned} p_{n, \beta_0} &= P(X_{(n)} - X_{(n-1)} > \beta_0) \\ &\geq n(P(B))^{n-1}P(A) \\ &\geq n(P(B))^nP(A). \end{aligned}$$

Factor and consider $(P(B))^n$ and $nP(A)$

separately. First,

$$(3.2.8) \quad P(B) \geq 1 - 1/n, \text{ by definition of } u_n.$$

$$(3.2.9) \quad (P(B))^n \geq (1 - 1/n)^n \geq \frac{1}{2}e^{-1}, \text{ since } n \geq 2.$$

Next,

$$(3.2.10) \quad \frac{P(A)}{1 - F(u_n)} \geq \alpha_0, \text{ by Condition 3.2, and}$$

$$(3.2.11) \quad \frac{1 - F(u_n)}{1 - F(u_n - \beta_0)} \geq \alpha_0, \text{ by Condition 3.2.}$$

Combining (3.2.10) and (3.2.11) yields

$$(3.2.12) \quad \frac{P(A)}{1 - F(u_n - \beta_0)} \geq \alpha_0^2.$$

But,

$$(3.2.13) \quad 1 - F(u_n - \beta_0) \geq 1/n,$$

so

$$(3.2.14) \quad \frac{P(A)}{1/n} \geq \alpha_0^2, \text{ or}$$

$$(3.2.15) \quad nP(A) \geq \alpha_0^2.$$

Now, combining (3.2.7), (3.2.9) and (3.2.15) yields

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

all $n \geq n_0$.

This completes the proof of (1).

Now prove

(2) If Condition 3.2 fails then Definition 3.2 fails.

Condition 3.2 failing means that for any

$\alpha > 0$, $\beta > 0$ there exists an x such that

$$(3.2.16) \quad \frac{1 - F(x+\beta)}{1 - F(x)} < \alpha.$$

Assume Condition 3.2 fails and prove that Definition 3.2 fails to hold. That is, for any constants $\epsilon > 0$ and $\delta > 0$ and any integer n_0 there exists an integer $n \geq n_0$ such that

$$(3.2.17) \quad P(X_{(n)} - X_{(n-1)} > \epsilon) < \delta.$$

Assume $\epsilon, \delta > 0$ and $n_0 \geq 2$ are given.

$$\text{Let } \beta = \epsilon, \alpha = \frac{\delta}{2} \min\left(\frac{1}{24}, \frac{\delta}{4}, \frac{1}{n_0}\right).$$

Assuming that Condition 3.2 fails implies that for α, β chosen as above there exists an x , call it x_0 , such that

$$(3.2.18) \quad \frac{1 - F(x_0 + \beta)}{1 - F(x_0)} < \alpha.$$

Let

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

(Again, heavy square brackets denote greatest integer function.)

Show that $n \geq n_0$ and

$$(3.2.19) \quad p_{n, \epsilon} = P(X_{(n)} - X_{(n-1)} > \epsilon) < \delta.$$

First,

$$(3.2.20) \quad n \geq n_0, \text{ since}$$

$$(3.2.21) \quad 1 - F(x_0 + \beta) \leq \frac{1 - F(x_0 + \beta)}{1 - F(x_0)} < \frac{\delta}{2n_0}, \text{ so}$$

$$(3.2.22) \quad n_0 < \frac{\delta}{2(1 - F(x_0 + \beta))}, \text{ while}$$

$$(3.2.23) \quad n = \left\lceil \frac{\delta}{2(1 - F(x_0 + \beta))} \right\rceil.$$

Define

$$A = (x_0 + \beta, \infty),$$

$$B = [-\infty, x_0].$$

Define a and b as follows:

$$a/n = 1 - F(x_0 + \beta),$$

$$b/n = 1 - F(x_0).$$

As in the proof of the first part of Theorem

3.1 we have

$$\begin{aligned} (3.2.24) \quad p_{n,\epsilon} &\leq P(X_{(n)} \in A) + P(X_{(n)} \in B) \\ &\quad + P(X_{(n-1)} \in B, X_{(n)} \notin B) \\ &\leq a + e^{-b} + b(1 - b/n)^{n-1} \\ &\leq a + 1/b, \text{ when } b \geq 12. \end{aligned}$$

Now, show that $p_{n,\epsilon} < \delta$.

It suffices to show that $a \leq \delta/2$, $b \geq 2/\delta$, $b \geq 12$.

First,

$a/n = 1 - F(x_0 + \beta)$. Therefore, by definition of n ,

$$(3.2.25) \quad \frac{a}{\delta/2(1 - F(x_0 + \beta))} \leq 1 - F(x_0 + \beta), \text{ so}$$

$$(3.2.26) \quad a \leq \delta/2.$$

Next,

$$(3.2.27) \quad b/n = 1 - P(B) = 1 - F(x_0),$$

while, by definition of n , (3.2.20) and (3.2.22)

$$(3.2.28) \quad n > \frac{\delta}{4(1 - F(x_0 + \beta))}.$$

From (3.2.18) we have

$$(3.2.29) \quad \frac{1 - F(x_0)}{1 - F(x_0 + \beta)} > 1/\alpha, \text{ or}$$

$$(3.2.30) \quad 1 - F(x_0) > \frac{1 - F(x_0 + \beta)}{\alpha}.$$

Therefore, from (3.2.27), (3.2.28) and (3.2.30),

$$(3.2.31) \quad b > \frac{n(1 - F(x_0 + \beta))}{\alpha} \\ > \frac{\delta(1 - F(x_0 + \beta))}{4\alpha(1 - F(x_0 + \beta))} = \frac{\delta}{4\alpha},$$

but, by definition,

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

so, from (3.2.31)

$$(3.2.32) \quad b > \frac{\delta}{4} \left(\frac{2}{\delta} \cdot 24\right) = 12, \text{ and}$$

$$(3.2.33) \quad b > \frac{\delta}{4} \left(\frac{2}{\delta} \cdot \frac{4}{\delta}\right) = \frac{2}{\delta}.$$

Thus, $a \leq \delta/2$, $b \geq 2/\delta$, $b \geq 12$ and $p_{n,\epsilon} < \delta$.

This completes the proof of (2) and also of Theorem 3.2.

3.3. A theorem on relative outlier-resistance.

DEFINITION 3.3. A distribution F will be said to be relatively outlier-resistant if for all $k > 1$ we have

$$(3.3.1) \quad P(X_{(n)}/X_{(n-1)} > k) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

CONDITION 3.3. $(1 - F(kx))/(1 - F(x)) \rightarrow 0$ as $x \rightarrow \infty$ for all $k > 1$.

THEOREM 3.3. Under Assumptions 1 and 2 about distribution F, Definition 3.3 holds if and only if Condition 3.3 holds.

PROOF. Theorem 3.3 follows directly from Theorem 3.1 if the following transformation is

made:

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

Under Assumption 2 Definition 3.3 holds for F_X if and only if Definition 3.1 holds for F_Y . Condition 3.3 holds for F_X if and only if Condition 3.1 holds for F_Y .

3.4. A theorem on relative outlier-proneness.

DEFINITION 3.4. 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

$$(3.4.1) \quad P(X_{(n)}/X_{(n-1)} > k) \geq \delta$$

for any integer $n \geq n_0$.

CONDITION 3.4. There exist constants $k > 1$, $\delta > 0$ such that

$$(3.4.2) \quad \frac{1 - F(kx)}{1 - F(x)} \geq \delta \text{ for all finite } x.$$

Note. Condition 3.4 is equivalent to

CONDITION 3.4a. There exist constants $k > 1$, $\delta > 0$ and x_0 such that

$$\frac{1 - F(kx)}{1 - F(x)} \geq \delta \text{ for all } x \geq x_0.$$

THEOREM 3.4. Under Assumptions 1 and 2 Definition 3.4 holds if and only if Condition 3.4 holds.

PROOF. Theorem 3.4 follows directly from Theorem 3.2 if the same transformation is made

as suggested in the proof of Theorem 3.3.

4. Classification of distributions according to their outlier properties. The ideas of relative and absolute 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.) We thus have a way of classifying distributions according to properties of their tails (right tails).

Any normal distribution is both relatively and absolutely outlier-resistant (i). (The lower case Roman numerals refer to Figure 1.) A Cauchy distribution is both relatively and absolutely outlier-prone (v). An exponential distribution is absolutely outlier-prone and relatively outlier-resistant (iii). There are six possible classes of distributions according to properties of the right tail. These classes are shown in Figure 1.

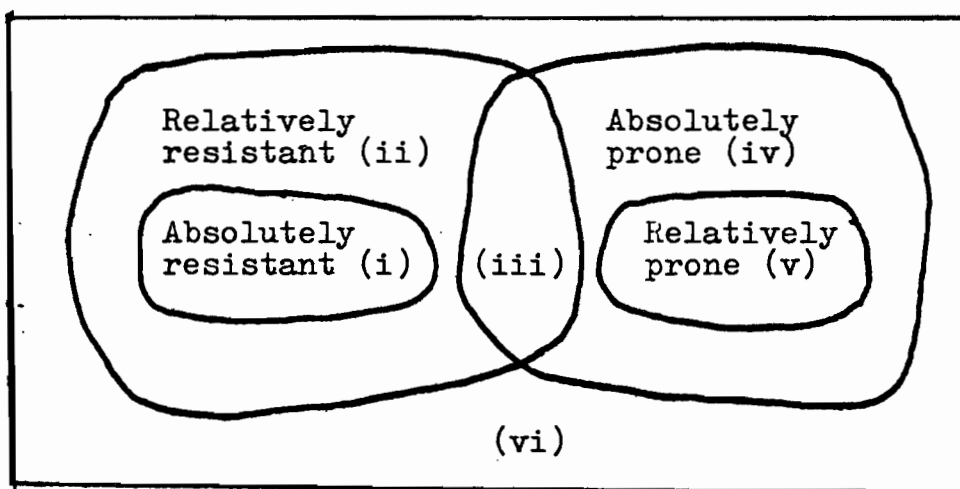


Figure 1 The six possible outlier classes

The Poisson distribution belongs in class (ii).
It was necessary to construct examples of distributions that belong to classes (iv) and (vi).

Example of class (iv). Suppose X takes the values shown in Table 1 with the probabilities listed. Then X is in class (iv).

Table 1.

	<u>Value</u>	<u>Probability</u>
1	{ 1	2^{-1}
2	{ 10	2^{-2}
	{ 11	2^{-3}
4	{ 100	2^{-4}
	{ 101	2^{-5}
	{ 102	2^{-6}
	{ 103	2^{-7}
8	{ 1000	2^{-8}
	{ 1001	2^{-9}

and so on.

Example of class (vi). Let

$$X = 2^k; k = 0, 1, 2, \dots$$

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

Then the distribution of X is in class (vi).

The purpose of Conditions 3.1, 3.2, 3.3 and 3.4 and Theorems 3.1, 3.2, 3.3 and 3.4 is to enable us to characterize distributions according to properties of their tails by looking at their distribution functions. In some cases we may be

given a density instead and we might like to have conditions similar to 3.1-3.4 in terms of densities.

5. Conditions on densities. We have the following sufficient conditions.

CONDITION 5.1. $f(x+\xi)/f(x) \rightarrow 0$ as $x \rightarrow \infty$ for all $\xi > 0$.

PROPOSITION 5.1. Condition 5.1 implies Condition 3.1.

CONDITION 5.2. There exist constants $\xi > 0$, $\delta > 0$, x_0 such that

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

PROPOSITION 5.2. Condition 5.2 implies Condition 3.2.

CONDITION 5.3. $f(kx)/f(x) \rightarrow 0$ as $x \rightarrow \infty$ for all $k > 1$.

PROPOSITION 5.3. Condition 5.3 implies Condition 3.3.

CONDITION 5.4. There exist constants $k > 1$, $\delta > 0$, x_0 such that

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

PROPOSITION 5.4. Condition 5.4 implies Condition 3.4.

PROOF OF PROPOSITION 5.1. Assume Condition 5.1 holds. Show that for any chosen $\epsilon, \delta > 0$ there exists an x_0 such that

$$(5.1) \quad \frac{1 - F(x+\xi)}{1 - F(x)} < \delta \text{ for } x \geq x_0.$$

Simply find x_0 such that

$$(5.2) \quad \frac{f(x+\epsilon)}{f(x)} < \delta \text{ for } x \geq x_0.$$

Then we have

$$(5.3) \quad \frac{1 - F(x+\epsilon)}{1 - F(x)} = \frac{\int_{x+\epsilon}^{\infty} f(z) dz}{\int_x^{\infty} f(z) dz}$$

$$= \frac{\int_x^{\infty} f(y+\epsilon) dy}{\int_x^{\infty} f(y) dy} < \delta \text{ for } x \geq x_0.$$

This completes the proof of Proposition 5.1.

The proofs of the other propositions are similar.

Conditions 5.1, 5.2, 5.3 and 5.4 are sufficient for Conditions 3.1, 3.2, 3.3 and 3.4 respectively, but they aren't necessary. Conditions 5.1 and 5.3 are necessary for Conditions 3.1 and 3.3 respectively if the following two assumptions are made.

ASSUMPTION 3. The distribution F has a density.

ASSUMPTION 4. The density of F has a monotone non-increasing right tail.

PROPOSITION 5.5. Under Assumptions 1, 2, 3 and 4 Condition 3.1 implies Condition 5.1.

Using the notation: $\bar{F}(x) = 1 - F(x)$,

Condition 3.1 says: $\frac{\bar{F}(x+\epsilon)}{\bar{F}(x)} \rightarrow 0$ as $x \rightarrow \infty$

for all $\epsilon > 0$, and

Condition 5.1 says: $\frac{f(x+\epsilon)}{f(x)} \rightarrow 0$ as $x \rightarrow \infty$

for all $\epsilon > 0$.

PROOF OF PROPOSITION 5.5. For any ϵ_0 picked

for Condition 5.1, use $\epsilon_0/2$ for ϵ in Condition 3.1. Start with x large enough so that we are in the (monotone) tail. The situation is illustrated in Figure 2.

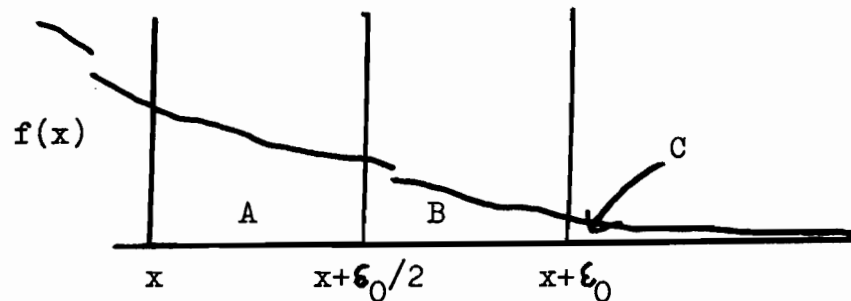


Figure 2 In the monotone tail

$$\text{Define } A = F(x + \epsilon_0/2) - F(x),$$

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

$$C = 1 - F(x + \epsilon_0). \quad \text{Note: } B \leq A.$$

$$\frac{f(x + \epsilon_0)}{f(x)} \leq \frac{B}{A} \leq \frac{B + C}{A + C} \leq \frac{2(B + C)}{A + B + C}$$

$$= 2 \frac{\bar{F}(x + \epsilon_0/2)}{\bar{F}(x)} \longrightarrow 0 \text{ as } x \longrightarrow \infty.$$

This completes the proof of Proposition 5.5.

PROPOSITION 5.6. Under Assumptions 1, 2, 3 and 4, Condition 3.3 implies Condition 5.3.

Condition 3.3 says: $\frac{\bar{F}(kx)}{\bar{F}(x)} \longrightarrow 0$ as $x \longrightarrow \infty$

for all $k > 1$, and

Condition 5.3 says: $\frac{f(kx)}{f(x)} \longrightarrow 0$ as $x \longrightarrow \infty$

for all $k > 1$.

PROOF. The proof of this proposition is similar to that of Proposition 5.

That Condition 3.1 and Assumptions 1, 2 and 3 are not sufficient for Condition 5.1 or that Condition 3.3 and Assumptions 1, 2 and 3 are not sufficient for Condition 5.3 is shown by the following

Example. Use the standard normal distribution $\Phi(x)$ to produce the new distribution as shown in Figure 3.

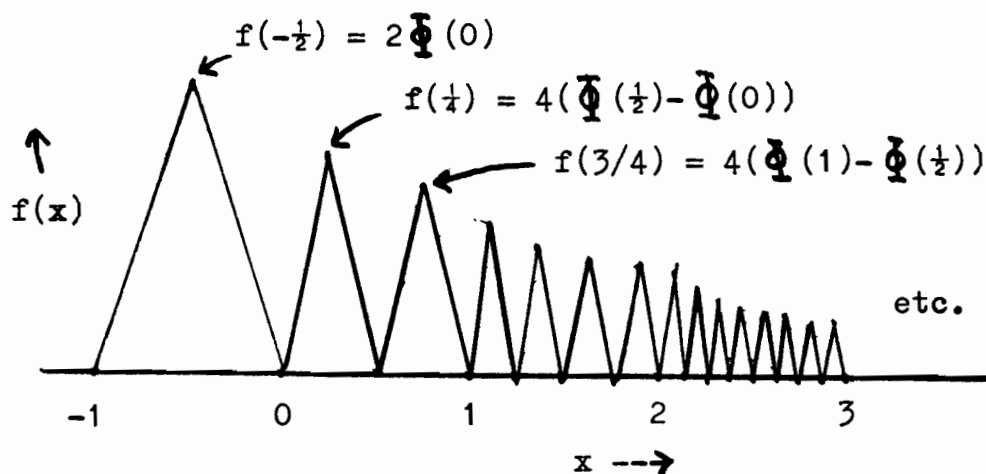


Figure 3 Saw-toothed normal

Each successive unit interval, beginning with $(-1, 0]$ is broken into twice as many subintervals as the preceding interval and the density takes the form of an isosceles triangle in each subinterval. The triangle with base $(a, b]$ will have height

$$\frac{2}{b-a}(\Phi(b) - \Phi(a))$$

except for $(-1, 0]$ which has height $2\Phi(0)$.

That Condition 3.2 and Assumptions 1, 2, 3 and 4 are not sufficient for Condition 5.2 or

that Condition 3.4 and Assumptions 1, 2, 3 and 4 are not sufficient for Condition 5.4 is shown by the following

Example. Let F be the distribution having density

$$f(x) = 0 \text{ for } x \leq 0,$$

$$f(x) = 2^{(-\frac{1}{2}k^2 + \frac{1}{2}k - 1)} \text{ for}$$

$$x \in \left(\sum_{i=1}^{k-1} 2^{(\frac{1}{2}i^2 - \frac{3}{2}i + 1)}, \sum_{i=1}^k 2^{(\frac{1}{2}i^2 - \frac{3}{2}i + 1)} \right]$$

$$k = 1, 2, 3, \dots$$

BIBLIOGRAPHY

- ANSCOMBE, F. J. (1960). Rejection of outliers.
Technometrics 2 123-147.
- DIXON, W. J. (1950). Analysis of extreme values.
Ann. Math. Statist. 21 488-506.
- DIXON, W. J. (1951). Ratios involving extreme values.
Ann. Math. Statist. 22 68-78.
- DIXON, W. J. (1953). Processing data for outliers.
Biometrics 9 74-89.
- 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.
- GRUBBS, F. E. (1969) Procedures for detecting outlying observations in samples. Technometrics 11 1-21.
- GUMBEL, E. J. (1958). Statistics of Extremes, Columbia University Press, New York.
- MISES, R. VON. (1923). Über die Variationsbreite einer Beobachtungsreihe. S.-B. Berlin Math. Ges. 22 3-8.

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