PNL-4690
UC-32

Applied Extreme Value Statistics

R. R. Kinnison

May 1983

Prepared for the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Operated for the U.S. Department of Energy
by Battelle Memorial Institute

Battelle

PNL-4690

DISCLAIMER

PACIFIC NORTHWEST LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC06-76RLO 1830

Printed in the United States of America
Available from
National Technical Information Service
United States Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22161

NTIS Price Codes
Microfiche A01

Printed Copy

Pages	Price Codes
001-025	A02
026-050	A03
051-075	A04
076-100	A05
101-125	A06
126-150	A07
151-175	A08
176-200	A09
201-225	A010
226-250	A011
251-275	A012
276-300	A013

PNL-4690
UC-32

APPLIED EXTREME VALUE STATISTICS

R.R. Kinnison
Senior Research Statistician

May 1983

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

PREFACE

Extreme value statistics is a statistical specialty that is seldom understood by researchers applying statistics to everyday problems. It is relevant to biological, engineering and environmental studies because often extreme or unusual conditions are more important than the usual conditions. For example, in carcinogenesis studies, the response to the maximum human dose of a chemical or the minimum level of radiation that could cause cancer is more important than the typical dose of the chemical or level of exposure. Extreme value statistical methods have been used to great advantage in hydrolic engineering and in architecture to predict floods or droughts, maximum wind gust force on buildings, and minimum breaking strength of materials.

Good high-level extreme value statistical theory books are available, such as those by Gumbel and Galambos. Most order statistics text also contain the mathematics of extreme value theory, and an occasional good article appears in journals of various special fields. However no text of applied methods for the professional without a statistical degree now exist. The basic concepts of extreme value statistics are simple, few in number, and have wide applicability. Extreme value statistics differ from 'ordinary' statistics more in the way data is collected than in data analysis. The data analysis aspects of extreme values use densities and distributions, estimators, probability plots, and many more statistical tools commonly found in all other types of statistical analysis. The level of presentation used here is for the scientist or engineer who uses statistics frequently, but who is not formally trained as a statistician.

With the wide use of automated data acquisition methods in the past few years, very large data sets have become common. Such data sets are troublesome to ordinary statistical methods because of the time needed to review and analize them, the size of computer required for storage and analysis, the accumulation of roundoff and truncation errors, and the

difficulty humans have in finding some kinds of data characteristics in charts and graphs of large numbers of data values. Extreme value statistics offers one way of simplifying massive amounts of data, by subdividing the data and analizing the extremes of the subdivisions. Some might argue that such a procedure throws away information, however the extremes cannot be found unless all the data values are examined. Such a procedure has a statistical advantage, the extremes of large sets or subsets of data have good statistical properties that are not strongly dependent upon the statistical properties of all the data.

This monograph has few examples and exercises. This is because the authors extreme value work has been with proprietary data, thus the examples and exercises had to be fabricated or taken from the literature. Contributions of data, examples, and exercises is solicited and material included in future editions will be acknowleged.

ACKNOWLEDGMENTS

This monograph is an extension, revision, and compilation of a series of tutorial papers prepared for limited distribution to radioecologists as TRAN-STAT issues 14-18. TRAN-STAT is supported by U.S. Department of Energy contract DE-AC06-76RL0 1830. Several of my associates at Pacific Northwest Laboratory provided encouragement and acted as editors for the initial series of papers. Special thanks are due Dr. Richard Gilbert who has contributed to all phases of the evolution of this work.

R.R.K.
January, 1983

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 PURPOSE

The statistical theory of extreme values is a well established part of theoretical statistics. Unfortunately, it is seldom part of applied statistics and is infrequently a part of statistical curricula except in advanced studies programs. This has resulted in the impression that it is difficult to understand and not of practical value. In recent environmental and pollution literature, several short articles have appeared with the purpose of documenting all that is necessary for the practical application of extreme value theory to field problems (for example, Roberts, 1979). These articles are so concise that only a statistician can recognise all the subtleties and assumptions necessary for the correct use of the material presented.

The intent of this text is to expand upon several recent articles, and to provide the necessary statistical background so that the non-statistician scientist can recognize an extreme value problem when it occurs in his work, be confident in handling simple extreme value problems himself, and know when the problem is statistically beyond his capabilities and requires consultation.

1.2 INTRODUCTION TO EXTREME VALUES

The purpose of the statistical theory of extreme values is to mathematically and logically explain observed extremes in samples of some specified size. In this text size of samples refers to the number of data points in a related group or set of values. It does not refer to the volume, weight or dimensions of the object being measured. The essential

conditions are that (1) the phenomena being measured is a statistical (stochastic) variable (what is commonly but erroneously called a parameter), (2) that the initial distribution from which the samples with extreme values have been drawn remains constant from one set of samples to the next (or that any change that occurs may be measured and a transformation of the data may be found to eliminate the effects of the change), and (3) that the observed extremes should be statistically independent. The literature is full of "practical rules" for dealing with the lack of independence, and claims of validity and lack of validity of these rules. Only through an understanding of some of the underlying statistical theory can the lack of independence be recognized and consistently managed. Environmental data is one of the most difficult kinds to analize for independence.

A literature search over the last 20 or so years will seem to indicate that there has been very little recent theoretical work by statisticians in extreme value problems. This is a not so, the study of this theory has simply been generalized and its name changed to order statistics. Order statistics is an extension from the study of the largest, or smallest, values of a sample to the study also of the second largest, and third largest, and so on. Extreme values are thus a subset of order statistics. A literature search on order statistics will yield a great deal of recent work and some fine contemporary textbooks (for example, David, H. A., 1970). Extreme values, being a special but important case of order statistics, are typically described in a chapter or two within such textbooks. A complete and rigorous study of extreme value statistics requires an understanding of order statistics in general. For the purpose of this text such a comprehensive understanding is not required, and the logical backround of the theory will be skipped and only those theorems and results that have practical application will be presented. It must be emphasized that this results in a 'cookbook' type presentation with its well known and real pitfalls.

The distributions of extremes may be characterized by certain statistics such as means, medians, modes, and a new statistic called the expected extreme. In ordinary statistics the common measure of central tendency is the mean because it has great advantages in most applied problems. In extreme value distributions the mode is preferred because it posseses advantages in extreme value problems. The initial distribution from which the samples containing the extremes are obtained and the size of these samples must be known in order to derive the exact extreme value theory for any specific problem. However, methods have been developed which require only a knowlege of sample size and the general type of initial distribution, and where forecasts are based exclusively on past observed extremes. Also, if the type of distribution is known and sample sizes are large, the asymptotic theory can be used. In practice, the asymptotic theory is almost exclusively used because it yields elegantly simple formulations for statistical tests on extreme values. This discussion of extreme value statistics will be concerned only with this asymptotic theory.

1.3 APPLICATIONS

1.3.1 Forecasting Floods

The prototype extreme value problem used by E. J. Gumbel (1941) was to predict annual floods. Hence, it is sometimes assumed that extreme value theory originated in hydrology. Section 1.4 of this chapter will explain that this is not the origin. However the study of floods was one of the early and very fruitful applications of the theory. The economic importance of accurately predicting floods has been realized since ancient times by agrarian societies. Today the Army Corps of Engineers is responsible for the management of rivers. Also, agriculture is dependent upon river management for both irrigation and avoidance of floods that destroy crops and soils.

Until the 1930's there were numerous attempts by engineers to find a mathematical formula for forecasting floods. In part their lack of success resulted from the endeavor to find mathematically exact solutions rather than statistical solutions. The statistical solutions were not then available. The engineers used instead arbitrary safety factors, such as double the largest flood that had occurred in the last 50 years. Such rules will, in the following chapters, be shown to be very conservative, and thus very costly to use as construction or design criteria.

Floods are the annual maxima of daily river discharges, and droughts are the annual minima. The analysis of droughts is essential in planning for irrigation, public health, and stream pollution. A key difference between floods and droughts is that droughts are bounded and floods are not. No matter how severe a flood one can always imagine a worse flood. But once a river runs dry there is no conceivable worse drought.

1.3.2 Environmental Pollution

Meterological phenomena are important in the study of air pollution. This is perhaps the field of study currently of greatest interest in extreme value theory studies. The major unknown, and the root of much controversy, is the relation of extreme pollutant concentrations to health effects in humans. The response of humans, or any biological system, to typical pollutant concentrations is itself an extreme value phenomena since only the few most sensitive persons respond. Thus environmental pollution can be conceived as a compounding of several extreme value distributions: those that describe when, where, and the magnitude of occurrence of extreme pollutant concentrations, and those that describe who will be where the maxima occurs and how they will respond to those maxima. At the present time there is no unified or general statistical theory for analyzing these compound problems. Each part must be treated as a separate and independent problem.

Currently in the medical and environmental literature there is a controversy about how to extrapolate toxicologic and carcinogenic data in order to predict dose-effect relationships at exposure levels much lower than those practical to use in laboratory studies. Linear extrapolation is often used to predict threshold levels. But the conclusion that saccarine is harmful in small amounts because it is harmful at very high doses has been challanged. There is the paradox that airplane travel is more hazardous than automobile travel because airplanes expose one to more cosmic rays than does automobile travel yet more people are killed in autos than in airplaines. These dilemmas should be recognized as arising from attempts to treat such phenomena deterministically rather than statistically. Even when it appears that statistics has been used, often the practitioners are unfamiliar with extreme value theory or unaware that they have an extreme value theory phenomena. This situation is analogous to that of the hydrologic engineers and dam builders before the use of extreme value theory to study floods. Recall that the deterministic study of floods yielded safety rules that were very conservative.

1.3.3 Strength of Materials

Two situations in which extreme value theory is being effectively used are to determine maximum wind gust and minimum breaking strength of materials. Both of these are important to aircraft designers and to architects of large buildings. Minimum strength of materials is important to all types of manufacturing from simple consumer products to heavy equipment. When something breaks, repair cost and down-time cost are usually substantial relative to the initial cost. Furthermore, human safety may be compromised. However, if too much extra strength is built into an item, an economic disadvantage results from the cost of excess materials.

1.3.4 Identifying Outlying Observations

The final application outlined, and the most statistical in nature, is the problem of identifying outlying observations. Every scientist has a favorite ad hoc "rule" for handling outliers that has advantages over those used by other scientists. Yet he is really somewhat uncomfortable with his rule, particularly when he reflect upon its logical and statistical foundations. Extreme value theory has much to contribute to the study of outliers, since an early motivation for statisticians to investigate extreme value problems was to identify outliers.

A major problem in identifying outliers in a data set, especially for small sample sizes, is that calculated means, standard deviations, and probabilities associated with some hypotheses are considerably influenced by the observed maxima and minima in the samples. These statistics are the basis for interpreting the data and for making forecasts, and such interpretations and forecasts should not be permitted to be erroneously influenced by invalid observations. On the other hand, the extremes may reflect important information. Perhaps they are a key to understanding the true principles governing the observed phenomena. Extreme value theory is the foundation of all sophisticated techniques for identifying outlying observations.

1.4 HISTORY

The first students of extreme value statistics were early astronomers who had the problem of deciding whether to accept or disregard a suspect (outlying) observation that appeared to differ greatly from the rest of a data set. Like many other statistical problems discovered by early astronomers, their mathematical tools were too crude to solve this problem. They can only be credited with the clear recognition and statement of the problem.

The modern history of extreme value statistics started in Germany in 1922 with a fundamental paper by L. von Bortkiewicz on the distribution of the range and the mean range in samples from a Gaussian (Normal) distribution as a function of sample size. These proved to be very difficult problems which were not solved in mathematical generality until recent times. Bortkiewicz found good numerical approximations, and called attention to the fact that the largest values of samples taken from Gaussian populations are new variables having separate distributions. Bortkiewicz thus deserves credit for being the first to clearly state the extreme value problem in statistical terms.

In the following year, 1923, R. von Mises, also in Germany, introduced the mathematically fundamental concept of the expected value of the largest member of a sample of observations. This was the start of the study of the asymptotic distribution of extreme values in samples from Gaussian distributions.

The founders of probability and statistical theory, such as Laplace, Pascal, Fermat, and Gauss, were too occupied with the general behavior of statistical masses to be interested in extreme values. The oldest remarks in the statistical literature about extreme values are perhaps those due to Fourier in 1824. He stated that for the Gaussian distribution, the probability of a deviation being more than 3 times the square root of 2 standard deviations from the mean is about 1 in 50,000, and the observation associated with this deviation could therefore be neglected. This seems to be the origin of the common but erroneous statistical "rule" that plus or minus 3 standard deviations from the mean should be considered the maximum range of valid sample values from a Gaussian distribution irrespective of the number of samples taken. In 1877, Helmert stated correctly that the probability of surpassing any specified value depends upon the size of the sample. The fallacy of the 3 standard deviations rule should be obvious. If the statistical distribution being sampled is unlimited, no matter how small the probability of the limits given by a rule, then the largest, or

smallest, sample value is also unlimited. As the sample size increases, the largest value encountered in a sample will likewise increase since there is more opportunity for inprobable values to occur. The statistical study of extreme values attempts to describe the relationship between sample size and magnitude of the observed extreme values. For small samples the "three sigma rule" is too conservative. For large samples it is too weak.

Largest values from distributions other than the Gaussian were first studied in 1923 by E. L. Dodd. A major step occurred in 1925 when L. H. C. Tippet published tables of the largest values and corresponding probabilities for various sample sizes from a Gaussian distribution, and the mean range of such samples (Tippet, L.H.C., 1925). In 1927 M. Frechet published, in a remote journal, the first paper to obtain the asymptotic distribution of the largest value from a class of individual distributions. The next year, 1928, R. A. Fisher and L. H. C. Tippet published the paper that is now considered the foundation of the asymptotic theory of extreme value distributions. They independently found Frechet's asymptotic distribution, and constructed two others. These three distributions have been found adequate to describe the extreme value distributions of all statistical distributions (Fisher, R. A., and Tippet, L.H.C., 1928). Fisher and Tippet, in this paper, were the first to stress the extremely slow convergence of the distribution of the largest value in samples from a Gaussian distribution toward its asymptote. Thus they showed the reason for the difficulties encountered by prior investigators.

The use of the Gaussian distribution as a starting point had hampered the development of the theory since none of the fundamental extreme value theorems are related in a simple way to the Gaussian distribution. It was reasonable to assume a Gaussian distribution for study purposes since this distribution is a foundation stone of much modern statistical reasoning. The theory of largest values ought to be based upon the Exponential distribution because it leads to simple development and expression of the

fundamental theorems of extreme value statistics. The results can then be generalized to other distributions.

The authors mentioned above were interested in extreme values only from the standpoint of statistical theory. In the middle 1930's E. J. Gumbel began studying the application of this theory, first in Germany, then in the U. S. when World War II engulfed Europe. Gumbel's first application was to old age, the consideration of the longest duration of life. He then showed that the statistical distribution of floods, long studied by engineers, can be understood using extreme value theory (Gumbel, 1941). These procedures have also been extensively applied to other meteorological phenomena, to stress and breaking strength of structural materials, and to the statistical problem of outlying observations.

1.5 SUMMARY

The history of extreme value statistics began late with respect to statistical history in general because early statisticians were concerned with the behavior of statistical masses rather than with the study of rare events. Fisher and Tippet (1928) made a major contribution by finding the asymptotic distributions of extremes. The application of extreme value theory began in the middle 1930's with the work of E. J. Gumbel. In contemporary times extreme value theory has become a part of the more generalized study of order statistics.

Several applications of current interest are discussed in this chapter. These include those classically associated with extreme values such as floods and the breaking strength of materials; applications which now use extreme value theory. Also discussed are some new applications in the biological and environmental sciences which currently do not yet use extreme value theory, but which have much to gain if this theory were effectively applied.

1.6 REFERENCES

David, H. A. 1970, Order Statistics, John Wiley & Sons, Inc.

Fisher, R. A., and Tippet, L.H.C. 1928, "Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample", Proc. Cambridge Phil. Soc., Vol. 24, pt. 2, pp. 180-190.

Gumbel,E. J., 1941, "The Return Period of Flood Flows", Ann. Math. Stat., Vol. 12, pp. 163-190.

Roberts,E. M., 1979, "Review of Statistics of Extreme Values with Applications to Air Quality Data", J. Air Poll. Control Assoc., Vol. 29, No. 6, pp. 632-637.

Tippet, L.H.C. 1925, "On the Extreme Individuals and the Range of Samples Taken from a Normal Distribution", Biometrika, Vol. 17, Pts. 3 and 4, pp. 364-387.

CHAPTER 2
DATA AND STATISTICS

2.1 INTRODUCTION

The research scientist first considering data as an extreme value situation usually finds several different procedures seem to be applicable. Only one procedure is usually applicable because of the 'scale' or 'level' of the measurements. The levels of the measurements inherent in a data set determine which statistics can be used with that data. Extreme value statistics can logically be divided into two types based upon scale of measurement. One type considers the number of extremes that occur. The second type considers the magnitude of extremes. In this division, the first type is relavent to what are called nominal scale measurements, and the second type to interval and ratio scale measurements. These scales are inherent characteristics of data and all statistical test procedures are valid only for particular data measurement scales. The ordinal scale, a fourth type of measurement, is not used in extreme value statistics, however it is important in the general study of order statistics. This chapter discusses these four types of measurement scales and reviews some statistical basics. This will provide part of the foundation needed for studying extreme value statistics. Readers already familiar with these topics can skip to chapter 3.

Associated with every statistical procedure is a mathematical model and some data. The procedure is valid under certain conditions or underlying assumptions. The model and measurement techniques specify these conditions. Even something as simple as calculating an average assumes that the average is a reasonable measure of the 'central tendency' of the population. Populations that are best modeled as skewed or multimodal have estimators of central tendency that are better than the common mean. All

statistical results inherently carry the qualification: 'If the model used was correct and if the measurement requirements were satisfied'. Such statistical assumptions and requirements are often violated in subtle but important ways. Statistics courses typically teach one to recognize model assumptions, but measurement requirements are rarely studied.

Mathematical models of data, called statistical distributions, and measurement theory are important for understanding extreme value statistics. Measurements and distributions in general are the topic of this chapter. The distributions unique to extreme value statistics are discussed in the following chapter. For extreme value work it is convenient to talk about statistical distributions in terms of the 'Exponential family', the 'Cauchy type', and distributions with limits. These descriptors are beyond the level of elementary statistics and will be explained in the next chapter.

2.2 MEASUREMENTS

Measurement characteristics can be divided into two parts; scale or level of measurement, and independence of observations. Correlation analysis is the usual technique for measuring independence, but correlation is not synonomous with independence. Independence is a population attribute, and correlation is a sample attribute. For small samples the correlation can be very different from the underlying dependence. However for large samples there is little practical difference. Since extreme value work typically uses large samples, this text will assume independence is well approximated by correlation. Observations are independent if the selection of one value from a population for inclusion in a sample does not influence the chances of any other value being selected for inclusion in the same sample. A common source of dependence in extreme value work comes from the use of serial data, such as a time series of data values. Daily maxima of pollutant concentrations are dependent (correlated) because the causes of pollution are phenomena that last for many days. Daily

temperatures are dependent because they are influenced by the length of day, which has an annual cycle. A later chapter will consider techniques for recognizing and eliminating data dependence.

2.3 LEVELS OF MEASUREMENT

Measurement is the activity of mapping or assigning numbers to objects or observations. Levels of measurement are a way of describing the characteristics of data obtained from measurements. the description contains information about the way data is collected and inherent characteristics of the things measured.

2.3.1 Nominal Scale

When numbers, or symbols, are used to identify groups or classes to which various objects belong, the scale of measurement is said to be nominal. The numbers are used only as a name for the group or category to which each observation belongs. In addition to group identification, nominal data typically includes a second part, a count of the number of items within each group. These counts are called frequencies. Frequencies can only assume integer values, there cannot be 2.5 persons in the Jones family.

Sports teams are identified by their home town. They could also be identified with 1, 2, ..., or 101, 102, ..., or A, B, C, ..., and so on. The identity of a team is a nominal scale data value and the number of players on a team is the corresponding frequency. Social security numbers can be considered as the group identification part of a nominal data value for which the frequency of each group is one. Arithmetic can be performed upon the frequencies but not on the group identification. The number of players on a football team is meaningful, but the sum of the numbers on thier jerseys contains no information.

2.3.2 Ordinal or Ranking Scale

The categories or classes into which objects are partitioned may stand in some kind of unmeasurable relationship to each other in addition to being identifiable as different categories. The essential feature of the ordinal scale is that the relative order of the objects or classes can be identified but not quantified. In a beauty contest first and second place contestants are identified, but one cannot say how much more beautiful the first place contestant is over the second place contestant. Ordinal relationships are typically assigned consecutive integer numbers for identification. These identifications are called ranks, 1, 2, 3, or first, second, third, and so on.

Air pollution indices are usually on an ordinal scale of measurement. Although such indices may appear to be more precise than ranks, they typically do not meet the requirements of the higher measurement scales that will be discussed next. A pollution index of 50 does not indicate that the air is twice as hazardous as air with an index of 25. The higher index indicates a more hazardous condition, but the magnitude of the difference cannot be quantitated.

An order preserving transformation of the category indices does not change the information contained in ordinal data. It does not make any difference whether the index 1 is assigned to last place, 2 to second from last, and on up; or the index 1 is assigned to first place with ranking downward.

The median is the statistic most appropriate for describing the 'central tendency' of measurements on an ordinal measurement scale. Sometimes the median value cannot be quantified; in a beauty contest the median is that contestant for which half the contestants are more beautiful and half are less beautiful.

2.3.3 Interval Scale

When measurements have all the characteristics of an ordinal scale, and in addition the interval sizes (distances) between objects or groups is measurable, the measurements are said to be at the interval level. An interval scale is characterized by a 'unit of measurement' which assigns a real number to the relationship (distances) between all pairs of objects or groups.

Temperature measurements are a good example of interval scale measurement. Fahrenheit, Celsius, and Kelvin scales are commonly used and these demonstrate the arbitrary nature of the zero point and the distances that are typical of an interval scale.

Any transformation or mathematical operation on interval scale data values must preserve not only the ordering of the objects but also the relative differences between the objects.

The interval scale is the first quantitative measurement scale presented. The nominal scale names and counts objects or attributes of objects, and the ordinal scale arranges objects.

2.3.4 Ratio Scale

A measurement that has all the characteristics of an interval scale plus a physically definable zero point is at the ratio measurement scale. For this scale, the ratio of any two measured values is independent of the units of measurement. Zero is the measure that defines the absence of a quantity. The ratio of the height to the width of a room is the same whether English or metric units are used. However the ratio of daily maximum temperature to the daily minimum changes from Fahrenheit to Celsius temperatures; thus length is a quantitative measure at the ratio level of measurement, and temperature is not. measurements of weight, mass, length,

width, and flow are typical ratio level measures.

2.3.5 Collapsing Measurement Scales

An important aspect of this system of measurement is that a higher level can always be collapsed into a lower level. For example, persons weights can always be grouped into underweight, ideal weight, and overweight. In this example, a ratio scale measurement has been collapsed into an ordinal scale. Note that this operation on the data has no inverse. That is, having groups of people classified as underweight or overweight does not allow reconstruction of their actual weights because the distinctions between groups are arbitrarly defined and are not always intuitively obvious. Weight groupings can depend upon age, sex, bone structure, and so on. Sometimes a 140 pound person is overweight, sometimes at ideal weight, and sometimes overweight.

A simple, but interesting, question is: what is the measurement level associated with the measurement of time (seconds, minutes, hours)?

2.4 STATISTICAL CONCEPTS

This section provides a brief review of important statistical concepts frequently used with extreme value analysis. Most statements resulting from scientific investigations are really inferences which are uncertain in character. Statistics is the formal study of this uncertainty, it attempts to both describe and to measure uncertainty. Probability is a measure of how likely is the occurrence of a chance event.

2.4.1 Data

An <u>experiment</u> is a carefully defined procedure whose outcome is observable but is not completely predictable in advance. Data is obtained when the observed outcomes are measured. The set of all possible outcomes is called the <u>sample space</u>. A <u>sample</u> is a particular set of data values

obtained when an experiment is repeated a number of times. The term 'experiment' is used both to refer to a procedure that yields a single data value, and to collectively refer to all such procedures that yield a data set.

2.4.2 Random Variables, Distributions, and Densities

A rule or mathematical function that associates a real number with each possible outcome of an experiment is called a random variable. A discrete random variable can take on only a finite or denumerable number of values, otherwise the random variable is continuous.

A density function is a mathematical rule which assigns a probability to each possible value of a (discrete) random variable. The density function is a link between the sample space and probabilities. Such rules for assigining probabilities have two distinct forms depending upon whether the random variable is discrete or continuous.

For discrete random variables the probability associated with each value, x, within the sample space of a random variable X may be enumerated. For each possible value x[i], the discrete density f(x[i]) assigns a specific probability

$$\text{Prob}(x[i]) = f(x[i]) \quad . \tag{2.1}$$

The axioms of probability impose the following restrictions on f(x[i]).

$$\begin{aligned} &0 \le f(x[i]) \le 1 \quad \text{for all } i \\ &\sum_{\text{all } i} f(x[i]) = 1 \end{aligned} \tag{2.2}$$

An alternate representation is the cumulative density function or distribution F(x[i]),

$$(2.3) \qquad F(x[i]) = \sum_{X \le x[i]} f(X)$$

The density, F, specifies the probability that the random variable, X, assumes a value less than or equal to x[i]. The axioms of probability require that

$$(2.4) \qquad \begin{aligned} &0 \le F(x[i]) \le 1 \quad \text{for all } i, \\ &F(-\infty) = 0, \\ &f(+\infty) = 1. \end{aligned}$$

Formula 2.3 relates the density to the distribution for discrete random variables. The distribution is a mathematical function that accumulates or integrates probabilities from the lowest possible value up to any specified value x[i].

For continuous random variables a different formulation of the density function is required. Since there is an infirite number of values within the sample space, it follows that

$$(2.5) \qquad f(x[i]) = 0 \quad \text{for each } i \ .$$

That is, the probability of any specific value is zero. This does not mean that a value is impossible, but that a value is extremely unlikely given the infinite number of alternate values. Also, the probability that a random variable assumes a value in the interval between two distinct points, say a and b, will generally not be zero. The points a and b can represent the precision of a measuring instrument. A scale that measures to the nearest gram can weigh objects with an irfinite number of possible weights between 20 and 21 grams, but the value recorded will be either 20 or 21 grams. Because of such considerations, the density as defined for a discrete random variable is replaced in the continuous case by a density

function f(x) defined by an integral,

$$(2.6) \qquad \text{Prob}(a \leq X \leq b) = \int_a^b f(X)dX \quad .$$

To be consistent with the axioms of probability, a continuous density function must satisfy the following conditions:

$$f(x) \geq 0 \ , \quad \text{and}$$

$$(2.7) \qquad \int_{-\infty}^{+\infty} f(x)dx = 1 \quad .$$

The distribution function, F(x), for the continuous case is defined as

$$(2.8) \qquad F(x) = \int_{-\infty}^{x} f(y)dy \quad .$$

The distribution F(x) defines the probability that a continuous random variable X assumes a value less than or equal to x.

2.4.3 Expected Value and Moments

Although a random variable is completely specified by either its density or distribution, it is often convenient to work with some descriptive measure or statistic which summarizes information about the random variable. The expected value of a random variable is such a summary. The expected value of any function of a random value is defined as the weighted average (weighted by the probability of occurrence) of the

function over the sample space. The symbol E[ ] is used to denote the expected value of whatever appears within the brackets. Thus the expected value of the function g() on the random variable X is denoted by E[g(X)].

For a discrete random variable X, with density f(), the expected value of a function g(X) is defined as

$$(2.9a) \qquad E[g(X)] = \sum_{\text{all } x} g(x)f(x) \quad .$$

For a continuous random variable the corresponding definition is

$$(2.9b) \qquad E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

The mean and variance of a random variable are special cases of the expected value function. The mean is a measure of centeral tendency and is defined by g(X) = x in equations 2.9a and 2.9b. The variance describes the spread or dispersion of the possible values of a random variable about the mean and is defined by

$$(2.10) \qquad g(x) = (x - E[X])^2 \quad .$$

The symbol V[X] is often used to denote the variance. The variance may also be interpreted as the average squared deviations from the mean. The standard deviation, s, is defined to be the positive square root of the variance, and has the advantage of having the same units of measurement as the mean.

The mean and standard deviation are often referred to as location and scale parameters respectively. It is common practice in statistics to express a random variable as a distance from its mean in multiples of its standard deviation. This is called a standardized random variable and formally is the transformation

$$(2.11) \qquad z = \frac{x - E[X]}{s} \quad .$$

The new random variable Z with values of z has a mean of zero and a variance of one.

A function of two random variables that has special importance in statistics is the product of their deviations from their corresponding means. The expected value of this function is the covariance, defined as

$$(2.12) \qquad Cov[X,Y] = E[(x - E[X])(y - E[Y])] \quad .$$

The covariance is important because it measures the linear association, if any, between the two random variables. If X has no influence on Y then X and Y are said to be independent and their covariance will be zero.

A measure of dependence which is related to the covariance is the correlation coefficient, r, defined as

$$(2.13) \qquad r = \frac{Cov[X,Y]}{(V[X]*V[Y])^{1/2}} \quad .$$

The correlation has a range from -1 to +1 with a value of zero indicating independence. A positive correlation indicates that Y tends to have values of the same sign as X and a negative correlation indicates that Y tends to have values of the opposite sign from X.

2.4.4 Prototype Random Variables

For extreme value statistics, two kinds of random variables are used to describe exceedances and magnitudes respectively. These classifications are discussed in Chapter 3. For nominal and ordinal level data the extreme value distributions available are limited to the distribution of counting exceedances. The theoretical basis of this distribution is the binomial family of distributions. For interval and ratio level data the magnitude

of extreme values can be measured and a second level of distributions becomes important. Three families of distributions are used in extreme value statistics: the Weibull family when the sample space is bounded, the Exponential family when the sampling is from the more common continuous random variables, and the Cauchy family for extremes from random variables with non-finite variance. A typical random variable of the Exponential family, which is the most important family, is the Gaussian or Normal distribution.

2.4.5 Sample and Population Moments

A major goal of statistical analysis is to make inferences about a population from a sample. Usually the density function f(x;p) is assumed known, but contains unknown parameters p. There are two kinds of statistical inference: estimates and hypothesis tests. Estimation is further divided into point estimation and interval estimation. A statistic is a mathematical function of sample values which does not contain any unknown parameters and in some sense extracts information from the sample. An estimator of a population parameter, p, is a statistic that is designed to produce numerical values that represent the numerical value of the parameter. An estimate is the numerical value produced by an estimator using sample data. A variety of statistical criteria are available in statistics for judging how representative an estimate is for a parameter. These criteria in turn use such concepts as 'unbiased', 'robust', 'minimum variance', 'consistent', and combinations of these.

An approximation of a population density function that is derived from a sample and has no unknown parameters is the empirical density. The corresponding empirical distribution is then an approximation to the distribution of the population being sampled. Let x[1], x[2], ...x[n] represent an ordered (from smallest to largest) sample of size n from a (continuous or discrete) random variable f(x;p). In this ordered form the x[i] are referred to as the order statistics. The empirical density

function is defined as

$$h(x) = \begin{cases} 1/n & \text{for } x=x[i],\ i=1,2,\ldots n \\ 0 & \text{elsewhere.} \end{cases} \tag{2.14}$$

The corresponding empirical distribution is defined as

$$H(x) = \begin{cases} 0 & \text{for } x < x[1] \\ i/n & \text{for } x[i] \leq x \leq x[i+1] \\ 1 & \text{for } x \geq x[n] \end{cases} \tag{2.15}$$

The empirical distribution is equal to the fraction of a sample that is less than or equal to any given value of x.

Sample moments are defined by substituting the empirical density, h(x), for the population density, f(x), in equation 2.9a. Sample moments are especially important statistics because the expected value of a sample moment is equal to the corresponding population moment. It is important to conceptually separate population moments and sample moments. It is always possible to compute a sample moment because the form of h(x) is always known and contains no unknown parameters. The population moments may not be computable because: 1) the form of the function f(x) is unknown, 2) f(x) contains unknown parameters, or 3) f(x) is of such a mathematically complex form that the expectation cannot be derived. The sample mean and sample variance are the most used (and often missused) sample statistics.

CHAPTER 3
FAMILIES OF DISTRIBUTIONS

3.1 INTRODUCTION

The properties of the most useful mathematical models for describing the random fluctuations of data are well presented in statistics courses. These models are the distribution functions defined in Chapter 2, and are given names such as: Normal, Gaussian, Students-t, Gamma, Binomial, Poisson, and so on. Extreme value statistics is not concerned with individual distributions, but rather with groups of distributions defined by common mathematical and statistical properties. It is necessary to understand these groupings or families of distributions before one can understand the statistical theory of extreme values. A verbal discussion with a few mathematical formulas should show adequately the structure and relationships within this statistical theory. Many textbooks are available to fill in details and add rigor.

There are three asymptotic extreme value distributions. These asymptotic distributions are not a good place to start a discussion for an understanding of the relationships between the several parts of extreme value theory. To start with, it is more important to note that extreme value statistical procedures can be divided into (I) those that count the number of occurrences of extreme events, and (II) those procedures that measure the magnitude of extreme events. An understanding of the differences between counts and magnitudes was the motivation for the portion of the previous chapter on measurement scales. This division of procedures also has an analogy to the division in statistics between discrete and continuous distributions.

3.2 COUNTING EXTREME VALUES

The first type of procedure is the enumeration of rare events. The data consist of the number of rare events that occurred in some time interval or perhaps in a group of experiments. The terminology used for this is 'the distribution of the number of exceedances'. The magnitude of a measurement which determines what is a rare event or an extreme value is defined, and then the number of such rare events is counted. The statistical procedures for analyzing exceedances will be given in the next three chapters.

There are two ways of specifying the definition of a rare event. The most common way is an arbitrary declaration derived from physical, chemical, or biological principles, such as 'exposure to more than 50 millirems radiation per year is hazardous'. Just as with the example of classifying people as over- or underweight, 50 millirems is sometimes hazardous and sometimes not. Setting standards of maximum allowable exposure is a common problem of environmental and health regulations. Once a standard has been set, exceedances can be counted, regardless of the validity of that standard. Even though some standards may have been based upon emotion, fear, or politics, instead of scientific evidence, exceedances can be statistically analyzed for any given standard. A current example is the controversy over how much saccarine is too much. Extreme value theory for exceedances is not concerned with how, why, or at what magnitude the standard is set. Of course, the statistician must emphasize that the resulting statistics are no more or no less reasonable than the standards themselves.

In some situations it may not be necessary to use fixed or arbitrary standards. A future data set can be compared to a past set. In this situation, exceedances are still counted, but the standards are not fixed quantities. This differs from a fixed standard as illustrated in the following example. Consider the study of air pollutant oxidant levels in a

city. Using an arbitrary standard, a concentration over one part per million is declared an exceedance and causes an air pollution alarm. Using past data the maximum oxidant levels recorded over the last, perhaps 20 years, is tabulated and then the number of times the current data exceeds the 20 year maximum is counted. The standard could also be chosen to be the second, or third highest level in the 20 year data. Here exceedances are still being counted, but the definition of an exceedance has changed. The exceedance definition is derived from past data and its magnitude is a sample value from a random variable.

Definition of exceedances using past data is, of course, only possible if past data exist. It does, however, offer one way to alleviate the problems inherent in using arbitrary standards. Note that this type of standard defines a transformation from an interval or ratio level of measurement to a nominal level.

Statistically, exceedances are modeled with discrete distributions. Depending upon the specifics of the problem, the Binomial, Poisson, Geometric, or Hypergeometric distribution could be used. These situations will be examined in detail in Chapter 5.

In elementary statistics courses the Poisson distribution is sometimes called 'the distribution of rare events'. The Poisson is a distribution that counts rare events and does not measure their magnitudes. Thus, it does not define a rare event. It models the number of occurrences that are defined so that the probability of an occurrence is extremely small. The probability of an occurrence is assumed to be fixed and known. Rare events are identified and then counted at the nominal measurement level, while extreme values are inherently at the ratio or interval level. Because measurement scales can be collapsed but not expanded, extreme values are rare events but rare events are not extreme values.

3.3 THE MAGNITUDE OF EXTREME VALUES

The second type of procedure for analyzing extreme values uses the magnitude of the rare event. In the oxidant example given above, the data would include not only a count of the number of oxidant levels observed to be over one part per million, but would also list all the readings in this class. Typically the single value that is the maximum is the most important item in such a list.

Extreme value data is typically collected in fixed-size groups, such as the 24 hourly oxidant levels in each day. When studying exceedances, the proportion of those 24 measurements that exceeded the one part per million criteria is determined. When studying extremes the single maximum of those 24 measurements is used. The grouping of the measurements is usually easy to define because groups are derived from uses of the data. Floods are grouped by annual maxima rather than daily maxima for obvious reasons. Oxidant levels are typically grouped on a daily basis because the acute health effects response to high oxidant levels develop over a day or so of exposure, rather than over years.

After a number of extreme values have been collected, such as daily maximum oxidant levels, weekly maximum radon levels in air over a uranium mill tailings pile, annual floods, and so on, these data may be modeled by a statistical distribution. The kinds of distributions used in this context are called extreme value distributions. Theorems that have been developed allow mathematical derivation of the extreme value distributions from the distribution of the original data and the sample size. The mathematical derivations are often extremely difficult, thus the appropriate asymptotic distribution is usually used. The details of how to derive these distributions are given in textbooks on order statistics. This text will use only the asymptotic distributions, except for a few simple cases.

A special characteristic of all extreme value distributions that facilitates applied work is the simple form of their asymptotes as the size of the group from which the extreme is extracted becomes large. These asymptotes converge to only three distributions. These three distributions were first derived by Fisher and Tippet in 1928. The history of this important contribution was given in the first chapter. The problem of choosing the correct extreme value distribution is made easier by the fact that the choice depends upon some general characteristics of the data distribution from which the extremes are extracted.

All continuous statistical distributions can be classified into three families; the Exponential family, the Cauchy family, and the Weibull family. One of the three extreme value distributions is associated with each of these three types of data distributions. (The three asymptotic extreme value distributions are members of the Exponential family). The classification of data distributions can be difficult. Fortunately, all the commonly used statistical distributions have already been classified.

The Exponential family of distributions permit unlimited values of the variables. The area under the tails of the distribution curve must converge to zero for large (positive or negative) values of the variate, at least as strongly as the area under the tail of the exponential function, EXP(-x). All moments exist for the members of this family. However, not all distributions where all moments exist are of the Exponential family. Most of the common statistical distributions belong to this family. The family includes the Normal (Gaussian), Exponential, logistic, log-normal, Gamma, and Chi-square distributions.

Other distributions, which are also unlimited, have a very long tail so that they converge less strongly than the exponential function. These distributions have no moments beyond a certain order. The Cauchy distribution is the only well known member of this family, which is called the Cauchy family. The members of this family are rarely encountered in

applied statistics, an exception is the distribution of ratios which often belong to this family.

The third family of extreme value asymptotic distributions is easy to recognize. These are the distributions with an upper or lower limit. The largest, or smallest, extreme value is thus bounded, the limits become parameters of the extreme value distribution. Higher-order moments also do not exist for the members of this family of distributions. The prototype of the limited distributions is the Weibull, which is used extensively in engineering stress problems. This family is called the Weibull family. Strength of materials, dielectric strength, and failure time of machines usually follow a Weibull distribution. The Beta distribution also belongs to this family. The Exponential, Cauchy, and Weibull distributions are defined in the appendix to this chapter.

Some further simplifications can be made. It is only necessary to consider data from one tail of a data distribution, either the largest extreme or the smallest extreme. A distribution often belongs to different families depending upon whether large values or small values are being studied. The Exponential data distribution is bounded below because only positive values of the variate are allowed, but it is unbounded above. Thus, even though it is the prototype for the Exponential family, the prototype behavior applies only to extreme large values. For extreme small values, the Exponential data distribution belongs to the Weibull family.

The distinction between the Exponential family and the Cauchy family of distributions is usually made by examining a data distribution function for the existence of higher moments. It is the parent data distribution that needs to be considered; that is, the distribution which models the physical, chemical, or biological mechanism from which sample measurements are derived. There is also a sampling distribution associated with every statistical problem. The most common example of this association is that samples from a Gaussian (Normal) distribution have a Students-t

distribution.

A sample data distribution is not adequate for such a determination. This usually means that the identification of this distribution must be derived from physical-chemical, or mathematical principles rather than from examination of data sets.

The existence of a sample variance does not mean that a corresponding population variance exists. This can be very confusing for the Cauchy distribution. Mathematical statistics teaches that the Cauchy distribution can be recognized by the fact that it has no variance, and yet a sample variance can always be computed for any sample from a Cauchy distribution. Why doesn't the sample variance tell us something about the parent distribution? An experiment can explain this unexpected characteristic of the Cauchy distribution. Suppose you were to take a series of samples, of increasingly larger size, from both a Gaussian and a Cauchy distribution, calculate the sample variances, then plot them against sample size. You would discover that for the Gaussian distribution the sample variances converge to the population variance as the sample size gets large. However, for the Cauchy distribution no such convergence will be seen. The sample variance will increase in an unbounded way as sample size gets large. This suggests that the variance is infinite for infinite-sized samples from a Cauchy distribution. It can be shown mathematically that this is true. These results of comparing sample variances of the Cauchy and Gaussian distributions could have been predicted from theormes of mathematical statistics. These theormes require a great deal of background to use, but the general ideas can be explained simply. First one must distinguish between sample variances and the variance of the population from which the sample was taken.

The population variance is defined as the second moment about the mean, u. Formally, for a statistical density function f(x), the variance v is:

$$v = \int_{-\infty}^{+\infty} (x-u)^2 f(x)dx \ .$$

When f(x) is replaced by the formula for a particular density function and the integration is performed, one may find that the variance is degenerate, infinite, or some other intractable mathematical phenomena. One is then led to ask how the commonly used estimator of the variance was obtained? This estimator is:

$$\hat{v} = S^2 = \Sigma (x-x)^2/(n-1)$$

This formula came from assuming f(x) to be the Gaussian density function. First the Gaussian density f(x) was substituted into the above integral to show that the variance existed, and is equal to that parameter of f(x) that has been named the variance. Then the maximum likelihood equations for a sample from a Gaussian distribution were solved in order to express the variance estimate in terms of sample data values. Finally the equation derived from maximizing the likelihood function was adjusted to be unbiased. This yielded the common formula for estimating the variance. It is important to note that the sample variance estimates a parameter of the Gaussian distribution and that other distributions do not contain this parameter. Often, but not always, an algebraic relationship can be found so that the sample variance can be used to estimate a parameter of a non-Gaussian distribution. The existence of such an algebraic relationship does not imply that the estimated parameter is a variance, or even that such a parameter measures the spread of a distribution. The sample variance has come to be used as a descriptive measure of the spread or amount of variability in a data set. It is convenient and useful for this purpose, but this utility does not imply any utility as a parameter estimator.

3.4 RELATIONSHIPS AMONG EXTREME VALUE DISTRIBUTIONS

The three extreme value distributions are related through logarithmic transformations. These transformations allow extreme values from both the Cauchy and Weibull families to be analyzed using statistics derived from the Exponential family. Because of these relationships, the extreme value distribution derived from the Exponential family is usually called 'the extreme value distribution'.

A logarithmic transformation of the data converts a Cauchy type extreme value to an Exponential type (Sarhan and Greenberg, 1962). Formally, if y has a Cauchy type extreme value distribution, then $x=\ln(y)$ has an Exponential type extreme value distribution ('ln' denotes the natural logarithm). This simple relationship of types of extremes could easily be confused with the relationship between the Gaussian and Log-normal distributions. Both the Normal and the Log-normal are members of the Exponential family of distributions. Extreme samples from these two distributions have extreme value distributions of the same form but with different parameters. Since the logarithms of a Log-normally distributed sample can be analyzed as a Gaussian distributed sample, algebraic relationships can be found between the parameters of these two distributions and the parameters of corresponding extreme value distributions. The logarithm of the extreme value from a Log-normal distribution is equivalent to the the extreme value from the corresponding Gaussian distribution.

The logarithms of the samples from a Cauchy distribution do not have a Gaussian, a Log-normal, or any other well known distribution. Nor do the logarithms of the extremes from a Cauchy type distribution correspond to the extremes from a Gaussian distribution in any definitive way. However the logarithms of the extremes from a Cauchy type distribution do have the same form of extreme value distribution (with unique parameter values) and thus the same general statistical properties as the extremes of any members

of the Exponential family.

The logarithmic transformation for extremes from the Weibull family are somewhat more complicated. Some of the parameters of the extreme value distribution must be known or estimated before the logarithmic transformation can be applied. Let z be the variable with the Weibull type distribution. Let x be the variable with an Exponential type extreme value distribution, derived from z through some logarithmic transformation. Also let w be an upper bound on z (a reflection will be obvious for bounding from below). Two parameters are used in this transformation: u is the mode of the x's, and v is the mode of the z's. The transformation is (Sarhan and Greenberg, 1962):

$$x = \ln\left(\frac{w - v}{w - z}\right) + u$$

This transformation is not easy to use because the values of u and v are usually unknown. The value of the bound, w, may or may not be known. Often the value of w is fixed by physical-chemical principles.

Modern computer algorithms for generalized functional maximization such as the Simplex method (O'Neill, 1971) can be used to simultaneously maximize the likelihood that x has an extreme value distribution and also to find the maximum likelihood estimates of u, v, and, if needed, of w. This algorithm will be used in the examples given in subsequent chapters. In order to use the Simplex method one must know computer programming and have access to a general purpose scientific computer.

3.5 REFERENCES

Fisher, R. A., and Tippet, L.H.C. 1928. "Limiting Forms of the Frequency Distribution of Largrsr or Smallest Members of a Sample," Proc. Cambridge Phil. Soc. Vol. 24, pt. 2, pp 180-190.

O'Neill, R. 1971, Algorithm AS 47, "Function Minimization Using a Simplex Procedure", Applied Statistics 20, pp 338-345.

Sarhan, A. E., and Greenberg, B. G. 1962, "Contributions to Order Statistics", John Wiley & Sons Inc., pp 67-70.

3.6 APPENDIX 3-A

Distributions and Densities

Exponential
Range: $0 < x < +\infty$
Mean = Standard Deviation = $1/b$, $b > 0$.
Mode = 0, Median = $\ln(2)/b$

Density

$$f(x) = b*\exp(-b*x)$$

Distribution

$$F(x) = 1 - \exp(-b*x)$$

Cauchy
Range $-\infty < x < +\infty$
Location parameter = a, the median and mode.
Scale parameter = b.
(Pi = 3.14159...)

Density

$$f(x) = 1/(Pi*b(((x - a)/b)^2 + 1)$$

Distribution

$$F(x) = 1/2 + Pi^{-1}*\text{atan}((x - a)/b)$$

Weibull
Range $0 < x < +\infty$
Scale parameter = b, $b>0$
Shape parameter = c, $c>0$.

Density

$$f(x) = (c*x^{c-1}/b\ *\exp(-(x/b)^c)$$

Distribution

$$F(x) = 1 - \exp(-(x/b)^c)$$

CHAPTER 4
COUNTING EXCEEDANCES

4.1 INTRODUCTION

The statistical theory used for studying exceedances is a union of many parts of statistics, some old and some new. The discussion in chapter 2 shows that exceedances are at the nominal measurement level. That is, the data consists of a count of events classified as extreme or exceeding a fixed criteria. Statistical methods for analyzing nominal data are some of the oldest, and are associated with the origins of statistics. Statistics and probability started in the eighteenth century when wealthy gamblers called upon mathematicians to determine the correct odds in their games, so they could find the best betting strategies. From this begining, probability theory developed what is now called the Bernoulli class of distributions. This class includes the Binomial, Negative Binomial, Multinomial, Geometric, Hypergeometric, Pascal, and Poisson distributions.

The statistics of exceedances are nonparametric or distribution free since the methods require only nominal level data from an underlying continuous distribution. It is assumed that the observations are independent and that the parameters of the underlying distribution do not change (over time) or that they change in a known way. The basic problem is to forecast the average number of cases that will exceed a specified value within the next N trials (or time periods).

The Bernoulli class will be emphasized and, in the next chapter, Exceedances distributions and statistical tests derived from them. These distributions consider counts or frequencies of extremes, and the time or number of samples between occurrences of extremes. They allow estimation of how often extremes may occur. They do not consider the magnitude of the

extremes. Many applied problems work with limited and unreliable data. Using the statistics of counts or frequencies allows the derivation of useful conclusions with a minimum of underlying assumptions.

A critical factor in the use of Bernoulli distributions is that the parameters of the distribution are known constants. In tossing a coin, throwing dice, or dealing cards, the probability of winning or losing can be determined exactly. Although substantial algebra may be required to determine these probabilities, they are mathematical consequences of simple well known physical facts: two sides to a coin, six sides on a die, 52 cards in a deck, and so on.

In the study of exceedances the probability of an exceedance occurring usually cannot be estimated as accurately as in a gambling situation. Typically the probability of an extreme event must be estimated from limited past data using the assumption of no time dependent changes, or from concomitant data which has measurement error. The Bernoulli distributions may be used when substantial information is available so that the probability of an exceedance occurring is well established. When this probability must be estimated from data and there is significant error in the estimate, the Bernoulli distributions should not be used. When error is significant a version of the 'Distribution of Exceedances' should be used. This distribution was first published by S. S. Wilk in 1927 (Wilk, 1942). Note that the distribution of exceedances came along two centuries after the Bernoulli distributions. The Distribution of Exceedances is the topic of chapter 5.

4.2 THE BERNOULLI DISTRIBUTIONS

This section reviews relationships between the common Bernoulli distributions. Elementary statistics and probability textbooks provide additional details. The best known of these distributions is the Binomial, which gives the probability of the number of 'successes' in a fixed number

of trials in which sampling with replacement is assumed. When there are more than two kinds of outcomes (e.g. highly polluted, slightly polluted, and not polluted), the Binomial generalizes into the Multinomial distribution. When sampling is without replacement, the probability of an outcome changes as each sample is taken, such as in the game of Bingo. Then the correct distribution is the Hypergeometric.

Instead of counting the number of successes in a fixed number of trials, suppose the number of successes is fixed and the number of trials is counted. The Geometric distribution counts the number of trials necessary to achieve the first success. For example, with a specified probability of a pollution episode, the Geometric distribution would count the days between episodes. The Pascal distribution is the extension of the Geometric distribution that counts the number of trials to achieve the mth success rather than the first success. The extensions to sampling without replacement have not been named but are discussed in probability textbooks. The negative Binomial distribution is a variation of the Pascal. It counts the number of failures before the mth success while the Pascal counts the number of trials up to and including the mth success.

There are also two important asymptotic extensions that start with the Binomial distribution, and consider what happens as the number of trials becomes very large. When the probability of success remains constant and the number of trials approaches infinity, the Binomial distribution asymptotically approaches a Gaussian (Normal) distribution with mean value Np and variance Np(1-p). N is the number of trials and p is the probability of success. In practical applications this approximation has been found reasonable for values of N as small as 20 if p is not very close to zero or one. In extreme value problems p is usually close to zero, so one should be cautious about using this approximation.

The second asymptotic situation is one in which the probability of success decreases as the sample size increases in such a way that the

product is constant. That is, as $N \to \infty$ and $p \to 0$, then $Np \to c$, where c is a constant. This asymptote yields the Poisson distribution with parameter c. The Poisson distribution is commonly called the distribution of rare events. More correctly, it should be called the distribution of the number of rare events because it is a distribution of how often rare events occur, not of the magnitude of rare events. A more comprehensive reference to statistical distributions is Hastings and Peacock (1974). These distributions are also included in the comprehensive statistical reference by Beyer (1966).

4.3 EXAMPLES USING THE BERNOULLI DISTRIBUTIONS

Example 1

A large industry claims that it emits perceptible smoke from its incinerator on 5 or less percent of the days. The city in which the industry is located has hired a consultant to investigate this claim. The consultant monitors the stack on 20 randomly chosen days over a summer and uses the decision rule to accept the industry claim if smoke is observed on zero or one observation day, and reject the claim if smoke is observed on two or more days. The consultant chose this claim because one day is 5% of 20 days of sampling. What is the probability that the consultant will reject the industrial claim even though it is correct? What is the probability that the consultant will accept the claim if the true probability is 0.1? The first of these questions concerns binomial confidence intervals, and the second concerns binomial power.

A table of binomial probabilities is used to answer these questions. Table 4.1 is a portion of such a table; N is the sample size and X is the observed frequency of success. The body of the table contains the probability that X successes occur in N trials where N=20. The columns of the table correspond to the probability of X successes in 20 trials with

the true probability of success equal to 0.05 and 0.10.

TABLE 4.1

P(x successes in N=20 trials)

X	0.05	0.10
0	.3585	.1216
1	.3774	.2702
2	.1887	.2852
3	.0596	.1901
4	.0133	.0898

The probability of zero or one success if the true probability of each success is 0.05 is the sum of the first two enteries in the 0.05 column, .3585 + .3774 = .7359. The probability is 1.0 - 0.7359 = 0.2641 that the consultant will reject the industry claim even though it is correct. Thus, the consultant has about a 26% chance of making an error against the industry. What should the consultant's rule be for him to have less than a 5% chance of making this error? Continue adding up the 0.05 column and subtracting the total from one until the answer is less than .0500. If his rule is to accept the industry claim if 0, 1, or 2 smokey days are observed, his error probability is about 0.08. If his rule is 3 or less smokey days his error probability is about 0.02. Thus to keep his error rate under 5% the consultant should not reject the industry claim until he observes smoke on more than 3 of his 20 days of observation.

Solving the same problem using a true probability of observing smoke of 0.10, no smoke in 20 observations will then occur with probability 0.12, smoke on 0 or 1 day would occur with probability 0.39, and on 0, 1, or 2 days with probability 0.68. Using the consultant's original rule, there is a probability of 0.39 of erroneously accepting the industry claim of 0.05 when the true probability of smoke is 0.10. Thus, for the original rule, the consultant has a 55% chance of making some kind of error (1 - (1 -

.26)(1 - .39) = .55). If the consultant protects himself against erroneously rejecting the industry claim by changing his rule to 3 or more days, he will erroneously accept with probability 0.87 when the true probability is 0.10.

The consultant should obviously take more than 20 samples in order to control both kinds of errors. The steps in this example can be repeated for successively larger sample sizes until a sample size and decision rule is found to give acceptable probabilities for both kinds of errors. For larger samples, the normal approximation to the binomial is useful.

It is important to see exactly how this example relates to the assertion that the Bernoulli distributions can be used only with known parameters. Here a known value for the probability of success is hypothesized, and data sets are compared with this hypothesis. The basic question was: could this data have come from a Bernoulli distribution with the hypothesized parameter? The probability of success was not estimated from this data set (or any other data set of similar size).

Example 2

Suppose a city must decide where to put a new sewage treatment plant. It has two possible sites, A and B. If A is chosen, the cost will be much higher, and if B is chosen, residents might object to the odor. The city council decides that B is acceptable if odors can be detected on five or fewer days of a year. The city engineer has good meteorological data and finds that the wind blows from the B site over residential areas on 1.4% of the days of the year. Most available tables of the Binomial distribution do not go beyond a sample size of 20, and calculating the binomial function for samples of 365 would be tedious. This is a situation where the Poisson approximation to the Binomial can be used since the sample size is large and the probability of success is small. (Recall that the Gaussian

approximation to the Binomial is used when the probability of success does not get small with increasing sample sizes.)

The parameter of the Poisson distribution is found by multiplying the Binomial probability of success (p) by the number of trials (N). In this example Np = 365*.014 = 5.11. This parameter is both the mean and the variance of the relevant Poisson distribution, so the standard deviation is 2.26. For plant site B there should be a yearly average of just over 5 days (exactly 5.11 days) of odor from the proposed plant. Hence, there is about a 50% chance that the city council's criteria will be satisfied in any one year. The city council might not consider this statistic much help in decision making. A 95% confidence interval can be approximated on the number of days of odor per year by taking the mean plus and minus two standard deviations, $5.11 \pm (2)(2.26)$ or 0.59 to 9.63 days. A confidence interval on Poisson observations should be stated as integers (without fractional parts). Hence, there is approximately a 0.95 probability that odor from site B will be detected by residents at least once but less than 10 times in a year.

Using a table of Poisson probabilities, it is easy to check the accuracy of this approximate confidence interval. Such tables are found in most statistics text. One section of such a table is reproduced as Table 4.2. The X column is the number of successes. The next column contains the probability of exactly X successes if the Poisson parameter is 5.1. The last column is the cumulative sum of the probabilities and is thus the probability of X or fewer successes.

TABLE 4.2

Poisson probabilities for a mean of 5.1

X	prob.	cumulative prob.
0	.0061	.0061
1	.0311	.0372
2	.0793	.1165
3	.1348	.2513
4	.1917	.4232
5	.1753	.5985
6	.1490	.7475
7	.1086	.8561
8	.0692	.9253
9	.0392	.9645
10	.0200	.9845

This table indicates that the confidence interval approximations are somewhat in error. First consider the mean. Previously a 50 - 50 chance was assumed that an annual count of odorous days would be below the mean of 5.1. The last column of Table 2 shows that the objectional odor has almost a 60% chance of occurring on 5 or fewer days per year. Perhaps the city council would accept 60 - 40 odds but not 50 - 50 odds of meeting their criteria. The mode of the distribution is 4, that is, observing 4 days of odor per year has the highest probability of all outcomes. The 95% confidence interval is obtained by studying the cumulative probabilities column and picking a set of X values for which the cumulative probability is between 2.5% and 97.5%. This gives 1 to 9 days rather than 1 to 10 days as found by the approximation.

Symmetric confidence bounds are used mainly out of habit learned from working with continuous distributions. With continuous distributions a 95% confidence interval is obtained with the interval from exactly the 2.5% value to exactly the 97.5% value. A 95% interval can also be obtained from 1% to 96%, or many other combinations. To choose between these many alternatives it is assumed that symmetry about the mean value is a

desirable attribute of confidence intervals. Since the Bernoulli distributions have discrete jumps in probability, exact symmetry or exactly 95% confidence levels cannot usually be achieved. The confidence interval of 1 to 9 successes is chosen on the basis of getting as close to symmetry as possible; that is, choosing the endpoints as close as possible to 2.5% and 97.5%. An examination of the cumulative probability column of the Poisson table shows that a 92.7% confidence level is actually achieved (96.45% - 3.72% = 92.73%). To find the true confidence level of the 1 to 10 successes derived from the approximation, subtract the cumulative probability of one success from the cumulative probability of 10 successes to get 94.7%. This is much closer to the desired 95% level than is the interval of 1 to 9 which was chosen for symmetry. The interval of 0 to 9 successes has an exact confidence of 95.8%. This example shows that there is some leeway in specifying confidence intervals for discrete distributions because no single criteria for choosing endpoints is universally applicable.

Four different intervals for this sample data have been illustrated. All are reasonable 95% confidence limits. The first was obtained from asymptotes. The second, from exact probabilities restricted to be as close to symmetrical (in probability) as possible. The third was from exact probabilities as close to 95% as possible. The fourth was from exact probabilities and was as close as possible to a confidence interval with <u>at least</u> 95% probability. It is important to indicate the criteria used to choose confidence intervals of discrete distributions.

4.4 SUMMARY

This chapter discusses counting extremes when the probability of an extreme value occurring is well established. This situation is handled statistically with the well known Bernoulli class of distributions.

The logical relationships between the members of the Bernoulli class were reviewed, and their uses were illustrated. The critical condition for use was emphasized; that the probability of the extreme occurring be known without significant error.

4.5 REFERENCES

Beyer, W.H. (Ed.), 1966, Handbook of Tables for Probability and Statistics, The Chemical Rubber Publishing Co.

Hastings, N.A.J., and J.B. Peacock, 1974, Statistical Distributions, John Wiley & Sons (Halstead Press), ISBN 0-470-35899-0.

Wilk, S. S., 1942, "Statistical Prediction with Specific Reference to the Problem of Tolerance Limits", Ann. Math. Stat., 13, pp 400-409.

CHAPTER 5
DISTRIBUTION OF THE NUMBER OF EXCEEDANCES

5.1 INTRODUCTION

Suppose a Bernoulli trial situation but the probability of a success is not known. It is necessary to estimate this probability from historical data. Historical data often is of poor quality or badly documented. Such estimates of probability obviously introduce a source of error in addition to sampling error. Sampling error is the only source of error assumed to exist by the Bernoulli distributions.

The formal solution of this problem can be found in the 'Bayesian Estimation' sections of contemporary mathematical statistics textbooks. First a compound, conditional distribution of the binomial event is derived given that the binomial probability parameter is a random variable. Then, the marginal distribution of the event may be derived by integration. This requires that the distribution of the parameter itself be known. The early workers in extreme value theory didn't have the tools of modern Bayesian theory. However they essentially performed the same steps in the derivation outlined in the next section.

5.2 DERIVING THE DISTRIBUTION

The early workers in extreme value statistics found a way of circumventing the explicit estimation of the probability of success from data. Their method has the desirable attribute of being nonparametric. Instead of explicitly estimating the probability of an exceedance or choosing a criteria for classifying observations as extremes, they expressed the unknown distribution parameters as functions of past observations. A historical or reference data set is examined to determine

how many values in this data set exceed a criteria. The criteria is expressed as the rank of the observation in the reference data set that is closest to the criteria. The question to be answered is: in how many cases, x, will the mth observation out of a total of n observations in the past data be equalled or exceeded in N future trials? The n reference observations are assumed ranked so that m=1 denotes the largest observation, and m=n is the smallest. Therefore the mth observation is the mth largest. A symmetry will be obvious so one could rank from the bottom and consider small observations. Because of this symmetry, only large extreme values will be discussed in detail. Note that all available sample values are used, the extremes are not picked out for analysis.

The sample size for which the forecast is wanted, N, is often not the same as n, the sample size of the reference data. The number of cases, x, called the number of exceedances, is a new statistical variate. Its density is denoted by w(x;n,m,N) where n, m, and N are parameters. The starting point is a special case of a distribution studied by Wilk (1942). A dichotomy is constructed based on the mth largest of the past n observations. The probability of a new data value being less than the mth past value is denoted by F and is unknown. The probability of an exceedance is 1-F. This is a Binomial situation except that the probability of a success (exceedance) is unknown. If a Binomial distribution with F as the probability of success is formulated and integrated over all possible values of F, density of Exceedances is obtained.

$$(5.1)\quad w(x;n,m,N) = \frac{m\binom{n}{m}\binom{N}{x}}{(N+n)\binom{N+n-1}{x+m-1}}, \qquad \sum_{x=0}^{N} w = 1.0$$

where $0 \leq x \leq N$ and $1 \leq m < n$. The large brackets represent the binomial coefficient:

$$\binom{x}{y} = \frac{x!}{y!(x-y)!} \quad .$$

The useful symmetry for extreme small values is:

(5.2) $$w(x;n,m,N) = w(N-x;n,n-m+1,N) \quad .$$

The cumulative density or distribution is denoted by W(x;n,m,N) where values of w are summed from 0 to x.

Restated in words, w is the probability that there will be exactly x values in a new sample of size N that will equal or exceed the mth value in the reference sample of size n. For W, 'exactly x' in the previous sentence is changed to 'x or fewer'.

Nothing is assumed known about the variate from which the two samples were taken except that it is continuous and does not change between the time of the two samplings. Thus the distribution of Exceedances is distribution free or nonparametric.

The probability is 1/2 that the largest (m=1) of N reference observations will not be exceeded (x=0) in N future observations.

$$w(0;N,1,N) = \frac{1*\binom{N}{1}\binom{N}{0}}{2N*\binom{2N-1}{0}} = \frac{\frac{N!}{1!(N-1)!}*\frac{N!}{0!(N-0)!}}{2N*\frac{(2N-1)!}{0!(2N-1-0)!}} \quad .$$

Since N!=N(N-1)!, this formula simplifies to:

$$w = \frac{\frac{N(N-1)!}{(N-1)!}*\frac{N!}{N!}}{2N*\frac{(2N-1)!}{(2N-1)!}} = \frac{N}{2N} = 1/2 \quad .$$

The probability that the largest of N past observations will always be exceeded can be calculated using x=n, or more simply by using the inverse probability:

$$P(\text{always}) = 1 - P(\text{never}) = 1-1/2 = 1/2.$$

By symmetry it is clear that the smallest of N past observations has a probability of 1/2 of never or always being exceeded in N future observations.

The formula for w reveals some interesting aspects of the Exceedances density for special values of x, m, n, and N. The probability that the mth largest value from n initial observations will be exceeded at least once in N new observations is:

$$(5.3) \quad P(x \geq 1) = 1 - \frac{n!(N+n-m)!}{(n-m)!(N+n)!} = 1 - P(x = 0) \quad .$$

When m=1 this is the probability that the largest value from the initial distribution will always be exceeded in N new observations, and is:

$$(5.4) \quad P = 1 - \frac{n}{n+N} = \frac{N}{n+N} \quad .$$

The probability that all N of the new observations will exceed the largest (m=1) of the original n observations is given by:

$$(5.5) \quad w(N;n,1,N) = \frac{n!*N!}{(N+n)!} \quad .$$

This is a very small probability for even small values of n and N.

If n (the reference sample size) is odd, then m=(n+1)/2 corresponds to the median of the initial variable. From the definition of median, it is

equally probable that the median of n past observations is exceeded x or N-x times in N future trials. The density of the number of exceedances above the median is symmetrical.

5.2.1 Example: Design of Experiments

Formulas 5.1 through 5.4 can be used to design an experiment to approximate the "worst case condition". The number of observations to take, n, such that x or more values greater than the largest of these n will occur with probability p in N future observations can be calculated from these equations. Typically x and p are set small. When N is known equation 5.1 can be solved by summing over the values of x. If x is restricted to x=1 then the problem is simplified since no summation is needed, and equation 5.4 can be used. Suppose a 90% chance that weekly pollution maxima will not exceed the largest value in a reference data set is desired. How many weeks of data should be collected to obtain this maximum value? To use formula 5.4 this 90% must be stated as a 10% chance of observing one or more values larger than the largest of the reference data set. Solving equation 5.4 for n with p=0.1 gives n=9N. Thus, if the maximum of 9 weeks of data is taken, there is a 90% certainty that another single (N=1) weekly maximum will not exceed this value. For a 90% certainty for all the weekly maxima over a year, 9 years of data is needed.

5.3 MOMENTS OF THE DISTRIBUTION OF EXCEEDANCES

The moments of this distribution may be obtained from properties of the hypergeometric and binomial functions (Gumbell, 1958, section 2.2.2). The mean number of exceedances over the mth largest value in N future trials is:

$$(5.6) \qquad \bar{x}_m = m^* \frac{N}{n+1} \quad .$$

The mean number of exceedances over the smallest value (m=n) is n times the mean number of exceedances over the largest value (m=1). Clearly the mean increases with m. If N=n+1, the mean number is m. If n is odd and m = (n+1)/2, the mean number of exceedances over the median is N/2. If both n and N are large, the mean number of exceedances over the largest value is approximately unity.

The variance of the number of exceedances over the mth largest value is:

$$(5.7) \qquad V_m = \frac{m*N*(n-m+1)(N+n+1)}{(n+2)(n+1)^2} \quad .$$

From this formula it can be seen that the variance increases with increasing N and decreases with increasing n. The variance is maximum for m = (n+1)/2; that is, for the median of the original observations.

The quotient of the variances of the number of exceedances above (greater in magnitude) the median, and above the extremes is:

$$(5.8) \qquad \frac{V_{(n+1)/2}}{V_1} = \frac{(n+2)^2}{4*n} = \frac{V_{(n+1)/2}}{V_n} \quad .$$

Consequently, the variance of the number of exceedances above the median is about n/4 times as large as the variance of the number of exceedances above the extremes. In this sense, the extremes are more reliable than the median and this quality increases with increasing sample size.

The variance of the exceedances is larger than the corresponding binomial variance because the probability is a known parameter for the binomial case, while for exceedances only the rank of the observation, m, that corresponds to the probability is known. For N=n+3 the variance of exceedances is approximately twice the variance of the corresponding Binomial distribution.

The coefficient of variation, CV, is obtained from:

$$(5.9) \qquad CV^2 = \frac{V_m}{\bar{x}_m^2} = \frac{(N+n+1)(n-m+1)}{N(n+2)m} \quad .$$

The following simple example using these formulas shows the effect of sample size on the accuracy of estimates. Suppose 9 readings of radon exposure to workers who have just finished working in an isotope storage building. Compute an estimate of how many of the 20 workers on the next shift will be exposed to more than the mean of the previous shift. Using formula 5.6 with N=20, n=9, and m=5 (the middle reading of the nine) gives a mean number of exceedances of 10. This is, of course, intuitive since the means of the past and future groups should be the same. If some change in exposure is suspected and a simple test for such a change is to be done, formula 5.7 could be used to obtain a standard deviation on the count of exceedances and if the actual count is more than two standard deviations from the expected count one would conclude that a change occurred. However, such a test may not be possible with small samples because the relative accuracy is large. Using formula 5.9 for this example yields a relative accuracy or coefficient of variation of 37%. This says the mean cannot be estimated very well. Ten workers times 37% is about 4, two standard deviations would be 8 workers. Thus a mean plus or minus two standard deviations would include a larger range than the sample size itself, not a very useful statistic.

If there had been 900 workers on the previous shift, and 2000 on the next, formula 5.4 can be applied to get a mean of 1000. Formula 5.9 gives a coefficient of variation of 4%, showing that the relative accuracy has increased greatly. The mean times the coefficient of variation gives a value of 40 workers. So while the relative accuracy increases, the absolute accuracy decreases.

The median number of exceedances is found by summing values of w over increasing values of x until the cumulative probability is 1/2. Let M be the value of x that corresponds to the median. Then the median can be found by solving the following equation for m.

$$(5.10) \qquad \sum_{z=0}^{M} w(z;n,m,N) = 1/2 = \sum_{z=N-M}^{N} w(z;n,n-m+1,N)$$

Such a number need not exist. For example, if N=n, then w(o;n,1,N) exceeds 1/2, and the distribution of the number of exceedances over the largest value does not posses a median.

5.4 EXAMPLE: Nitrous Oxides in Urban Air

Typical nitrogen oxide levels for urban areas were used to simulate the data in the following computational example.

Suppose a small industrial area has a good air pollution control system. Their regulations require that some types of industrial operations shut down until weather conditions change when daily nitrogen oxide levels exceed 0.1 part per million. From the previous summer's data one finds that 90 daily measurements were made and on 12 of these days the criteria was exceeded. During the coming summer how many times per month (30 days) should the industries expect to have to shut down?

Using Formula 5.6, the mean number of exceedances is estimated to be 3.95. Formula 5.7 gives a corresponding standard deviation of 2.13. These give an approximate 95% confidence interval of -.21 to 8.12 exceedances. Rounding this interval to the nearest integers gives 0 to 8 exceedances.

These approximations can be checked using formula 5.1 to compute the exact probability of any specified number of exceedances. Let m=12, N=30, n=90, and x vary from zero to 12 or 15 to include all significant

probabilities. The computational difficulties caused by large factorials are a significant consideration with these formulas. In such a sequence of probabilities a recursive formula can often be found for obtaining a probability in a simple way from the previous member of the series. Formula 5.1 we can be used to compute the probability, w, for x=0, then the only part of the formula that changes for values of x=1,2,3,... is the bottom term in one of the binomial coefficients of the numerator and in one of the denominator. A little algebra with binomial coefficients shows that:

$$\binom{I}{J} = \binom{I}{J-1} * \frac{I-J+1}{J} \quad .$$

This formula gives a simple recursive relationship for calculating successive values of w as x increases sequentially. The numerator and denominator of formula 5.1 are each multiplied by a simple fraction. This was done to produce Table 5.1.

TABLE 5.1
Distribution of Exceedances for
Nitrous Oxide Example

x	Probability	Cumulative Prob.
0	.02598	.02598
1	.08660	.11258
2	.15256	.26514
3	.18806	.45320
4	.18134	.63454
5	.14507	.77961
6	.09977	.87938
7	.06036	.93974
8	.03265	.97239
9	.01596	.98835
10	.00711	.99546
11	.00290	.99836
12	.00109	.99945
13	.00038	.99983

For values of x greater than 13 the probabilities become so small that these numbers of exceedances are of no consequence. This table shows that the mode occurs at x=3 (the largest single probability). The median can be found by interpolating between x=3 and x=4 to find a "point" at which the 50% cumulative probability would occur; this value is 3.26. The theoretical mean can be found by summing the product of the number of exceedances multiplied by the corresponding probability; this sum is 3.952, very close to the answer given by formula 5.6. The variance can likewise be calculated from formulas for the second moment about the mean. The variance thus estimated is 4.46, which yields a standard deviation of 2.11. This is close to the value given by formula 5.7.

This table can also be used to find confidence limits on the number of exceedances expected in a future 30 day period. Suppose a 95% confidence interval is of interest. The table is searched for values of the cumulative probability close to 0.025 and 0.975. For x=0 the probability slightly exceeds 0.025, so this value of x should be included in the interval. Now the 0.95 point rather than a 0.975 point is required because the interval does not exclude any values at the lower end. This leads to two choices for the upper end, 7 or 8 exceedances. Seven exceedances has a cumulative probability of 0.94 which is closer to 0.95 than the 0.97 cumulative probability associated with 8 exceedances. However 8 would be chosen if the confidence interval is to be <u>at least</u> 95%. There is another possibility. The interval 1 to 8 exceedances has a probability of 0.946 and this is as close to 95% as can be achieved. This is a similar situation to that discussed in Section 4.3 where a variety of possible confidence intervals was found when working with Bernoulli distributions. The simulation presented in the following paragraphs suggest that the 1 to 8 exceedances is the best interval, but generally, simulations do not provide a very strong justification.

A simple computer program was written to perform 500 simulations of 30 day nitrogen oxide readings and to count the occurrences exceeding 0.1 ppm. The results of this simulation are given in Table 5.2.

TABLE 5.2

Summary of Simulated Nitrous Oxide Exceedances

Count	Observed Frequency	Cumulative Frequency	Observed Probability	Cumulative Probability
0	8	8	.016	.016
1	36	44	.072	.088
2	65	109	.130	.218
3	79	188	.158	.376
4	98	286	.196	.572
5	81	367	.162	.734
6	56	423	.112	.846
7	34	457	.068	.914
8	23	480	.046	.960
9	12	492	.024	.984
10	5	497	.010	.994
11	3	500	.006	1.00
12	0	500	.000	1.00
13	0	500	.000	1.00

The number of exceedances in 30 days of simulated nitrogen oxides measurements ranged from 0 to 11, with a mean of 4.30, a mode of 4, and a standard deviation of 2.25. The median can be approximated by linear interpolation between the counts of 3 and 4 to find a value associated with a cumulative probability of 0.5. This value is 3.63. A confidence interval of 1 to 8 is closest to 95% (actually 94.4%).

Table 5.1 gives the theoretical values that should occur in the probability columns of Table 5.2. The theoretical values of the mean, median, mode, and standard deviation are just slightly smaller than the "observed" values. To see how well the data in Table 5.2 fit the

theoretical distribution, a Chi-square goodness of fit test can be used. First the probabilities in Table 5.1 are multiplied by 500 to get expected frequencies. These are compared with the observed frequencies in Table 5.2 with the Chi-square test. The Chi-square value is 15.6 with 11 degrees of freedom. This corresponds to about an 85% confidence level, so the fit is acceptable but not really good. The 11 degrees of freedom from the 14 rows in the tables is a result of grouping rows for 10 through 13 to avoid small frequencies in the Chi-square calculation.

5.5 SUMMARY

The probability that the mth largest among n observations will be exceeded x times in N future trials is given by the Distribution of Exceedances. It is analogous to a Binomial distribution except that the probability of success is a variable quantity. The mean number of exceedances is the same as the mean of the corresponding Binomial distribution. However, the variance is larger. This variance of the number of exceedances is largest for the median value of the initial distribution, and smallest at its extremes. This advantage of extremes increases with sample size.

In 1/2 of all cases the largest (or smallest) of n reference observations will never (always) be exceeded in n future trials.

5.6 REFERENCES

Gumbel, E.J. 1958, Statistics of Extremes, Columbia University Press, N.Y., N.Y., Library of Congress No. 57-10160.

Hastings, N.A.J., and J.B. Peacock, 1974, Statistical Distributions, John Wiley & Sons (Halstead Press), ISBN 0-470-35899-0.

Wilk, S.S., 1942, "Statistical Prediction with Specific Reference to the Problem of Tolerance Limits", Ann. Math. Stat., 13, pp 400-409.

CHAPTER 6
MORE ABOUT EXCEEDANCES

6.1 INTRODUCTION

The previous chapter introduced a general form of the Distribution of Exceedances, emphasizing that this distribution is a generalization of the Binomial distribution, but using an estimate of the probability of success. The asymptotic behavior of this distribution was not given for the case when the two sample sizes are large and rare exceedances are of interest. These asymptotes are one of the subjects of this chapter.

6.2 EXTRAPOLATION FROM SMALL SAMPLES

It is common with environmental studies to make rather sweeping statements about future events based on limited previous information. In terms of the Distribution of Exceedances, this is equivalent to assuming a large N and a small n. Typically m is also small. Instead of using x, $q = x/N$, $(0 < q \leq 1)$, the proportion of future exceedances is used. Since N is large, x and therefore q can be approximated as continuous variables. Gumbell (1958, section 2.2.5) shows that the distribution of q is given by

$$(6.1) \qquad f(q;n,m,N) = N*m*\binom{n}{m} \frac{N!(qN+m-1)!(n-qN+n-m)!}{(N+n)!(qN)!(N-qn)!} \quad .$$

Stirling's formula leads to the approximation

$$(6.2) \qquad f(q;n,m) = \binom{n}{m} * mq^{m-1}(1-q)^{n-m} \quad .$$

The associated cumulative distribution function is

$$(6.3) \qquad F = \int_0^q f(t;n,m)dt \ .$$

The solution of this integration is a recursive integral function of m and n. Appendix A shows how to find the value of this interval from tables of the Incomplete Beta distribution.

The probability that at most some proportion q of the new observations will exceed the smallest of the n previous observations is

$$(6.4) \qquad F(q;n,n) = q^n \ .$$

The probability that in a future large sample at most some proportion q will exceed all of the old observations is

$$(6.5) \qquad F(q;n,1) = 1-(1-q)^n \ .$$

By symmetry this is also the probability that at most a fraction q will be less than the smallest value in the original sample.

6.2.1 Example; Design of Experiments

Formulas 6.1 through 6.5 can be used to design an experiment when N (a new sample size) is unspecified but known to be very large. Suppose one wishes to collect enough reference data to get a 90% chance that at most 10% of future data will exceed the largest value in the reference data set. Using formula 6.5 this problem can be set up as q = 0.1 and F = 0.9. Solving for n

$$0.9 = 1-(1-.1)^n \quad \text{or} \quad n = \log(.1)/\log(.9) = 21.85$$

Applying this to the example of weekly air pollution maxima, the maximum of 26 weeks of data has a 90% chance that, of all future weekly maxima, only 10% will exceed this maxima.

6.3 THE LAW OF RARE EXCEEDANCES

If both n and N become large, two special cases are of interest. In the first, the rank m increases with n such that the quotient m/n approaches a constant value, and the mth value is near the median. In the second, m is constant and much smaller than n so that m indexes extreme or rare values.

For the first case consider the situation where n = N = 2k-1. Since N and n are large then k also will be large. Then m=k is the rank of the median of the initial distribution, and to a very close approximation, m = N/2 = n/2. Gumbel (1958, section 2.2.6) shows that for large N and n, and m in the neighborhood of the median, the number of exceedances over the mth value is asymptotically Normally distributed with both mean and variance equal to k. This variance is very large relative to the mean. This is called the Distribution of Normal Exceedances.

In the second case, N and n are large, and m and x are small. Gumbel shows that

$$(6.6) \qquad w(x;n,m,N) \simeq \binom{x+m-1}{x} \frac{n^m N^x}{(N+n)^{m+x}} .$$

The probability that the mth value is never exceeded is the situation where x=0;

$$(6.7) \qquad w(0;n,m,N) \simeq \left(\frac{n}{N+n}\right)^m .$$

The probability that the largest value is exceeded in x future observations is

$$(6.8) \qquad w(x;n,1,N) \simeq \frac{n}{N+n} * \left(\frac{N}{N+n}\right)^x .$$

This is a geometric series decreasing with x. When N=n, formula 6.6 becomes

$$(6,9) \qquad w(x,n,m,n) \simeq \binom{x+m-1}{x} * (1/2)^{m+x}.$$

which is the asymptotic probability that the mth largest value will be exceeded x times in N future trials. This probability is independent of n and contains the single parameter m. Since m is small compared to n, this is called the Law of Rare Exceedances, and denoted simply as

$$(6.10) \qquad w(x;m) = \binom{x+m-1}{x} * (1/2)^{m+x}.$$

The probability that the largest value will be exceeded x times in future observations is obtained by substituting m=1 into 6.10, giving

$$(6.11) \qquad w(x,1) = (1/2)^{x+1} \quad .$$

It follows that the probability that the largest value previously observed will be exceeded at most x times in future observations is

$$(6.12) \qquad w = 1 - (1/2)^{x+1} \quad .$$

This probability converges rapidly to unity as x increases. These asymptotic formulas are useful because they are independent of sample sizes. But they can be misinterpreted if one forgets that they assume large samples for both the reference and the future data sets, and that the underlying distribution is constant over time.

The mean and variance of the distribution of rare exceedances can be obtained from formulas 4.7 and 4.8. The mean is m and the variance is 2m. Thus, this distribution is similar to a Poisson distribution except that the variance of rare exceedances is twice that of a corresponding Poisson variate. The difference is intuitively justified. If the Poisson law were

applied to rare exceedances, it must be assumed that the mean number of exceedances is known. With rare exceedances a sample estimate of this mean is used. Consequently, the variance must be larger than for the Poisson case.

Both the distribution of rare events (Poisson) and the distribution of rare exceedances may be standardized by y = (x - mean)/SD, so that y converges to a standardized Gaussian distribution.

The variance of rare exceedances, 2m, is much smaller than the variance of normal exceedances, N/2. The variance of rare exceedances is smallest for m=1, the largest value observed.

6.4 RETURN PERIOD

The concepts presented so far will now be used to develop useful tools for the next chapter, which introduces the magnitudes of extreme values. The first of these tools is the return period, important when time is a statistical variable of interest. Flood control engineers are interested in the time interval between floods; the mean of these intervals is the return period for floods.

Consider first a discrete variate generator, for example dice. The probability that any specified face occurs on a toss is 1/6. Therefore the specified face is expected, in the long run, and on the average, once in six trials. For a continuous variate there is no probability for any specific value of the variate, such as x, so a dichotomy is constructed. The probability of observations equal to or larger than x is 1-F(x), where F is the distribution (cumulative density) function of the variate x. Observations are made at regular intervals of time and an experiment stops when a specified value X of x has been exceeded once. The probability that this exceedance happens on trial v is to be found. The variable v has a geometric distribution with probability parameter p = 1-F(X). The <u>return</u>

period is defined to be the mean value of v, denote this mean by T(x).

$$(6.13) \qquad \bar{v} = \frac{1}{p} = T(x) = \frac{1}{1-F(x)}$$

The variance of T(x) is T^2-T, and the median number of v is

$$(6.14) \qquad \frac{.69315}{-\ln(1-1/T)} \approx 0.69315*T - 0.34657 \quad .$$

The mean is about 44% larger than the median, and there is as much chance for the event to happen prior to .69*T(x) as after.

Every distribution has a return period function, and every return period has an associated distribution. Thus it is incorrect to write down an arbitrary function and call it a return period.

The return period is most interesting if observations are made at equidistant intervals of time. Then the return period can be identified as a number of observations. This is the origin of the name.

As an example, suppose a measurement is mace daily and the largest value in one year is of interest. The return period is the number of 365-day periods (years) that would on the average elapse before an exceedance of the specified magnitude would occur again.

The return period of the median is 2, of the upper quartile is 4, and so on. Starting at the median, the return period increases with increasing values of the variate. For values smaller than the the median, the return period is smaller than 2. For the first quartile it is 4/3. The return period converges to unity for decreasing values of the variate.

The value of x for which the return period is doubled, D, is found from

(6.15) $F(D) = (1+F(X))/2$.

Conversely, the return period of 2X is the solution for

(6.16) $T(2X) = 1/(1-F(2X))$.

For a given distribution, F, D is obtained as a function of X, and T(2X) as a function of T(x) since X is a function of T.

6.5 EXPECTED EXTREMES AND EXTRAPOLATION TO LOW DOSES

Let F be the cumulative density function of the previous section, and let n be the number of observations in a (large) sample. Then a specific large value of the variate, call it u[n], is uniquely defined by stating that its cumulative probability is defined by

(6.17) $F(u[n]) = 1-1/n$.

This equation is another way of writing the return period since n is analogous to T(x). The equation may be rewritten as

(6.18) $n(1-F(u[n])) = 1$.

In this form the product on the left side is the number of values equal to or exceeding u[n]. Since this product is unity, u[n] is called the <u>expected largest value</u>. Note that the expected largest value is not the mean largest value (which in the next section will be shown to be determined from $F(u) = 1 - 1/(n+1)$). By symmetry the expected smallest value is

(6.19) $F(u[1]) = 1/n$.

The two percentiles u[n] and u[1] are functions of n and differ for different distribution types and for parameter values within a given distribution type. If the initial data distribution is symmetric, the two

expected extremes are equal in size about the mean but differ in sign. Equation 6.19 is used implicitly by authors who use the "weakest link" argument to establish environmental criteria. This argument contends that the criteria should be such that even the most susceptible of a population should not be affected by the pollutant. With an assumed form for F at small values of the argument, and an estimate of population size, 6.19 can be used, and u[1] becomes the basis of pollutant criteria.

The current controversy about extrapolation to low doses for carcinogenicity criteria can be interpreted as a disagreement about the form of the density function F. Usually reasonable arguments can be found to establish n, but agreement on F is seldom realized. This is important because this procedure uses extrapolation to the tails of the curve. By using the asymptotic theory for magnitudes of response (to be studied in later chapters) this problem can be studied with limited information about the function F. It has proven difficult for many persons to accept the concept that the asymptotic theory allows one to establish criteria without complete knowlege of the response function. For example, in carcinogenesis studies arguments are common about how the response at low doses should be modeled; by linear extrapolation, by quadratic extrapolation, or if a backround response should be considered, and so on. The asymptotic theory shows that such details can be irrelevant. Agreement is necessary only on the type of distribution: Exponential class, Cauchy class, or Weibull class.

6.6 PLOTTING POSITION

For this section, it is convenient to change the notation so that observations are ranked from the smallest to the largest. In order to eliminate as much confusion as possible, the symbol r will denote this new ranking, and the symbol m will be retained as the rank from the largest to smallest. For a sample of size n, m = n-r+1 and r = n-m+1.

Probability plotting, which is described in Section 7.2, is an extremely useful graphical tool for working with the Extreme Value distribution. It is easily done with ordinary graph paper and an electronic hand calculator with scientific features. For this tool it is necessary to discuss the ways available to calculate plotting position, and choose the best one for extreme values. In probability plotting the data itself is plotted on one axis, and on the other axis is plotted an expected probability transformed by the inverse of the density function. The standardized Extreme Value distribution is a double exponential (see appendix of Chapter 3). The inverse is calculated as a double natural logarithm of the appropriate percentile. It is not easy to decide how this percentile or expected probability should be determined. There is a substantial amount of published literature about this problem, and most of it is applicable only to the Gaussian (Normal) distribution.

Of the many proposed formulas for calculating the expected probability, the following three are in common use because of their simplicity and near optimum statistical properties:

1) $p = (r-1/2)/n$
2) $p = r/(n+1)$
3) $p = (r-3/8)/(n+1/4)$.

Kimball (1960) discusses these and some others that are rather difficult to compute.

If the first of these probabilities is used to compute an expected return period for the largest observation, it leads to a logical contradiction. In formula 6.13 replace the cumulative probability, $F(x)$, by the percentile of the largest observation, $p = (n-1/2)/n$. This gives

$$T(x_n) = \frac{1}{1 - \frac{n-1/2}{n}} = 2n \quad ,$$

which claims that an event which has happened once in n trials will on the average occur once in 2n trials.

The second choice of expected probabilities is based upon percentiles. The expected value of the rth value from a sample of size n from a uniform distribution on the unit interval (rectangular distribution on the interval 0 to 1) is r/(n+1). A statistical theorem (Mood & Graybill, 1963, Theorem 6.1) gives the following rule for relating any density function to the density of the unit uniform:

Theorem
Any density for a continuous variate X may be transformed to the uniform density f(y) = 1, 0 < y < 1, by letting Y = F(X), where F(x) is the cumulative distribution of X.

Then for any distribution represented by its density function F, solving r/(n+1) = p = F(x) for x yields the expected value of the pth percentile of that distribution.

The third choice of expected probabilities was developed for graphical estimation of parameter values of a Gaussian distribution. The slope of the line on a Gaussian probability plot can be used to estimate the standard deviation. This third choice was developed to give almost unbiased estimates of the standard deviation from the slope of a regression line on these plots.

Kimball (1960) found that the third choice is best for estimating the variance of a Gaussian distribution. All three choices do reasonably well for estimating the mean of a Gaussian distribution. The second choice was found best for extreme value work because it gives almost unbiased graphical estimates of the parameters of the Extreme Value distribution. For the remainder of this text only p = r/(n+1) will be used.

6.7 SUMMARY

This chapter extends the Distribution of Exceedances to the consideration of counting the frequencies of rare events. The concepts that are introduced were then used to develop some tools that will be useful in the next chapters. These tools are the Return Period, Rare Exceedances, Expected Extremes, and plotting position.

Also, these statistical tools are _sometimes the only reasonable statistics available for many applied problems. With current environmentally sensitive and politically active demands upon science, one could be asked to analyze extremely sparse and incomplete data. What conclusions can be reached, for example, when the available data is that three toxicity cases have been observed in 10 to 20 thousand workers exposed to compound X? Ordinary statistics are of no value in such situations. Nothing might be known about the magnitudes of exposure or the distribution of responses. Still, an estimate of a return period and its standard deviation is useful, expected extremes can be discussed, and the laws of Rare Exceedances can be used.

6.8 REFERENCES

David, H.A., 1970, Order Statistics, John Wiley & Sons Inc. N.Y. N.Y.

Gumbell,E.J., 1958, Statistics of Extremes, Columbia University Press, N.Y. N.Y.

Kimball, B.F., 1960, "On the Choice of Plotting Positions on Probability Paper", J. American Statistical Assoc. 55, pp. 546-60.

Mood, A.M., & Graybill,F.A. 1963, Introduction to the Theory of Statistics, McGraw-Hill, N.Y. N.Y.

6.9 APPENDIX 6-A

A useful identity allows one to find the cumulative probability of formula 5.3 from tables of the Incomplete Beta distribution. Formulas 5.2 and 5.3 are:

$$f(q;n,m) = m\binom{n}{m}q^{m-1}(1-q)^{n-m},$$

$$F(q;n,m) = \int_0^q f(t;n,m)dt = m\binom{n}{m}*\int_0^q t^{m-1}(1-t)^{n-m}dt \quad .$$

For typographical purposes indicate the Gamma function with G( ). Then the incomplete beta function is:

$$I(q;a,b) = \frac{G(a+b)}{G(a)G(b)}*\int_0^q t^{a-1}(1-t)^{b-1}dt \quad .$$

Equating powers within the integrals for I(q;a,b) and F(q;n,m) gives:

$$m-1 = a-1 \text{ or } a = m$$
$$b-1 = n-m \text{ or } b = n-m+1 \ .$$

Next consider the multipliers of the integrals.

$$m\binom{n}{m} = m\frac{n!}{m!(n-m)!} = \frac{m*n!}{m(m-1)!(n-m)!} = \frac{n!}{(m-1)!(n-m)!} \ ,$$

$$\frac{G(a+b)}{G(a)G(b)} = \frac{G(m+n-m+1)}{G(m)G(n-m+1)} = K \quad .$$

For integer values, G(a) = (a-1)!,

$$K = \frac{(m+n-m+1-1)!}{(m-1)!(n-m+1-1)!} = \frac{n!}{(m-1)!(n-m)!} \ ,$$

and this is identical to the multiplier for F( ).
Thus, F(q;n,m) = I(q;m,n-m+1) .

Tables of the Incomplete Beta distribution are given in Beyer (1966). A useful identity for the Incomplete Beta distribution is:

If $I(q;a,b) = P$
then $I(1-q;b,a) = 1-P$.

As an example of using this identity consider the probability $F(.25;8,1)$. From formula 6.5 this value is:

$$1-(1-.25)^8 = .9 .$$

The tables in Beyer (1966) must be used in an inverse interpolation mode. Beyer has tables of $I(q;a,b) = P$ with successive tables for various values of P, q in the body of the table, and a and b indexing the rows and columns of the tables. Look in these tables in rows and columns of 8 and 1 respectively for an entry of approximately 0.25, no such value can be found. Invoking the 'useful identity' look in rows and columns of 1 and 8 for an entry of approximately .75, it is found in the table for P=10%. That is:

If $I(1-q;b,a) = 1-P$ then $I(.75;1,8) = .1$
and if $I(q;a,b) = P$ then $I(.25;8,1) = .9$

which is the desired result.

Appendix Reference

Beyer, W.H. (Ed.), 1966, Handbook of Tables for Probability and Statistics, The Chemical Rubber Publishing Co.

CHAPTER 7
THE MAGNITUDE OF EXTREME VALUES

7.1 INTRODUCTION

The previous chapters have concentrated upon counting the number of extreme values. This chapter begins the study of the magnitude of extremes, but first, a word of caution. The magnitude of the exposure (e.g. the level of styrene vapors in air) can be measured or the number of persons ill from the insult can be counted. However, it is important to distinguish that this is not the same as the magnitude of the response, e.g. the degree of illness. The error of equating the magnitude of the exposure to the magnitude of the response is often found in articles and reports.

This chapter assumes that a ratio or interval scale measurement is available and appropriately defined for the problem at hand.

7.2 EXPLORATORY DATA ANALYSIS OF EXTREMES

Simple graphical methods are a good starting point for any data analysis. Probability plotting is easy in extreme value distribution work. In a sample of n independent observations, one of them (or perhaps several identical ones) is the smallest or the largest. If N such samples of size n are gathered, a sample of N extreme values is obtained. The distribution of this sample is of interest under the conditions that n is large, that the variate, say x, is unlimited in the direction of the extreme under consideration (largest or smallest values), and that the initial distribution sampled is from the Exponential family. It was noted in the Chapter 2 that a transformation can change variables from the Cauchy family and from the Weibull family to the Exponential family.

Let f(x) be a density function, and let F(x) be the corresponding distribution function. In Chapter 6 a large value, u, was defined as the expected extreme using the cumulative probability formula

(7.1) $F(u) = 1 - 1/n$.

Define a new parameter, a, by

(7.2) $1/a = n*f(u)$.

Then, for the exponential family, the asymptotic probability (distribution) for the largest value, denoted as x[n], is (Gumbel, 1958, Chapter 5)

(7.3) $H(x[n]) = \exp(-\exp(-y))$, where

(7.3') $y = (x[n] - u)/a$.

The variable y is defined to be the reduced variate. This is analogous to the familiar standardized Gaussian (Normal) variate. The parameter u is a measure of the central tendency of the extreme value distribution, but it's not the mean of that distribution. Likewise, the parameter a (or more exactly, 1/a) is a measure of dispersion, but it's not the standard deviation. Usually the indicator of sample size is dropped from the notation unless it is variable in the problem at hand, and x[n] is replaced with x.

Formula 7.3 can be defined as a function of y rather than of x. Then, just as for the standardized Gaussian distribution, a single reduced extreme value distribution can represent all possible extreme value distributions. The reduced distribution is denoted by the expression:

(7.4) $H(y) = \exp(-\exp(-y))$.

This reduced extreme value probability distribution function has an

important advantage over the standardized Gaussian distribution; namely, the inverse of the extreme value distribution is easy to compute,

(7.5) $y = -\ln(-\ln(H(y)))$.

The corresponding inverse of the standardized Gaussian distribution is the inverse of a non-analytical exponential integral.

The asymptotic probabilities of the smallest values are obtained by changing y into -y and H(x) into 1 - H(x). Thus, only the largest extreme values need be considered.

The parameter definitions 7.1 and 7.2 require knowledge of the parent density or distribution function. Since such knowledge is lacking in most cases, a method is needed to estimate these parameters from the observed largest sample values alone. A mathematical study of the Extreme Value distribution shows that u is the modal largest value in a sample of size N, and that 1/a is the rate of increase of the most probable largest value with the natural logarithm of the number of samples N, and is proportional to the standard deviation of the extremes.

If the data are from any distribution that is in the Exponential family, then the N observed extreme values x[m] (m=1,2,3...N), ordered in increasing magnitude, should be scattered about a straight line when plotted against their expected cumulative relative frequencies. The quantity used for the expected cumulative relative frequency is obtained by substituting

(7.6) $H(y) = \overline{H(\bar{x}[\bar{m}])} = m/(N + 1)$

into equation 7.5. (The over-line indicates average or sample expected value.) The rationale for this was discussed in Section 6.6.

Typically, the observed magnitudes of the N extremes, x, are plotted vertically, and the corresponding y values, the solution to equations 7.6 and 7.5, are plotted horizontally. This arrangement, opposite that usually employed in statistics for plotting cumulative distributions, has been adopted in extreme value statistics in order to have sampling variation operative in the vertical direction only, as is customary for curve fitting. The values of x depend upon the experimental measurements, and the values of y are determined by the sample size, N, and the index associated with the order of the data values. The values of y should usually be within the range of -2 to +8. These (x,y) pairs should then scatter about the line

(7.7) $x = u + a*y$.

Using the Return Period defined in Section 6.4, and equations 7.3, 7.5, and 7.6, the return period can be defined as:

(7.8) $T(x[m]) = 1/(1 - H(y)) = 1/(1 - m/(N+1)) = (N+1)/(N-m+1)$.

This gives the average number of observations necessary to obtain one value equal to or larger than x. For large values of x, the return period converges towards exp(y).

These equations facilitate probability plotting of extreme values on ordinary linear-linear graph paper. Extreme value probability paper can sometimes be found, it was an important tool before scientific hand-held calculators made exponentiation a trivial operation. These special probability papers have linear scales for the observed variate x and the reduced variate y. They also include, parallel to the y scale, two nonlinear scales for the return period and cumulative probability or frequency. The relationship between these three variables is shown in the following BASIC computer program, which prints a table of corresponding values of the three quantities. In this program Y is the reduced variate,

P is the cumulative probability, and T is the return period. This program is an implementation of equations 7.3, 7.5, and 7.8.

```
10 FOR Y=8 TO -2 STEP -.1
20 P=EXP(-EXP(-Y))
30 T=1./(1.-P)
40 PRINT Y,P,T
50 NEXT Y
60 END
```

A probability plot of an extreme value data set is always advisable, even when the data is automatically collected or is a replicate of previous data. A deviation from a straight line plot can easily be spotted. Whenever a curved line is suspected, first check if the extremes were correctly collected and recorded. Then one of the logarithmic transformations of the x variable discussed in Section 3.4, should be plotted to see if the data then plot as a straight line. A formal statistical test to determine if the data conforms to a Cauchy, Weibull, or an Exponential extreme value distribution is beyond the level of this text. One possible test is to use a general algorithm for maximizing the likelihood function for the observed data with each of the three extreme value distributions, then compare the goodness-of-fit using the likelihood ratio test.

7.2.1 Probability Plot Example

In an urban area, the annual maxima of weekly average parts per million nitrous oxide levels for each of 10 years were: 0.108, 0.063, 0.111, 0.077, 0.081, 0.085, 0.097 0.083, 0.078, 0.062. These 'x' values were ranked from smallest to largest, their cumulative relative frequencies calculated using equation 7.6, and their reduced variates, y, calculated using equation 7.5. The results are:

TABLE 7.1
Observed Data and Reduced Variate

m	m/(N+1)	x	y
1	0.091	0.062	-0.875
2	0.182	0.063	-0.533
3	0.273	0.077	-0.262
4	0.364	0.078	-0.012
5	0.455	0.081	+0.238
6	0.545	0.083	0.501
7	0.636	0.085	0.794
8	0.727	0.097	1.144
9	0.818	0.108	1.606
10	0.909	0.111	2.351

Figure 7.1 is a plot of the data in Table 7.1 with the x values on the vertical axis and y values on the horozontal axis.

This plot suggests a linear relationship of x and y. A least-squares linear regression of x on y gives: $x = 0.0767 + 0.0162y$. The standard errors of these regression coefficients are 0.0014 and 0.0013, respectively. Equation 7.7 relates these regression estimates to the scale, a, and position, u, parameters of the Extreme Value distribution.

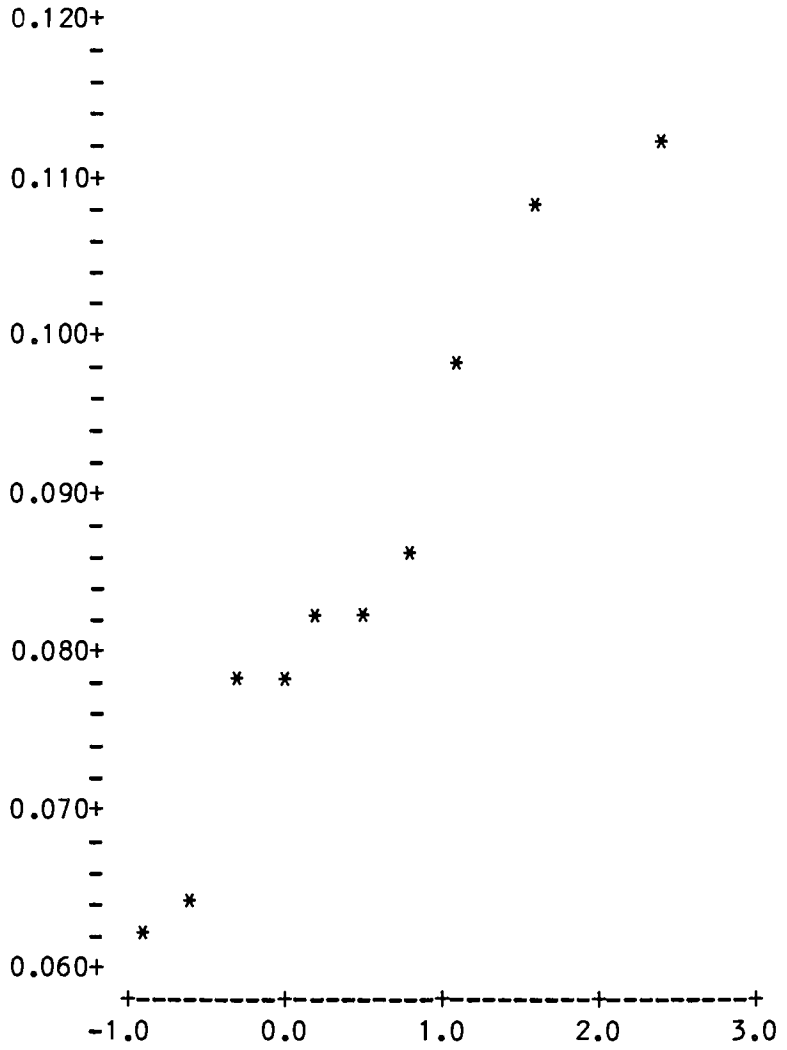


FIGURE 7.1
Observed Nitrous Oxide in ppm on Ordinate,
Reduced Variate on Abscissa.

7.3 MAXIMUM LIKELIHOOD ESTIMATES

Although these graphical and regression estimates of the parameters of the extreme value distribution are easy to compute, they are accompanied by an important statistical question of bias. For example, there is good reason to argue that a weighted regression should be used. Most computer centers have algorithms available for the generalized maximization of any

analytic (well behaved) function. These algorithms have such names as: Davidon-Fletcher-Powell, Fletcher-Reeves, Conjugate Gradients, Newton-Raphson, and Nedler-Mead Simplex. For illustration the Nedler-Mead Simplex algorithm (O'Neill, 1971) will be used to directly maximize the likelihood function of the extreme value distribution.

The extreme value density is obtained by differentiating the distribution 7.3 to get:

$$h(x) = \exp(-y - \exp(-y))/a \quad , \tag{7.9}$$

$$\text{where } y = (x - u)/a \quad . \tag{7.9'}$$

(Equation 7.9' is identical to 7.3'.) The likelihood function is the product of the h(x)'s for the observed values of x:

$$L = \prod_{i=1}^{N} h(x[i]) \quad . \tag{7.10}$$

Typically, the log-likelihood function is maximized,

$$l = \ln(L) = \sum_{i=1}^{N} \ln(h(x[i]) \quad . \tag{7.11}$$

The N = 10 data points x[i] of Example 7.2.1 (Table 7.1) were input to the SIMPLEX algorithm along with formulas 7.11 and 7.9. The algorithm searched for those values of a and u that maximized the likelihood for this data set. A simplification of 7.11 is possible:

$$l = N*\ln(1/a) - \sum_{i=1}^{N} (y + \exp(-y)) \quad . \tag{7.12}$$

Since all generalized functional maximization algorithms are iterative, they require initial parameter estimates. The results of the linear regression are reasonable initial estimates; the likelihood

maximization will tend to eliminate any biases. The results of this maximization for the example data are:

Parameter	Value	Standard Error
u	0.0771	0.0064
a	0.0136	0.0047

A comparison of these values with those obtained from the least-squares linear regression shows that the parameter values are about the same, but the standard deviations estimated by the regression are too small by a factor of almost five.

The asymptotic correlation between the maximum likelihood estimates for u and a is 0.313 (Johnson and Kotz, Chapter 21). The estimated correlation for these parameters in this example is 0.324.

7.4 FORMAL PROPERTIES

No attempt will be made to show the derivation of the formulas presented in this section. Some depend upon material presented in several chapters of mathematical development in Gumbell's book (1958).

7.4.1 Reduced Variate

The structural similarity of the reduced variate used in extreme value work, and the standardized Gaussian or Normal variate used in many well known applications of statistics is obvious. The reduced variate, $y = (x - u)/a$ given in equation 7.3', has location parameter u and scale parameter a. A standardized normal variate is obtained by subtracting the mean from a data value and then dividing by the standard deviation, for example

(7.13) $\quad t = (x - M)/SD$.

If the mean M and standard deviation SD are obtained from a sample of extreme values, then the reduced variate y, is related to the corresponding standardized variate t by:

(7.14) $t = SQRT(6)*(y-E)/pi$,
where E is Euler's constant 0.5772156649,
and pi = 3.1415927.... .

The mean value of a reduced variable, y, (the expected largest value) is Euler's constant. Then the expected largest value in terms of the sample extremes can be derived from equation 7.3'. Let ME signify the Mean Extreme value and SDE denote the Standard Deviation of Extreme values. Then,

(7.15) $ME = u + aE$

(7.16) $SDE = pi*a/SQRT(6)$.

The mean and standard deviation of a sample of extremes can be computed, then equations 7.16 and 7.15 can be used to estimate a and then u, the parameters of the reduced variable.

(7.15a) $\hat{u} = \bar{x} - aE$

(7.16a) $\hat{a} = s*SQRT(6)/pi$.

The variance of these estimates based on sample moments are approximately (Johnson and Kotz, Chapter 21)

$$V(\hat{u}) \simeq 1.1678\hat{a}^2/n$$

$$V(\hat{a}) \simeq 1.1\hat{a}^2/n \quad .$$

The effeciency of these estimators, relative to the maximum likelihood estimators, is about 95 percent for u and only about 55 percent for a.

The mean and standard deviation of the sample of 10 extremes used in Example 7.2.1 are 0.0845 and 0.0166, which yield estimates of a = 0.0130 and u = 0.0770. These estimates are a little closer to the maximum likelihood estimates than are those derived from the least-squares equation for this particular data set.

The return period may be estimated for large values of the reduced variate from the asymptotic relationship:

(7.17) $T(x) = \exp(y)$,

which Gumbell claims to be reasonably good for $y > 5$.

7.4.2 Relation of Parent Distribution to Extreme Values

If the mean, m, and variance, v, of a density function are known (small sample estimates will not suffice), then an upper limit on the mean extreme value of samples of size n from that density is given by

(7.18) $ME \leq m + v*(n - 1)/SQRT(2n - 1)$.

The variance of a set of extreme values is smaller than the variance of the parent distribution. However, one should never use extremes as a tool to make inferences about parent distributions because the extremes do not contain information about the central tendency of the parent distribution. Extremes are used to make inferences about other possible extremes. Appendic 7-C presents an example that illustrates how much in error conclusions can be if this warning is ignored.

It is valid to derive inferences about extremes from a known parent distribution. In practice, one would have a hypothesized distribution and wish to explore the behavior of extremes under the condition that the hypothesis is correct. In general such an investigation involves difficult

algebra and asymptotic theory, and has been analytically solved for only a few common parent distributions. If x(i) is a sequence of independent identically distributed (i.i.d.) Gaussian (Normal) random variables with a mean of m and standard deviation of s, then the maximum of n such variables W(n), when n is large, has the following Extreme Value distribution:

(7.19) $P(W(n) \leq z) = \exp(-\exp(-(z-u(n))/a(n)))$
where $a(n) = s*u(n)/SQRT(2*ln(n))$
$u(n) = s*c(n) + m$
$c(n) = SQRT(2*ln(n)) - (ln(ln(n)) + ln(4*pi))/(2*SQRT(2*ln(n)))$

Furthermore, the expected value of W(n) is:

$$EV(W(n)) = u(n) + E*a(n) .$$

It may be more convenient to standardize the parent distribution before determining the distribution of extremes. Then the following is equivalent to equation 7.19. If x(i) is a sequence of i.i.d. standardized Gaussian random variables (mean=0, standard deviation=1) then the maximum of n such variables W(n) has the following Extreme Value distribution:

(7.20) $P(a(n)*(W(n) - u(n)) \leq z) = \exp(-\exp(-z))$
where $a(n) = SQRT(2*ln(n))$
$u(n) = SQRT(2*ln(n)) - (ln(4*pi) + ln(ln(n)))/(2*SQRT(2*ln(n)))$.

7.4.3 Sample Size

Samples from an Extreme Value distribution also have an Extreme Value distribution with the same scale parameter. If x[i], i = 1,2,...n are n extreme values each with mode u and scale parameter a, then MAX(x[i]) has an Extreme Value distribution with

(7.21) mode = u + a*ln(n), and
(7.21') scale parameter = a .

Equivalently, if y is the reduced variate corresponding to the x's, then

the reduced variate corresponding to the maximum of n x's is

(7.22) $y' = y - \ln(n)$.

This equation is derived as follows:

$$\text{new reduced variate} = \frac{(x - (u + a*\ln(n)))}{a}$$

$$= \frac{x - u - a*\ln(n)}{a} = \frac{x - u}{a} - \ln(n)$$

$$= \text{old reduced variate} - \ln(n)$$

Equation 7.22 can be used to pass from one sample size of extremes to another that is not necessarily an integer multiple of the original sample size. If the x's are extremes from samples of size m then the maximum of n x's is the extreme from a sample of size m*n. On extreme value probability plots a change in sample size appears as a shift in the line to larger (or smaller) values of the variate, but not a change in the slope of the line. This allows for the extrapolation of the magnitude of extremes to larger (or smaller) populations than the one from which the sample of extremes was obtained. It is important to make the distinction between this and the counting of exceedances, discussed in Chapters 5 and 6, as population size increases. (Also, both situations are distinct from the situation of recognizing more exceedances in a fixed sample size because measurement methods are better.)

Suppose, from the example of maxima of weekly nitrous oxide averages for each of 10 years in Section 7.2.1, a prediction of a maxima for 25 years of data is required. Using the maximum likelihood parameter estimates and equations 7.21 gives

$$\text{mode} = 0.0769 + 0.0136*\ln(2.5) = 0.0896 .$$

The scale parameter remains at 0.0136. That is, if 10 years of data

resulted in a mode of 0.0769, then 25 years of data should give a mode of maximum weekly values of about 0.0896.

7.4.4 Other Statistics

The median extreme value, in the scale of the original variable x, can be calculated from

(7.23) $$\text{median} = u - a*\ln(\ln(2)) .$$

The corresponding values for the reduced extreme variate can be derived from the equations given above. They are tabulated in Table 7.2.

TABLE 7.2

Statistics of any Reduced Extreme Variate

Statistic	Value
Mean	Euler's Constant = 0.57722
Median	$-\ln(\ln(2)) = 0.36651$
Mode	0.0
Standard Deviation	pi/SQRT(6) = 1.28255

Finally, a table of cumulative probabilities of the reduced Extreme Value variate is not needed because they can easily be computed on a scientific hand calculator using the equation:

(7.24) $$P(y \leq k) = \exp(-\exp(-k)) .$$

Two commonly used probability statements are: the one-sided upper 95% confidence interval which has its limit at a y value of 2.97, and the 99% interval at 4.60. The two - sided 95% probability interval for the Extreme Value distribution is $-1.3 < y < 3.7$. Plus and minus two standard deviations of a sample of extremes about their mean encompasses 92.6% of the probability. For example 7.2.1, find the upper 95% limit for 25 yearly

maxima of weekly nitrous oxide averages. Substituting the mode of 0.0896, the scale parameter of 0.0136, and a y (reduced variate) value of 2.97 into equation 7.3', and solving for x gives:

$$2.97 = (x - 0.0896)/0.0136,$$
$$x = 0.130 .$$

This says that for 25 yearly maxima of weekly averages of nitrous oxide levels, there is 95% confidence that the overall maxima will not be greater than 0.130 parts per million.

7.5 GENERALIZED EXTREME VALUE DISTRIBUTION

On occasion none of the three asymptotic extreme value distributions will be applicable to a particular problem at hand. Maritz and Munro (1967) present a Generalized Extreme Value distribution which can describe extremes of small samples as well as large ones. This distribution is the three parameter function:

$$(7.25) \qquad F(x) = \exp\left[- \left[\frac{h(x - a)}{b} \right]^{1/h} \right] .$$

As h approaches zero this distribution approaches the Extreme Value distribution, it is of the Cauchy family for h less than zero, and of the Weibull family for h greater than zero.

Parameter estimates can be obtained using tables given in Maritz and Munro, or by using the generalized maximization of a likelihood function, discussed in Section 7.3. Special care is required when computing the likelihood function if h approaches zero, in this case roundoff error will cause severe computational problems. For some arbitrary small constant e, which depends upon the computer being used, a switch should be made from the Generalized Extreme Value density to the Extreme Value density whenever the estimated value of h is smaller than e.

7.6 SUMMARY

This chapter contains basic tools for working with the magnitude of extreme values. Since the Extreme Value distribution is just another statistical distribution like the Gaussian or the Student-t distributions, statistical tools like the mean, median, mode, standard deviation, and probability plotting can be used. In extreme value work emphasis is placed upon the mode rather than the mean, and on a scale parameter rather than the variance. Overall, the biggest difference between extreme value and the usual statistical procedures is in the way the data is collected; extreme value inference is concerned with only a small subset of all the data.

This chapter started with a discussion of data plotting, an important step in any data analysis. Then the reduced variate and the meaning of return period were discussed. These concepts were illustrated with an air pollution example. Maximum likelihood was presented as a good way of obtaining unbiased parameter estimates. And finally, many of the formal properties of the Extreme Value distribution were outlined. Beach (1975) has an interesting report that uses some of the statistics presented in this chapter.

7.7 REFERENCES

Beach, S. L., 1975, 'The Identification of a Homogeneous Critical Group Using Statistical Extreme Value Theory: Applications to Laverbread Consumers and the Windscale Liquid Effluent Discharges', Health Physics, Vol. 29, pp 171 - 179.

Changery, M.J., 1982. 'Historical Extreme Winds for the United States - Atlantic and Gulf of Mexico Coastlines', U.S. Nuclear Regulatory Commission, NUREG/CR-2639.

Galambos,J., 1978, The Asymptotic Theory of Extreme Order Statistics, John Wiley and Sons.

Gumbel, E. J., 1958, Statistics of Extremes, Columbia University Press.

Johnson, N.L., and Kotz, S., 1970, Continuous Univariate Distributions - 1, Houghton Mifflin.

Maritz, J.S., and Munro, A.H., 1967, 'On the Use of the Generalized Extreme-Value Distribution in Estimating Extreme Percentiles', Biometrics, Vol. 23, pp 79-103.

O'Neill, R., 1971, 'Function Minimization Using a Simplex Procedure', Applied Statistics, Vol. 20, pp 338 - 345.

Additional Reading

Singpurwalla, N. D., 1972, 'Extreme Values from a Lognormal Law With Applications to Air Pollution Problems', Technometrics, Vol. 14, No. 3, pp 703 - 711.

Epstein, B., 1960, 'Elements of the Theory of Extreme Values', Technometrics, Vol. 2, No. 1, pp 27 - 41.

7.8 EXERCISES

The following data sets are taken from Changery (1982). The first exercise is worked, only the data is presented for the remainder. For each data set, estimate the location and scale parameters, make a probability plot, and calculate the 2, 5, 10, 20, 50, 100, 200, and 500 year return period wind speeds.

7.8.1 Maximum Annual Wind Speeds for New London, Connecticut.

YEAR	MAX WIND(MPH)	YEAR	MAX WIND(MPH)
1873	70	1885	47
1874	41	1886	47
1875	48	1887	60
1876	59	1888	46
1877	54	1889	51
1878	59	1890	60
1879	42	1891	51
1880	42	1892	38
1881	50	1893	54
1882	42	1894	43
1883	45	1895	44
1884	53		

Mean = 49.826, St.Dev = 7.89, N = 23

Moment estimates of parameters

$\hat{a}$ = SD*SQRT(6)/pi = 7.89*0.7797 = 6.1513

$\hat{u}$ = mean - a*E = 49.826 - 6.1513*0.57722 = 46.2755

Regression estimates of parameters

The independent regression variable is

Y = -ln(-ln(m/(N+1)))

where m is the rank of the wind speed

THE REGRESSION EQUATION IS

Y = 46.1 + 7.14*Max Wind

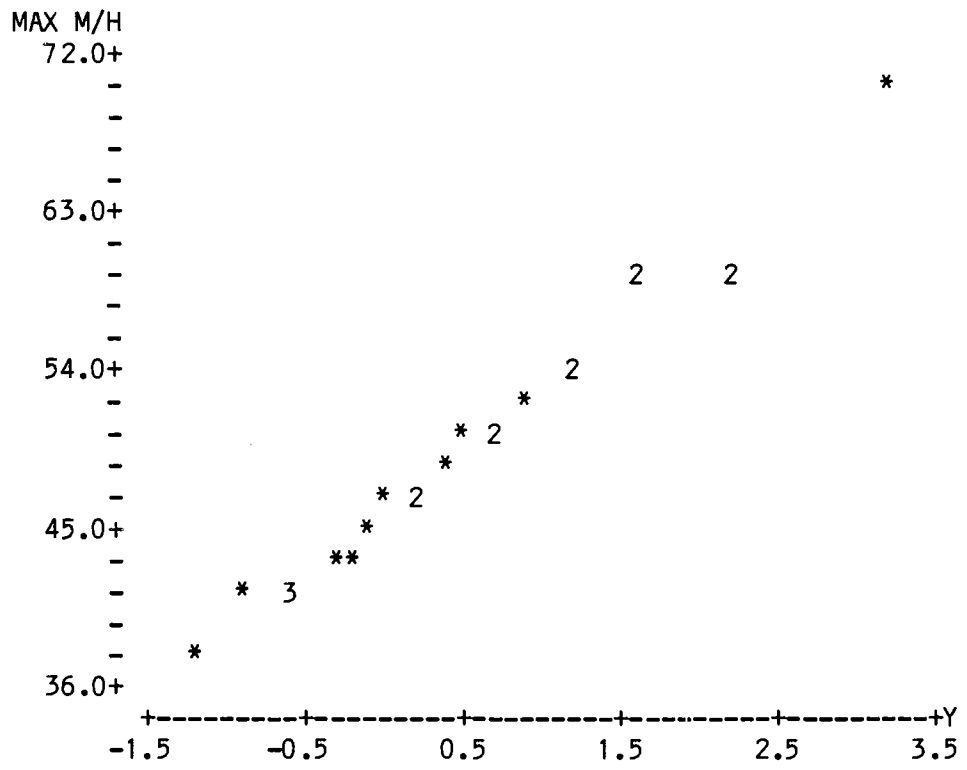
PARAMETER	ESTIMATE	ST. DEV. OF ESTM.	T-RATIO = ESTM/S.D.
u	46.0685	0.2213	208.08
a	7.1401	0.1852	38.54

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.9533 WITH (23-2) = 21 DEGREES OF FREEDOM

R-SQUARED = 98.6 PERCENT

ANALYSIS OF VARIANCE

DUE TO	DF	SS	MS=SS/DF
REGRESSION	1	1350.219	1350.219
RESIDUAL	21	19.084	0.908
TOTAL	22	1369.304	



Probability plot

Return Period

Return period = T = 1/(1 - Prob)
then Prob = 1 - 1/T
Y = -ln(-ln(Prob))
X = u + aY = wind speed predicted from return period.
M = X calculated from moment estimates of u and a,
R = x calculated from regression estimates of u and a.

Return period	Prob	Y	M	R
2	.500	.3665	49	49
5	.800	1.500	56	57
10	.900	2.250	60	62
20	.950	2.970	65	67
50	.980	3.902	70	74
100	.990	4.600	75	79
200	.995	5.296	79	84
500	.998	6.214	84	90

7.8.2 Maximum Annual Wind Speeds for New Haven, Connecticut.

YEAR	MAX WIND(MPH)	YEAR	MAX WIND(MPH)
1944	38	1956	33
1945	37	1957	37
1946	39	1958	34
1947	41	1959	49
1948	25	1960	42
1949	33	1961	45
1950	55	1962	45
1951	40	1963	51
1952	37	1964	44
1953	35	1965	43
1954	45	1966	49
1955	42	1967	45
		1968	44

7.8.3 Maximum Annual Wind Speeds for Apalachicola, Florida

YEAR	MAX WIND(MPH)	YEAR	MAX WIND(MPH)
1975	32	1978	32
1976	26	1979	31
1977	30		

7.8.4 Maximum Annual Wind Speeds for Fort Myers, Florida

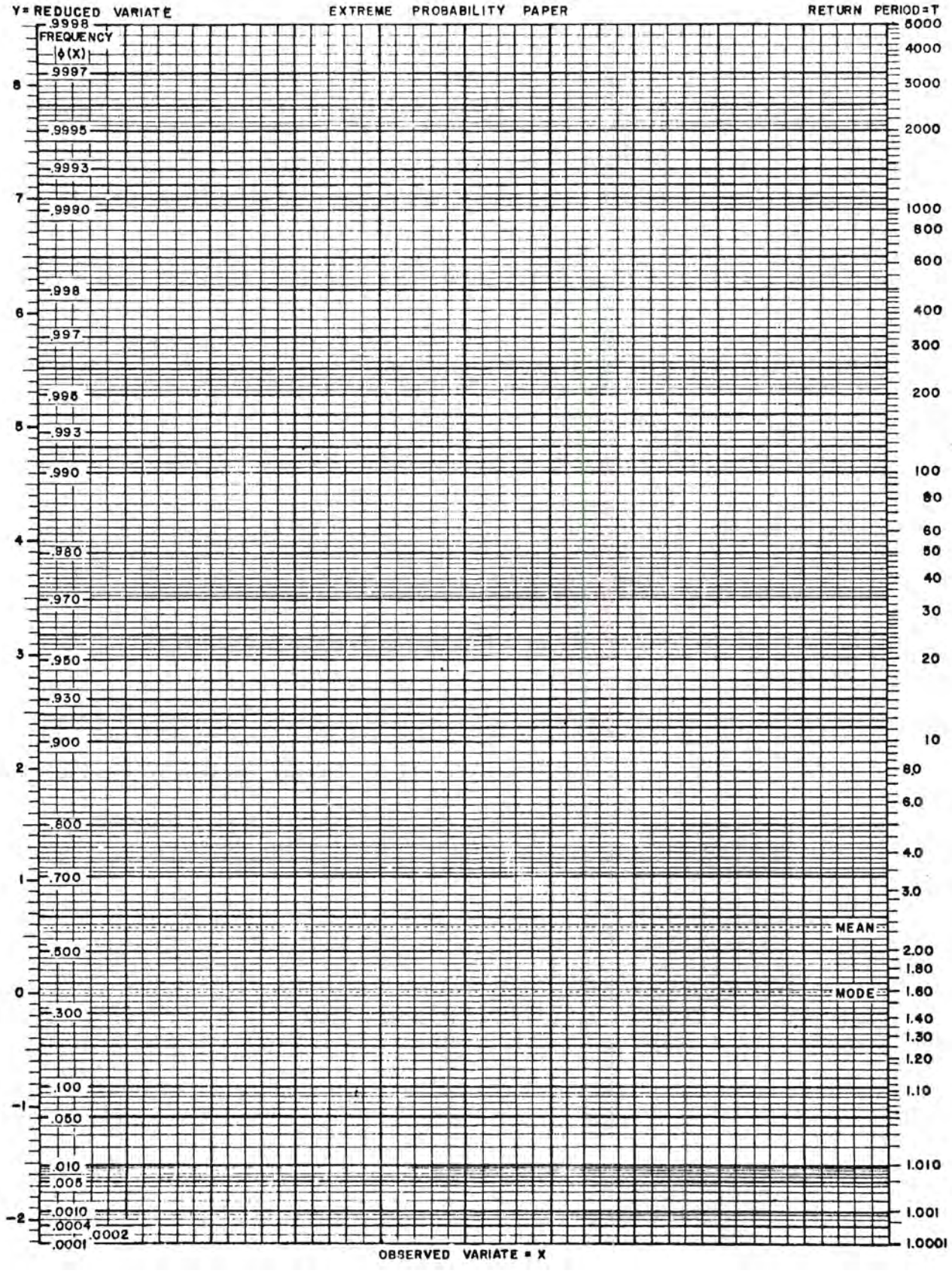
YEAR	MAX WIND(MPH)	YEAR	MAX WIND(MPH)
1920	40	1927	39
1921	48	1928	64
1922	36	1929	61
1923	33	1930	37
1924	57	1931	39
1925	40	1932	47
1926	65		

7.8.5 Maximum Annual Wind Speeds for Hartford, Connecticut

YEAR	MAX WIND(MPH)	YEAR	MAX WIND(MPH)
1940	34	1960	47
1941	43	1961	43
1942	39	1962	43
1943	43	1963	45
1944	59	1964	55
1945	43	1965	42
1946	50	1966	39
1947	47	1967	58
1948	39	1968	44
1949	42	1969	40
1950	67	1970	46
1951	37	1971	51
1952	54	1972	54
1953	48	1973	37
1954	48	1974	46
1955	43	1975	40
1956	43	1976	46
1957	39	1977	43
1958	43	1978	54
1959	42	1979	70

7.9 APPENDIX 7-A

A sheet of Extreme Value Probability Paper


Y = REDUCED VARIATE
EXTREME PROBABILITY PAPER
RETURN PERIOD = T
FREQUENCY
φ(X)
.9998
.9997
.9995
.9993
.9990
.998
.997
.995
.993
.990
.980
.970
.950
.930
.900
.800
.700
.500
.300
.100
.050
.010
.005
.0010
.0004
.0002
.0001
8
7
6
5
4
3
2
1
0
-1
-2
6000
4000
3000
2000
1000
800
600
400
300
200
100
80
60
50
40
30
20
10
8.0
6.0
4.0
3.0
2.00
1.80
1.60
1.40
1.30
1.20
1.10
1.010
1.001
1.0001
MEAN
MODE
OBSERVED VARIATE = X

7.10 APPENDIX 7-B

Goodness of Fit Test for the Extreme Value Distribution

Whenever data are analyzed the statistical methods are based upon several underlying assumptions that are rarely stated except in elementary textbooks. Testing the validity of such assumptions is rarely mentioned, although such tests are an essential part of good statistical analysis. That the randomness in the data may be described by some specified statistical distribution function is an assumption common to all parametric statistical methods. Goodness of fit tests are used to test this type of assumption. The Chi Square test is the best known of these tests, however, it should not be used with small sample sizes because it is sensitive to the way in which the data are divided into groups. The Komogorov-Smirnov test is very powerful for all sample sizes. However it requires known or hypothesized values rather than estimates for the distribution parameters. The Shapiro-Wilk test is most universally applicable. However, it suffers somewhat from not yet being available in elementary textbooks, being rather tedious to compute, and having references that are difficult to obtain since they are now over 10 years old. Confidence bands for the extreme value distribution, which serve a similar statistical function as goodness of fit tests, are described by Cheng and Iles (1983).

Another goodness-of-fit test, asymptotically equivalent to the Shapiro-Wilk test, has been independently proposed by Filliben (1975) and by Ryan et. al. (1980). This test is the Correlation Coefficient Goodness of Fit Test. Its main advantage is that it is easy to compute, and its main disadvantage is that it requires special probability tables. Such tables have only been published for testing goodness of fit to the Gaussian (Normal) distribution. Table 7.B.1 is a new probability table for testing goodness of fit to an exponential type Extreme Value distribution. The papers by Filliben and by Ryan should be consulted for the theoretical

background of such correlation tests and their relationship to other goodness of fit tests.

The correlation coefficient goodness-of-fit test is performed by computing the Pearson product-moment correlation coefficient between data values and the corresponding expected values of a reduced variate, computed from equations 7.5 and 7.6; these expected values are called 'scores'. The hypothesis of a good fit is then evaluated by comparing the computed correlation with an appropriate table of critical values. Some may object to calling the computed value a correlation coefficient because the scores are not random variables. The word correlation in this test is used to describe the computational procedure, not the statistical characteristics of the numbers used in the computation. It must be emphasized that none of the test statistics applicable to true correlation coefficients, such as a test for no correlation, are applicable to these goodness-of-fit correlations.

Table 7.B.1 was derived from Monte Carlo simulations using a standard

Cheng, R. C. H., and Iles, T. C., 1983, 'Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables', Technometrics, Vol. 25, No. 1, pp 77 - 86.

Table 7.B.1
Approximate Critical Values, Correlation Coefficient
Goodness of Fit Test to an Exponential Type
Extreme Value Distribution

	Lower Tail Area		
n	.01	.05	.10
5	.815	.872	.898
10	.854	.904	.925
15	.874	.921	.939

is less than a one percent chance that the data have an Extreme Value distribution. Also for a sample size of 5, a correlation of 0.840 means that there is between a one percent and a five percent chance that the data have an Extreme Value distribution.

Details of Performing the Test

The n extreme data values of a sample are ranked from 1 to n (smallest to largest). Let i denote the rank associated with data point x(i). Its corresponding score is the corresponding standard reduced Extreme Value distribution value:

$$y(i) = -\ln(-\ln(i/(n+1))) .$$

A plot of x(i) versus y(i) is the Extreme Value probability plot discussed in section 7.2. The Pearson product-moment correlation coefficient between x(i) and y(i), i = 1,2,...n, is computed. The goodness-of-fit to an exponential type Extreme value distribution is evaluated by comparing this computed correlation to the critical values given in Table 7.B.1.

References to Appendix B

Filliben, J. J., 1975, 'The Probability Plot Correlation Coefficient Test for Normality', Technometrics, Vol. 17, No. 1, pp 111 - 117.

Ryan, T. A., Joiner B. L., and Ryan, B. L., 1980, Minitab Reference Manual, Statistics Dept., Pennsylvania State U., (section 11.7).

Cheng, R. C. H., and Iles, T. C., 1983, 'Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables', Technometrics, Vol. 25, No. 1, pp 77 - 86.

Table 7.B.1
Approximate Critical Values, Correlation Coefficient
Goodness of Fit Test to an Exponential Type
Extreme Value Distribution

	Lower Tail Area		
n	.01	.05	.10
5	.815	.872	.898
10	.854	.904	.925
15	.874	.921	.939
20	.888	.931	.948
25	.898	.939	.954
30	.906	.946	.959
40	.918	.953	.965
50	.927	.959	.970
60	.933	.963	.973
70	.939	.967	.976
80	.943	.969	.978
90	.947	.972	.980
100	.951	.974	.981
200	.970	.983	.988

7.11 APPENDIX 7-C

Extremes and Population Inference

The reason for not making inferences about a parent population from extreme values is illustrated in the following example from Galambos (1978, page 90). The level of mathematics required to follow the computations presented here is higher than that required for the regular chapter material, however the conclusions are easy to appreciate.

If X is a lognormally distributed random variable (that is, ln(x) has a standard Gaussian distribution), then the distribution of the function

$$(7.21) \qquad y = \frac{x^e - 1}{e}$$

converges to a standard Gaussian distribution as e approaches zero. For small values of e it will, for practical purposes, be impossible to distinguish between samples from the distribution of Y and from a standard Gaussian distribution.

Suppose an experimenter collects a sample of size 50 from the disrtibution of Y with e=0.1. A goodness of fit test is performed and accepts the hypothesis that the parent distribution is Gaussian. Find the probability that the maximum of the sample has a value less than 2.6.

When it is (incorrectly) assumed that the parent distribution is Gaussian, Equation 7.20 gives the probability for any given value of an arbitrary value z. For n=50, a(50)=2.797, and u(20=2.101, then

$$P(2.797*(w - 2.101) \leq z) = \exp(-\exp(-z)) .$$

This may be converted algebraically to

(7.22) $P(x \leq 0.3575*z + 2.101) = \exp(-\exp(-z))$.

Since the problem calls for $P(w \leq 2.6)$, solve for z in

$$2.6 = 0.3575*z + 2.101$$

which gives z=1.396. The desired probability is then $\exp(-\exp(-1.396)) = 0.78$. In words, if the data are really from a standard Gaussian distribution, then there is a 78 percent chance that the largest of a sample of 50 values will be less than 2.6.

But the data is not from a Gaussian distribution, rather it is from a distribution that is indistinguishable from a Gaussian. The correct probability is obtained by starting with the normalizing formula for the lognormal distribution and working through the transform Equation 7.21. Let u'=exp(u(n)) and a'=u'/a(n). Then the equation equivalent to 7.22 is

$$P(w \leq u' + a'*z) = \exp(-\exp(-z)) .$$

For n=50, u'=8.174 and a'=2.922.Let w' be the lognormal equivalent of the w used above for the Gaussian distribution. The value of w' is obtained by replacing y in Equation 7.21 with w and x with w'. Since e=0.1,

$$w = 10(w')^{.1} - 10 .$$

An equivalent to 7.22 is

$$P(w \leq 10(w')^{.1} - 10) = \exp(-\exp(-z)) .$$

Since $P(w \leq 2.6)$ is required, solve for z in

$$2.6 = 10(8.174 + 2.922*z)^{.1} - 10 ,$$

which gives z=0.6544. Then $\exp(-\exp(-0.6544))=0.59$. Thus, assuming a

Gaussian distribution rather than a look alike transformed lognormal distribution, causes an error of about 0.2 (0.78 - 0.59) in probability when computing the probability that the largest value of a sample of size 50 will be less than 2.6.

This example shows how sensitive the extremes are to subtle changes in the parent distribution. The converse is that inferences about the parent population will also be very sensitive to subtle differences in the shape of the tails of the distribution.

CHAPTER 8
EXTREMES OF DATA CONTAINING TRENDS

8.1 INTRODUCTION

In Chapter 7 the 'classic' extreme value situation was presented, in which all samples are identically and independently distributed. Such a situation rarely describes any real data set. One must deal with lack of independence between observations, measurement errors, and many other practical aspects of data collection and analysis. This chapter considers some methods for handling the most common cause of lack of independence: correlations caused by time trends in the data.

The Extreme Value distribution is robust against correlations within a data set. Berman (1964) has shown that extreme value theory can be applied to stationary autocorrelated Gaussian sequences provided that:

$$\sum_{i=1}^{\infty} r(i)^2 < \infty \quad ,$$

where r(i) is the autocorrelation of order i. Autocorrelations can be computed and examined to see if they satisfy Berman's condition. For example, suppose some autocorrelations are examined and

$$r(i)^2 \leq 1/2^i \quad .$$

Since

$$\sum_{i=1}^{\infty} 1/2^i = 1.0 \ ,$$

these autocorrelations satisfy Berman's condition and the data may analyzed relying upon the robustness of extreme value procedures. In general, if a plot of the autocorrelation function indicates significant autocorrelation

for only a finite number of lags, then this condition is satisfied. Berman's result applies only to the form of the limiting distribution (Extreme Value distribution from a Gaussian sequence). The possibility that autocorrelations might result in biased parameter estimates is not addressed in his work. Berman's condition is in general not satisfied if there are periodicities in the data (sine or cosine functions describing the trend) since there would be infinite values of i for which of r(i) is significantly different from zero.

If the autocorrelations of all orders are high, the variance of an average (square of the standard error) can actually increase as sample size increases. If V is the varianve of a single reading, then the variance of the average of n readings is

$$V[n] = \frac{V}{n}(1 + (n-1)*\bar{r}*n)$$

where $\bar{r}$ is the average autocorre ation between the n data values.

When the average autocorrelation is zero this equation simplifies to the familiar equation for standard error. When the average autocorrelation is 1/n the variance of an average of n samples is the same as the varianve of a single sample. For average autocorrelations larger than 1/n the variance of the mean increases with increasing n. If Berman's condition holds, the average autocorrelation is zero since the sum of an infinite number of autocorrelations squarred must be a constant less than infinity. Gardenier (1982) shows that the expected number of exceedances is also significantly increased by autocorrelations, thus trends must also be considered when using the statistics presented in Chapters 4, 5, and 6.

As a general rule of statistical analysis, one should remove all the correlations that can be found. The tools of ordinary statistics for removing correlations in data can also be used with extreme value data. The most common of these tools is data transformation. When multivariate

data is being analyzed, rotations of axes are often used to gain independence. Principal Components and Factor Analysis are typically used for such rotations. These procedures have the added advantage of reducing the dimension of the multivariate problem. Often, only one Component or Factor is used in order to reduce the multivariate problem to a univariate problem. A later chapter will discuss multivariate extreme values.

Clearly, the maximum of a stationary continuous random process is at least as large as the maximum of any sampled values. It is important to know if these two maxima can be significantly different in magnitude, or if the sample data has a distribution of extremes different from that of the continuous maximum. Leadbetter (1977) found that, under very weak regularity conditions (which are typically satisfied in any practical data analysis), the distribution of extremes of a continuous process and sample extremes from such a process obey the same extreme value law and may be treated as independent samples.

Intuitively a finely spaced sampling scheme upon a continuous stationary random process should guarantee approximate equality of the continuous and sample maxima. If samples are independent rather than from a stationary process, then there are many (however small) intervals between samples in which values above the sampled maxima are possible. Even though such high values in any one interval are very unlikely, the large number of such intervals can lead to significant differences between the extremes of the samples and the phenomena being sampled. On the other hand, if a continuous stationary random process is being sampled, it follows that for sufficiently fine sampling intervals, there will be little difference between the maximum of the samples and the maximum of the phenomena because the sample values must be locally highly correlated. The underlying random process may be envisioned as a high frequency filter which removes most of the variability between sampling times. Air pollutant levels are filtered by diffusion and mixing time processes. The robustness of the extreme value distribution is due to the local correlations caused by these

continuous stationary random processes. When a trend is introduced, correlations also occur because the underlying phenomena becomes a nonstationary random process. This chapter discusses some of the special problems that occur when a nonstationary process is decomposed into a stationary process and a trend before the data is analyzed for extreme values.

8.2 REMOVING TRENDS BEFORE ANALYSIS

The first step in a data analysis is to examine the data for the existance of potential trends. This is usually done using graphic techniques. Typically a smoothing algorithm is used to mask the visual effects of randomness in the data. If no trends are apparent, then the techniques of Chapter 7 are immediately applicable. The most direct way of handling an obvious trend is to subtract it out of the data and then use the techniques of Chapter 7 on the residuals. This is equivalent to an extension of the extreme value distribution so that the mode is a function of sampling time. The reduced variate of formula 7.3 would then be written $y = (x(t) - u(t))/a$ indicating that the observed values x and the location parameter u are both functions of sampling time.

Often the mode u is assumed to be a polynomial function over time such as

$$u(t) = a + bt + ct^2 \quad . \tag{8.1}$$

Environmental work is often concerned with repetitive yearly cycles which are typically modeled with a harmonic (trigonometric) series,

$$u(t) = a + b*\sin(ct - d) + e*\sin(2ct - d) + \ldots . \tag{8.2}$$

For short intervals of a cyclic trend, such as daily maximum one hour average ozone concentrations over a single year, most authors prefer a quadratic polynomial rather than a harmonic function. One could combine

equations 8.1 and 8.2 to get a mixture of the two kinds of trends.

To accomplish removal of trend, standard regression procedures are used on the data before the analysis for extreme value parameters. The trend is then subtracted out of the data. Finally, the extreme value analysis is done. For example, suppose 50 years of flood (peak flow) data are available from a local stream. The flow is suspected to have been decreasing because of diversion of water for urban use. The first step in this data analysis is to plot the yearly floods versus year on linear-linear graph paper. On such a plot, the data may seem to show a linear decline. Then a straight line would be fit to the data and this line gives estimates of u(t).

Such a linear regression model cannot be used for u(t) because u(t) is the mode of the extreme data and least squares procedures estimate the mean. However, a linear (mean) trend can be subtracted from the data to eliminate the correlations induced by the trend. Then the extreme value analysis described in Chapter 7 can be performed. The resulting value computed for u will be the difference between the mode and the mean and is constant over time if the trend has been removed. Section 7.4.4 shows that the magnitude of the difference between the mean and mode should equal Euler's constant times the scale parameter (except for the effects of random error).

Trends based upon measures other than time should be considered. Suppose a plot of the flood data shows some rather sharp declines, with level intervals between the declines. This would be difficult to fit to a polynomial. However the sharp declines might correspond to the startup times of new industries in the area. A plot of floods versus population size would then show a linear trend, and population size would be a more meaningful measure of trend.

Another methodology for describing trend that is becoming increasingly more popular is to use an 'autoregressive process'. In its simplest form, autoregression describes the current data value as a function of the previous value(s) plus an independent random error:

(8.3) $$x(t) = a*x(t-1) + \text{error} \quad .$$

The reader is referred to the many texts on autoregression and moving averages for a detailed explanation (e.g. Box and Jenkins (1976), and Nelson (1973)).

Autoregressive theory yields algorithms for estimating the parameter a, and the magnitude of the error in equation 8.3. Of course, much more complicated forms than 8.3 are typically used for real data analyses. This theory also yields 'filters' for removing the trend from an autoregressive process.

A simple method for removing trends in extreme value data is to select out that portion of the data in which the extremes are expected to occur. For example, if annual high temperatures are of interest, one would collect data only during summer months to remove seasonal variability from the data. The 'cost' of this method is a smaller sample size from which to choose the extreme. There is also some chance that the true annual maxima will occur outside the chosen sampling period. The advantage of this method is its simplicity, it requires no mathematical description of the trend, and untrained persons can perform such an analysis.

8.3 INCLUDING TREND IN THE DATA ANALYSIS

When trend is removed from the data a subtle problem is created: the observed maxima are then the extreme deviations from the trend rather than the extremes over time. This may or may not be the variable of interest. For example, weekly maxima of one hour ozone concentrations are usually

assumed to have a lognormal distribution and also have a harmonic trend peaking in the late summer (Larsen, 1969). To remove trend from such data, one would regress a periodic function (sines and cosines of multiples of the time variable) on the logarithms of the data. When this periodic function is subtracted from the data a sequence of identically distributed extreme values result. The maxima of these detrended sample values might well occur in the middle of the winter, even though the annual maxima of ozone concentrationsoccurs in the summer.

The detrended sample values are studied to test the adequacy of the underlying assumptions of the statistical procedures, but the numbers of practical significance usually include the trend. The rest of this section presents a technique due to Horowitz (1980) which simultaneously treats the trend and the maximum values.

Let $x(t)$, $t=1,2,...n$, be a sequence of samples from a continuous random process of the form

(8.4) $x(t) = f(t) + e(t)$

where:

1) $f(t)$ is a bounded deterministic function
2) the sequence $e(t)$ is a Gaussian stationary process satisfying:
   a) $E(e(t)) = 0$ for all t
   b) $E(e(t)*e(t)) = v$ = variance, a constant
   c) $E(e(t)*e(t+k))/v = r(k)$ for all t and $k \geq 1$
   d) $\sum_{k=1}^{\infty} r(k)^2 < \infty$.

$E(\ )$ is the statistical expectation function. In equation 8.4 the $x(t)$ can either be the data or a transformation of the data, such as the logarithm of air pollutant concentrations mentioned previously. The assumption that the $e(t)$ have a Gaussian distribution may seem to defeat the purposes since most of the data of interest are extremes from a Gaussian distribution rather than data from a Gaussian. Typicall variables of interest are such

things as how annual floods change over the years, how weekly maxima of one hour ozone averages change over the year, and so on. River depths and ozone averages (or their logarithms) have approximately a Gaussian distribution, thus their maxima have an Extreme Value distribution. Leadbetter (1977, theorem 4.1) shows that an Extreme Value result that holds for a parent distribution also holds for sample extremes from that distribution. Leadbetter's theorem is an important addition to Horowitz's work since it allows the same statistical analysis techniques to be applied to samples and to extremes of samples.

Define z(n) to be the maximum of n observations,

(8.5) $$z(n) = \max(x(1),x(2),...x(n)) \quad .$$

Then Horowitz shows that an asymptotic approximation for the probability distribution of z(n) that is valid for large n 's given by:

(8.6) $$P(z(n) \leq Z) = \exp(-\exp(-(Z-b(n))/a(n))) \quad ,$$

where:

(8.7)
$$d(n) = \mathrm{SQRT}(2*\ln(n))$$
$$s = \mathrm{SQRT}(v)$$
$$a(n) = s*b(n)/d(n)$$
$$b(n) = s*c(n) + g(n) \quad @$$
$$c(n) = d(n) - (\ln(\ln(n))+\ln(4*\pi))/(2*d(n))$$
$$g(n) = h(n)/d(n)$$
$$h(n) = -\ln(n)+\ln\left(\sum_{t=1}^{n}\exp(d(n)*f(t)/s)\right) \quad .$$

@ The formula for b(n) given by Horowitz (1980) includes an exponentiation which is omitted here. Horowitz restricts his derivation to the distribution of the natural logarithm of the data rather than the data values. Study of similar work by Leadbetter (1977, section 11), Epstein (1960, pp 39 - 40), and Singpurwalla (1972, Appendix) indicates that the formula for b(n) can be used without exponentiation to analize data.

Some algebra shows that equations 8.7 are identical to equations 7.19 if f(t) is a constant. The key difference between the independent identically distributed samples of equations 7.19 and the samples from a nonstationary random process of equations 8.7 is the term h(n), which itself includes a term that is a summation over the expected values of the data, f(t). The expected value of z(n) is b(n) + E*a(n), where E is Euler's constant.

8.3.1 Example, Trend in Annual Floods.

Twenty years of flood data were simulated by creating a sample from an Extreme Value distribution, applying an autoregressive process to the sample, and finally adding a linear trend. This data is plotted in Figure 8.1 and listed in Table 8.1.

Table 8.1
Annual Floods versus Years

TIME	DATA	TIME	DATA
1	26.3717	11	22.0139
2	20.8157	12	20.1129
3	20.4614	13	20.3941
4	20.7928	14	18.8047
5	20.4814	15	19.5968
6	27.9141	16	18.5526
7	26.1495	17	21.0049
8	19.3770	18	20.6239
9	23.3212	19	17.3935
10	19.5338	20	16.0694

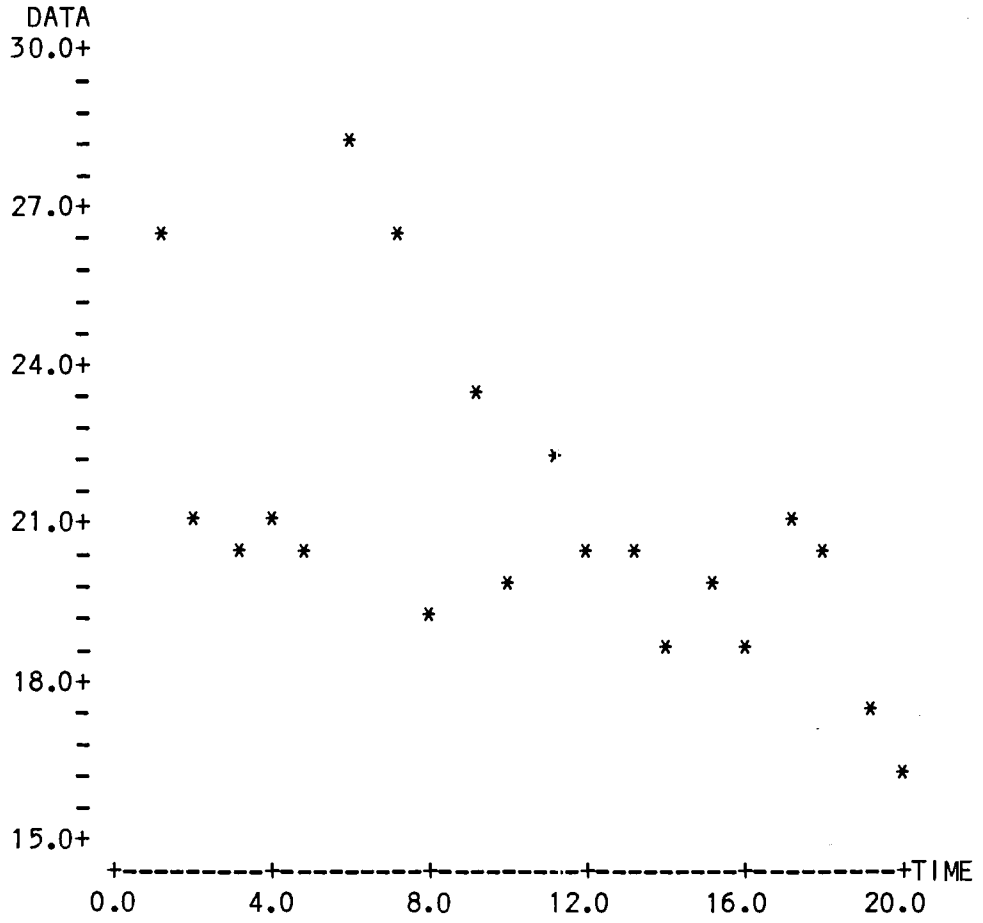


Figure 8.1
Annual Floods versus Years

A sample of size 20 is much smaller than any real data set, and too small to really test for significance of statistical trend processes within the data. Such a small sample is used here to allow the reader to easily repeat the computations.

Figure 8.1 suggests that a linear, and possibly a quadratic, function might describe the trend. These regressions were done, and compared to each other and also to a no trend model (zero slope) using the General Linear Hypothesis to test for significant differences between trend models.

These tests showed that a linear trend is significantly better than the no trend model (confidence = 99.5%) and that a quadratic model is not significantly better than the linear (confidence = 50%). The estimated linear trend model is:

$$DATA = 24.13 - 0.299*TIME$$

The R-squared value for this regression is 35.8%.

A time series analysis of the residuals (MINITAB ARIMA command) indicated no significant autoregressive pattern. Thus, one can conclude that the data show a linear trend with independent random errors.

Next the residuals from the regression are analyzed to determine if they can be described by an Extreme Value distribution. Figure 8.2 is an Extreme Value probability plot of these residuals. This type of plot was discussed in section 7.3. The residuals appear to be close to a straight line.

Applying the correlation test for goodness-of-fit discussed in Appendix 7-B yields a correlation of 0.990. This value falls close to the 90th percentile of the distribution of correlations for sample size 20, indicating a good fit. A correlation test for goodness of fit was also done for the Gaussian distribution (Ryan et. al., 1980, section 11.7), giving a correlation of 0.963 which falls at about the 10th percentile of the corresponding distribution. Thus one may conclude that the residuals from the linear trend model can be described by either an Extreme Value distribution or a Gaussian distribution. The Extreme Value distribution is a slightly better fit.

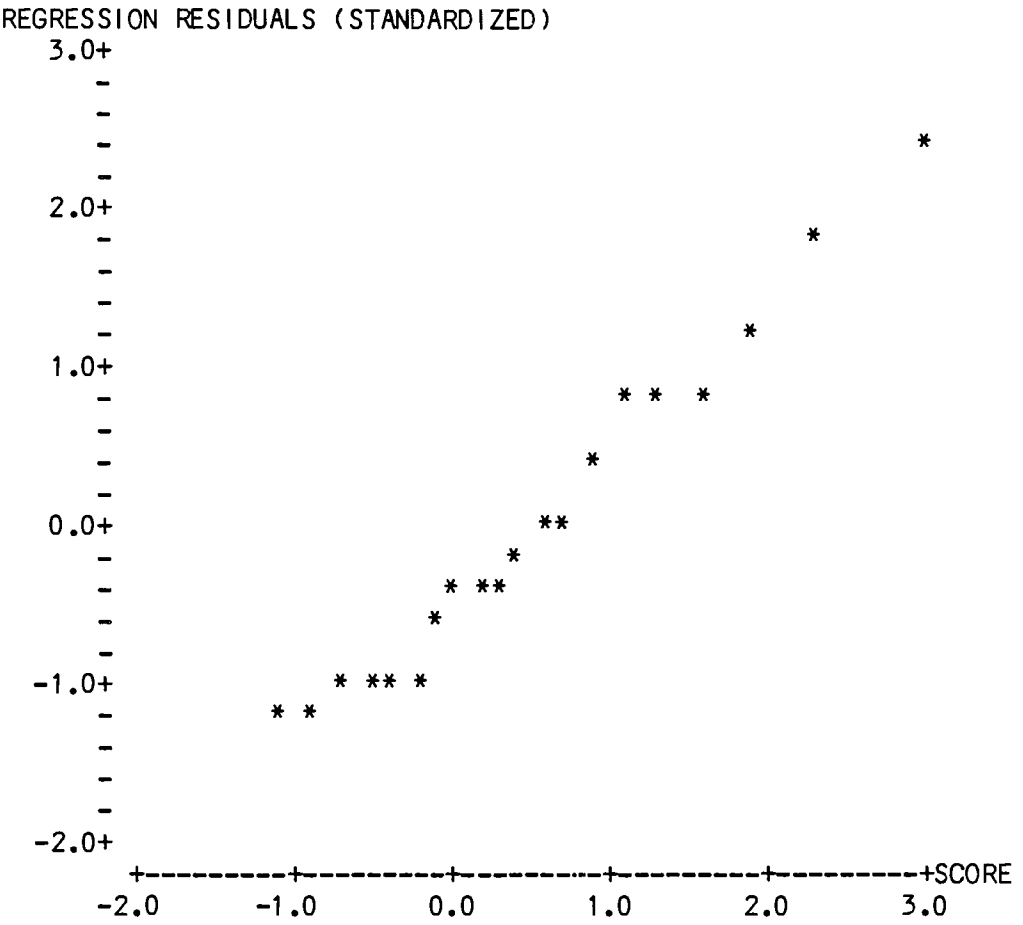


Figure 8.2
Extreme Value Probability Plot of
Linear Regression Residuals

If the trend information were ignored when computing the parameters of the Extreme Value distribution of floods equations 7.19 would be used, with the overall mean and variance of flood data, to compute the distribution of the maximum over 20 years. The description of the flood data is:

sample size = n = 20
mean = 20.989
standard deviation = 2.96

Applying equations 7.19 yields:

$c(n) = 1.707$
$u(n) = 26.02465$
$a(n) = 31.471$

These results show that, for the flood data, the mode of the extremes of 20 observations is just over 26, which is reasonably close to the observed maximum of 28. The value of the scale factor a(n) relative to u(n) suggest that there is a great deal of uncertainty in any estimate derived from this data. Applying equations 7.15 and 7.16 yields a mean of 44.19 and a standard deviation of 40.36, which along with u(n) and a(n) shows a long-tailed distribution over the high values. The mean (44.19) is approximately twice the mode (18.17, computed from Equation 7.15).

When the trend is included in the analysis, equations 8.7 are used rather than 7.19. Separating the computations into parts gives:

$n = 20$
$d(n) = SQRT(2*ln(20)) = 2.44775$
$f(t)$ = expected value at time t from the linear regression
$= 24.13 - 0.2995*t$
s = square root of linear regression mean-square-error
$= 2.436$

$$\sum_{t=1}^{n} \exp(d(n)*f(t)/s) = 7.8566E10$$

$h(n) = -ln(20) + ln(7.8566E10) = 22.091473$
$g(n) = h(n)/d(n) = 9.0252$
$c(n) = d(n) - 3.6282129/(2*d(n)) = 1.70655$
$u(n) = 2.436(c(n) + g(n)) = 26.1425$
$a(n) = 26.017$.

The mean and standard deviation computed from equations 7.15 and 7.16 after adjustment for the trend are 45.8 and 33.4, respectively. Thus, in this example, adjusting for the trend gives essentially the same mode and mean as without adjustment, but using the trend reduces the scale factor and standard deviation. With large sample sizes one would expect

(Horowitz, 1980) that excluding trend information would cause a substantial overestimation of the magnitude of extremes. This occurs because without the trend removed the estimate of e(t) is inflated by the effect of the trend. In analysis of variance terminology, this is combining the between time period effects with the within time periods effects.

The reduced spread produced by using trend information will result in a substantially smaller upper confidence limit on the expected extreme values. The relations given in section 7.4.4 are used to calculate the upper 95% confidence limit for these floods. Accounting for the trend gives a limit of 103, while ignoring it gives a confidence limit of 119. If a flood control project were being designed, the upper 95% confidence limit is a reasonable design criteria. A 16 foot difference in flood level is of substantial economic importance when engineering for flood control.

8.4 OTHER CONSIDERATIONS

Equations 8.4 to 8.7 offer an efficient way to obtain estimated maxima over long duration cyclic trends. Suppose one wishes to estimate the distribution of maximum annual pollutant concentrations, assuming there is no trend other than the annual seasonal cycles. One way to find such a distribution would be to collect data for many years, select out the maxima from each years data, and apply the techniques given in Chapter 7. Equations 8.4 to 8.7 can achieve the same goal with only one years data.

Another interesting application of these formulas is to explore the effects of various hypothetical trends upon a (known) stationary process. Environmentalists often predict the changes that would be expected if a new industry came into a region, or if a cleanup strategy is implemented. These predictions are usually estimated changes in average values. Such estimated means could be added to f(t) in equations 8.4 to 8.7 in order to predict the accompanying changes in the extremes.

8.5 SUMMARY

The commonly used procedure for estimating the Extreme Value distribution of a sequence of measurements implicitly assumes the samples are from a stationary random process. Ignoring trends results in an overestimate the magnitude of the extremes and their associated statistics. This chapter presents procedures for including such trends within the parameter estimation algorithm.

8.6 REFERENCES

Berman, S. N., 1964, 'Limit Theorems for the Maximum Term in Stationary Sequences', Ann. Math. Stat., Vol. 35, pp 502 - 516.

Box, G. E. P., and Jenkins, G. M., 1976, Time Series Analysis, (Revised) Holden-Day, San Francisco.

Epstein, B., 1960, 'Elements of the Theory of Extreme Values', Technometrics, Vol. 2, No. 1, pp 27 - 41.

Gardenier, T. K., 1982, 'Moving Averages for Environmental Standards', Simulation, Vol. 39, No. 2, pp 49-58.

Horowitz, J., 1980, 'Extreme Values from a Nonstationary Stochastic Process', Technometrics, Vol. 22, No. 4, pp 469 - 482.

Larsen, R. I., 1969, 'A New Mathematical Model of Air Pollutant Concentration Averaging Time and Frequency', J. Air Pollution Control Association, Vol. 19, pp 24 - 30.

Leadbetter, M. R., 1977, 'On Extreme Values of Sampled and Continuous Stochastic Data', Stanford University Dept. of Statistics Technical Report No. 10, NTIS C00-2874-21.

Nelson, C. R., 1973, Applied Time Series Analysis, Holden-Day, San Francisco.

Ryan, T. A., Joiner, B. L., and Ryan, B. L., 1980, Minitab Reference Manual, Statistics Dept., Pennsylvania State Univ.

Singpurwalla, N. D., 1972, 'Extreme Values from a Lognormal Law With Applications to Air Pollution Problems', Technometrics, Vol. 14, No. 3, pp 703 - 711.

CHAPTER 9
PARAMETER ESTIMATION

9.1 INTRODUCTION

In Chapter 7 three ways of estimating the location and scale parameters of the Extreme Value distribution are given: 1) by linear regression on the cumulative probability plot, 2) by transformation of the extreme's sample mean and variance (method of moments), and 3) by maximum likelihood. The discussion in that chapter concludes that the maximum likelihood estimates are the best, but are difficult to compute. Transformation of the mean and variance is computationally easy and gives values close to the maximum likelihood estimates, but is inefficient for the scale parameter estimate. This method is preferred when simplicity is desired. The regression estimates have good logical and intuitive basis, but are biased. All the above parameter estimating methods require the raw data values and no missing values.

There are many special cases in which additional parameter estimation methods are desirable. There may be so much data that it is not practical to use it all in computations. This could occur if one were analyzing the data on a small computer with limited memory. Data is sometimes collected in such a way that it is censored. In this chapter a method that uses censored samples and a method using information about quantiles of the data are presented. These are arbitrary selections from many estimation methods available in the literature. Along with these selected methods, it will be shown how Gumbel's regression method can be used to solve, in a crude but easily-computed way, the same problems.

9.2 ESTIMATING PARAMETERS FROM LARGEST OBSERVATIONS

There is a wealth of recent literature describing estimation techniques for censored samples. Most of these methods are associated with the Weibull distribution and with life testing of mechanical systems or of consumer products. These methods can be adapted to the Extreme Value distribution through the logarithmic relationship between Extreme Value variables and Weibull variables discussed in Section 3.4. A transform of data values and an inverse transform of parameter values often introduces statistical bias. A prominent feature of most Weibull distribution methods is that they require special tables of coefficients. The method presented in this section is less efficient than the Weibull-type methods, but it requires no tables and is simple enough for hand computation.

Suppose the 10 maximum yearly ozone concentrations from the 50 largest urban areas of the United States are available, and inferences about the yearly maxima over all 50 areas are desired. Formula 7.21 can be used to make inferences about larger sample sizes from a given data set. However, this formula is not applicable to this urban area problem because it assumes the data available is a random subset of all possible data values. For this example, the data reported is not random. The 10 largest values must be analyzed using methods derived from the theory of order statistics.

9.2.1 Regression Estimators.

The simplest method for estimating the location and scale parameters from a censored subset of the data is to use Gumbel's regression technique discussed in Section 7.3, but use only those values of the regressor, $i/(n+1)$, for which data are available. For the urban areas example, let the 10 known values, ranked from largest to smallest, be $x(1),x(2),...x(10)$. These would be plotted and regressed on the predictor variables $y(1),...y(10)$ where the y's are the extreme value scores given in formulas 7.5 and 7.6, $y(i) = -\ln(-\ln(i/(n+1)))$. Here, $n = 50$ and i ranges

from 41 to 50. The difference between this regression and the method presented in Section 7.3 is that in Chapter 7, i ranges from 1 to n. The intercept and slope from the regression are estimates of the location and scale parameters, respectively. (The reason for reversing the usual roles of the x's and y's was discussed in Section 7.2.)

This regression technique can be used for any kind of censoring or any pattern of missing data if the total sample size is known and if the known values can be assigned ranks within the total sample. One chooses the values of i to be the ranks associated with the known data values. Perhaps instead of the 10 highest yearly ozone maximum values of 50 urban areas, the 5 highest and 5 lowest were given. Values of i equal to 1 to 5 and 46 to 50 would be used. This regression method is the only one available that can be generally applied in censored and missing data values situations. One should be especially aware that missing data values can introduce substantial bias into the estimates calculated by regression programs, particularly into the estimated standard deviations of the parameter values.

9.2.2 Minimum Variance Unbiased Estimates.

For the case in which the k largest data values from an extreme value sample are known, Weissman (1978) gives formulas for maximum likelihood estimates and minimum variance unbiased estimates of the location parameter a and scale parameter b. Let x(1),x(2),...x(k) be the ordered (largest to smallest) k largest values from a sample of size n. Let m(k) be the mean of these k values, and define:

$$S(k) = \sum_{j=1}^{k-1} j^{-1} \qquad (9.1)$$

$$V(k) = \frac{pi^2}{6} - \sum_{j=1}^{k-1} j^{-2} \qquad (9.2)$$

where pi = 3.14159...

Maximum likelihood estimates of the Extreme Value distribution location parameter a and scale parameter b are:

(9.3) $\hat{a} = \hat{b}*ln(k) + x(k)$

(9.4) $\hat{b} = m(k) - x(k)$.

Equation 9.3 is identical to equation 7.21, except for a change in the meaning of the letters denoting the statistical quantities. Let E represent Euler's constant (0.57721...). Minimum variance unbiased estimates are:

(9.5) $\tilde{a} = \tilde{b}*(S(k) - E) + x(k)$

(9.6) $\tilde{b} = m(k-1) - x(k)$

and the corresponding variances are:

(9.7) $V(\tilde{a}) = \tilde{b}*((S(k) - E)^2/(k-1) + V(k))$

(9.8) $V(\tilde{b}) = \tilde{b}/(k-1)$.

These equations seem to emphasize the kth data point over all other values. (Points k+1 to n are unknown.) This importance is apparent rather than real, because the actual value used for x(k) is unknown until all the data are collected and ranked. Thus, x(k) is a random variable conditioned upon all other data values.

9.3 ESTIMATING PARAMETERS FROM SAMPLE QUANTILES

Large data sets are often recorded as frequencies that occur within groups or intervals of measurement values. This data summarizing reduces the volume of data, but it doesn't allow application of statistical methods that use raw data values. Such grouping of data can be handled by likelihood maximization methods. Generalized likelihood maximization techniques (algorithms such as Simplex, which is discussed in section 7.3.1) are easy to adapt to grouped data and to the Extreme Value distribution. The likelihood function to be maximized is:

(9.9) $$L = \prod_{i=1}^{k} p(i)^{n(i)}$$

where
k = number of groups of data,
n(i) = number of observations in the ith class, and

$$p(i) = \int_{x(i-1)}^{x(i)} f(z)dz = F(x(i)) - F(x(i-1)) \quad .$$

For the Extreme Value distribution, F(x) is given by equations 7.3; F(x) = H(x) = exp(-exp(-y)) with y = (x - u)/a. The scale parameter a, should be relatively large compared to the length of the intervals, x(i) - x(i-1). This condition is also satisfied if the standard deviation of the x's is large compared to interval length. Computationally, it is convenient to maximize the logarithm of the likelihood function:

$$\ln(L) = \sum_{i=1}^{k} n(i)*\ln(p(i)) \quad .$$

Gumbel's method of regression can also be used with grouped data to estimate the location and scale parameters. The extreme value scores used as the predictor variable are calculated by replacing the term 1/(n+1) with the value of the quantiles. For example, suppose that in a large number of daily maxima of hourly nitrous oxide measurements, 75% of the values are less than 0.1 part per million. The data point used for the Gumbel regression would be 0.1 for the x component and -ln(-ln(.75)) for the y component. The estimates of variances of the parameters produced by most regression programs are invalid in this situation because they do not correctly account for the degrees of freedom. They typically assume each data point represents only one observation. Also, the correlation coefficient goodness of fit test given in Appendix 7-B is not applicable because it is not based on grouped data. Grouping smooths the data and

produces correlations that are biased towards high values. The estimators for the location and scale parameters using Gumbel's regression technique on grouped data are simple enough to calculate on most programmable hand calculators. Also, elementary statistics textbooks give formulas for calculating the mean and standard deviation from grouped data. These, with formulas 7.15a and 7.16a, can be used to estimate the Extreme Value location and scale parameters from grouped data. The efficiency and bias in such methods have not been studied, thus, such methods cannot be recommended unless computational simplicity is essential.

Published techniques for evaluating the Extreme Value distribution with grouped data require special tables of coefficients. The method of Hassanein (1972) is the germ of many subsequent papers on grouped data parameter estimation. The papers published since Hassanein are primarily devoted to the analysis of the Weibull distribution and to eliminating biases. Some of these are discussed in Chapter 12.

Hassanein proposes that the Extreme Value location and scale parameters be estimated by linear combinations of order statistics. He chose the particular order statistics that maximize the relative efficiency of the estimates, and tabulates the coefficients needed to form these linear combinations. The user selects the number of order statistics used in the estimation. Hassanein's results are asymptotic. Mann and Fertig (1977) give bias corrections to Hassanein's equations for small to moderate sample sizes.

Quantile estimators reduce the computational burden by making use of selected observations. Assume a large ordered sample of size N is taken from data following the Extreme Value distribution. Let $x(N,1)$, $x(N,2),\ldots x(N,k)$ be the k sample quantiles to be used in forming estimates of the location parameter u, and the scale parameter a. Let $(N,I) = [N*p(I)+1]$, $I = 1,2,\ldots k$, where $[z]$ denotes the largest integer not exceeding z. Assume the data is ordered from smallest to largest, so that

x(1) is the smallest extreme, x(2) is the second smallest, and so on. (This ordering is opposite from the ordering used for making probability plots of extreme values.) Hassanein gives optimum choices of p(i) and the corresponding coefficients for the linear combinations. He also gives factors to be used in computing variances of the estimators. Tables of these factors and coefficients are given in Appendix 9-A. The estimators are linear combinations of the form:

$$\hat{u} = \sum_{i=1}^{k} c(i) * y[N*p(i)+1] \qquad (9.10)$$

$$\hat{a} = \sum_{i=1}^{k} d(i) * y[N*p(i)+1] \quad . \qquad (9.11)$$

The values of c(i), d(i), and p(i) are tabulated in Appendix 9-A for k = 1 to 7.

Hassanein also gives multipliers for determining the variances and covariances of the location and scale parameters. These are given in Appendix 9-A as E(1), E(2), and E(3), where:

$$V(\hat{u}) = \frac{\hat{a}^2}{N} * E(1) \quad , \qquad (9.12)$$

$$V(\hat{a}) = \frac{\hat{a}^2}{N} * E(2) \quad , \text{ and} \qquad (9.13)$$

$$Cov(\hat{u},\hat{a}) = - \frac{\hat{a}^2}{N} * E(3) \quad . \qquad (9.14)$$

9.3.1 Numerical Example.

Hassanein's equations are used to analyze the maximum radium concentrations, in picocuries per liter, for 485 drinking water wells. The data is summarized in Table 9.1. (The notation [0.2,0.4) means that the data group includes values from 0.2 picocuries up to but not including 0.4 picocuries per liter.)

Table 9.1
Maximum Radium Levels in Drinking Water Wells (pCi/l)

Radium Concentration	Frequency	Cumulative Frequency
[0.2,0.4)	4	4
[0.4,0.6)	11	15
[0.6,0.8)	27	42
[0.8,1.0)	48	90
[1.0,1.2)	62	152
[1.2,1.4)	58	210
[1.4,1.6)	55	265
[1.6,1.8)	60	325
[1.8,2.0)	61	386
[2.0,2.2)	36	422
[2.2,2.4)	17	439
[2.4,2.6)	18	457
[2.6,2.8)	8	465
[2.8,3.0)	7	472
[3.0,3.2)	6	478
[3.2,3.4)	3	481
[3.4,3.6)	1	482
[3.6,3.8)	2	484
[3.8,4.0)	1	485

Suppose the location and scale estimators are calculated using two quantiles; k = 2. From Table 7.B.1 the quantiles that give maximum efficiency are: p(1) = 0.087 and p(2) = 0.734. Next determine which observations best estimate these quantiles using [N*p(i)+1]. N = 485, so

$$[N*p(1)+1] = 43,$$
$$[N*p(2)+1] = 356.$$

Since x(43) and x(356) are not distinct among the groups given in Table 9.1, it is necessary to interpolate to recover their approximate values. x(356) falls at about the middle of the interval between x(325) = 1.8 pCi/l and x(386) = 2.0 pCi/l. This interval has a width of 0.2, an upper bound of 2.0, and a lower bound of 1.8. Using linear interpolation, the fractional distance within the interval of x(356) is

$$(356 - 325)/(386 - 325) = 31/61 = 0.5082 .$$

The fractional distance times interval width plus lower bound is

$$0.5082 * 0.2 + 1.8 = 1.9016 .$$

Thus, the value of x(356) reconstructed by linear interpolation is approximately 1.902 picocuries of radium per liter of drinking water. Similarly, x(43) = 0.804.

The values have been determined for the most efficient two quantiles for estimating the location and scale parameters. Next these quantiles are used with the coefficients given in Table 9.A.2 to form the linear combinations that estimate the parameters. From Table 9.A.1, c(1) = 0.5680, c(2) = 0.4320, and from Table 9.A.3 d(1) = -0.4839, d(2) = 0.4839. The scale and location parameter estimates are then:

$$\hat{u} = 0.5680 * 0.8042 + 0.4320 * 1.9016 = 1.2783$$

$$\hat{a} = -.4839 * 0.8042 + 0.4839 * 1.9016 = 0.5311$$

That is, based on the grouped data the Extreme Value distribution mode (location parameter) is 1.3 picocuries per liter and the scale parameter is 0.5 picocuries per liter of Radium in well water.

The variances and covariances of these estimates are calculated using the multipliers given in Table 9.A.4 and equations 9.12 through 9.14. From this table, E(1) = 1.5106, E(2) = 1.0749, and E(3) = -0.3401.

$$V(\hat{u}) = (0.5311^2/485)*1.5106 = 0.00088$$

$$V(\hat{a}) = (0.5311^2/485)*1.0749 = 0.00063$$

$$Cov(\hat{u},\hat{a}) = -(0.5311^2/485)*-0.3401 = 0.0002$$

and the corresponding standard deviations are:

$$SD(\hat{u}) = 0.0296$$

$$SD(\hat{a}) = 0.0250$$

$$Correlation(\hat{u},\hat{a}) = 0.27 \quad .$$

The location and scale parameters are significantly correlated. Assuming that a 95% confidence interval or the distribution of the parameter values is ± 2 standard deviations, the 95% confidence interval for the location parameter is $1.2783 \pm 2*0.0296$ picocuries per liter, or 1.22 to 1.34. Similarly the 95% confidence interval for the scale parameter is 0.48 to 0.58 picocuries per liter.

9.4 SUMMARY

This chapter presents some of the available methods for obtaining parameter estimates for an Extreme Value distribution from grouped or censored data. A method is given to obtain minimum variance unbiased estimates when only the largest extremes are reported. A simple regression method for the same kind of data is given. When sample size is large, the data is sometimes reported as frequencies within specified intervals or groups. Several methods are discussed to obtain parameter estimates from grouped data. Finally, a rigorous method of calculating the location and scale parameters from sample quantiles is presented.

9.5 REFERENCES

Hassanein, K. M., 1972, 'Simultaneous Estimation of the Parameters of the Extreme Value Distribution by Sample Quantiles', Technometrics, Vol. 14, pp 63 - 70.

Mann, N. R., and Fertig, K. W., 1977, 'Efficient Unbiased Quantile Estimators for Moderate-Size Complete Samples from Extreme Value and Weibull Distributions, Confidence Bounds and Tolerance and Prediction Intervals', Technometrics, Vol. 19, No. 1, pp 87 - 93.

Schupbach, M., and Husler, J., 1983, 'Simple Estimators for the Parameters of the Extreme-Value Distribution Based on Censored Data', Technometrics, Vol. 25, No. 2, pp 189 - 192.

Weissman, I., 1978, 'Estimation of Parameters and Large Quantiles Based on the k Largest Observations', J. Am. Stat. Assoc., Vol. 73, pp 812 - 815.

9.6 APPENDIX 9-A

EXTREME VALUE DISTRIBUTION
Tables for Calculation of Location and Scale Parameters from Grouped Data

Table 9.A.1
Coefficients for Choosing Quantiles for Maximum Effeciency
k = number of quantiles to be used
I = index of coefficient
Body of Table contains p(I)

	k					
I	2	3	4	5	6	7
1	.087	.055	.028	.018	.011	.008
2	.734	.439	.193	.114	.071	.047
3		.850	.604	.404	.251	.163
4			.896	.726	.547	.396
5				.931	.799	.652
6					.951	.849
7						.964

Table 9.A.2
Linear Combination Coefficients for Determining Location Parameter
k = number of quantiles to be used
I = index of coefficient
Body of table contains c(I)

	k					
I	2	3	4	5	6	7
1	.5680	.3386	.1566	.0994	.0623	.0439
2	.4320	.5184	.4316	.3030	.2027	.1382
3		.1430	.3250	.3673	.3315	.2649
4			.0868	.1804	.2564	.2813
5				.0499	.1144	.1727
6					.0327	.0764
7						.0226

Table 9.A.3
Linear Combination Coefficients for
Determining Scale Parameter
k = number of quantiles to be used
I = index of coefficient
Body of Table Contains d(I)

	k					
i	2	3	4	5	6	7
1	-.4839	-.4372	-.2845	-.2047	-.1454	-.1112
2	0.4839	0.1602	-.1526	-.2236	-.2189	-.1854
3		0.2770	0.2651	0.1012	-.0481	-.1254
4			0.1720	0.2208	0.1733	0.0780
5				0.1063	0.1673	0.1680
6					0.0718	0.1251
7						0.0508

Table 9.A.4
Multipliers for Asymptotic Variances and Covariances
of Location and Scale Parameters
k = number of quantiles to be used

	k					
	2	3	4	5	6	7
Var(u)	1.5106	1.2971	1.2287	1.1924	1.1706	1.1567
Var(a)	1.0749	0.9028	0.7933	0.7374	0.7043	0.6825
Cov(u,a)	-.3401	-.2579	-.2570	-.2674	-.2657	-.2638

CHAPTER 10
EXTREMES OF SMALL SAMPLES

10.1 INTRODUCTION

In previous chapters it is assumed that the samples from which the extremes were selected are large. Fisher and Tippet show that for samples from a Gaussian distribution, the sample size has to be infinite for the Extreme Value distribution to hold exactly (Chapter 1). In statistics, approximate asymptotic properties are usually adequate. However there are some situations in which the asymptotic distributions yield significantly biased results. No general rules exist that allow one to determine when asymptotics are adequate.

This chapter outlines statistics that can be used when sample sizes are small. The specialties of 'Order Statistics' and 'Simultaneous Inference' contain the mathematical tools necessary to analyze small sample extremes. The results are not as simple as for the asymptotic distributions, but this is the price of small samples. Order statistics gives the densities and distributions of the largest and smallest members of a sample of known size. Simultaneous inference shows how to collectively consider a set of (possibly correlated) probability statements.

The principal references for this chapter are the textbooks by David (1970) and by Miller (1966). Most mathematical statistics textbooks contain the basic ideas of order statistics, often indexed by such terms as 'the distribution of the range', or 'the distribution of quantiles'.

10.2 ORDER STATISTICS

Order statistics studies the statistical properties of maximum and minimum values, the range, extreme deviates from the mean, quantiles, the second from largest value, and the joint distributions of these statistics. In this short overview only the largest extreme from a small sample will be considered.

If X(1),X(2),...X(n) is a random sample from a continuous population, the r<u>th</u> largest of these values is called the r<u>th</u> order statistic, its value will be denoted as x[r]. Thus, the smallest sample value is x[1] and the largest is x[n]. Since the distribution of the X's, F(x), may be interpreted as the probability that X has a value less than or equal to some specified value x, the probability that exactly j of the X's lie in the closed interval $(-\infty,x]$ and (n-j) lie in the open interval (x,∞) is obtained from substituting F(x) for the probability in the Binomial series:

$$\binom{n}{j}F^{j}(x)(1 - F(x))^{n-j} \quad . \tag{10.1}$$

The event $x[r] \leq Z$ occurs if and only if r or more of the X(i)'s lie in the interval $(-\infty,Z]$. Thus,

$$F(x[r]) = P(x[r] \leq Z) = \sum_{j=r}^{n}\binom{n}{j}F^{j}(Z)(1 - F(Z))^{n-j} \quad . \tag{10.2}$$

In particular, the distribution function of the largest and smallest members of a sample from a population with distribution F(X) are:

$$F(x[n]) = (F(X))^{n} \quad , \text{ and} \tag{10.3}$$

$$F(x[1]) = 1 - (1 - F(x))^{n} \quad . \tag{10.4}$$

The corresponding density functions are found by differentiation to be

$$f(x[n]) = nf(X)F(X)^{n-1} \quad \text{, and} \tag{10.5}$$

$$f(x[1]) = nf(X)(1 - F(X))^{n-1} \quad . \tag{10.6}$$

This statistical reasoning can be extended to show that the distribution of the r$\underline{\text{th}}$ order statistic is

$$f(x[r]) = \frac{n!}{(r-1)!(n-r)!} f(X)F(X)^{r-1}(1 - F(X))^{n-r} \quad . \tag{10.7}$$

An interesting result of order statistics concerns the sampling distribution of the median of the density f(x). Let M(x) denote the median. Then for large n, the sampling distribution of M(x) for random samples of size (2n+1) is approximately Gaussian with mean equal to the population median and variance (Wilk, 1947, Chapter 4)

$$v = \frac{1}{8(f(M(X))^2 n} \quad . \tag{10.8}$$

If X is Gaussian distributed, then the variance of the median is approximately

$$pi*s^2/(4n)$$

(where s is the standard deviation of the distribution and pi = 3.14159... .) Comparing this with the variance of the mean, which for samples of size (2n+1) is

$$s^2/(2n+1) \quad ,$$

shows that, for large samples from a Gaussian population, the mean has a smaller variance than the median.

10.2.1 Approximating the Distribution of a Single Order Statistic.

Consider the summation of binomial terms of equation 10.2, the probability that j of the x(i) are less than or equal to some specified value. There is a relationship between binomial sums and the incomplete beta function I, that gives

(10.9) $$F(x[r] \leq z) = I(F(z),r,n-r+1) \quad .$$

Tables of the incomplete beta function are available in Beyer (1966). In order to use such tables it is often necessary to employ the inversion relationship:

(10.10) $$I(P,a,b) = I(1-P,b,a) \quad .$$

As an example, suppose one wishes to find the upper 5% limit of the fourth order statistic, x[4], from a sample of size 5, from a standardized (mean of 0.0, variance of 1.0) Gaussian distribution. Equations 10.2 and 10.9 show that this is equivalent to finding z such that

$$I(F(z),4,2) = 0.95 \text{ , or}$$
$$I(1-F(z),2,4) = 0.05 \text{ .}$$

The table on page 210 of Beyer gives the lower 5% point of the incomplete beta function, which by the inversion formula 10.10 is equivalent to the upper 95% point. To read the table use v(1) = 2*b = 2*4 = 8, and v(2) = 2*a = 2*2 = 4. The tabled value is 0.07644 which is the desired value for 1-F(z). Next use a table of the Gaussian distribution to find that value of z that corresponds to F(z) = 1 - 0.07644 = 0.92356. This value is very close to z = 1.43. Then for a standardized Gaussian distribution, the second largest observation in a sample of size 5 will be less than or equal to 1.43 with probability of 0.95.

Frequently several simple statistical tests are performed on the data from a single experiment. Suppose river water is sampled and tested for concentrations of 10 pollutants using a Student's t-test with 14 degrees of freedon for each pollutant. An experiment-wide 95% confidence (probability of a type I error) is desired. That is, with 95% confidence the statement is to be made that the combined t-tests indicate that no significant levels of any pollutants was found. This is equivalent to the statement: the maximum t-value of the 10 t-tests performed is within the 95% confidence limit of the largest order statistic in a sample of size 10 of a t-distribution with 14 degrees of freedom. Using the incomplete beta function, this may be formulated as: find t such that

$$I(F(t),10,1) = 0.95 ,$$

where F(t) is a Student's t distribution function with 14 degrees of freedom. In order to use the tables of the incomplete beta function it is necessary to apply the inversion relationship,

$$I(1-F(t),1,10) = 0.05 .$$

The tabulated value for 1 - F(t) is 0.028358. From a table of the t-distribution, the value of t that corresponds to F(t) = 0.9716 for 14 degrees of freedom is approximately 2.068. Thus, to be 95% confident that all of 10 t-tests are simultaneously not significant, the maximum of those ten 14 degree of freedom tests has to have a t-value less than 2.068, which is the 97% significance critical value for a single test. A 2% penalty is payed for considering the 10 tests as a single experiment. The t-value for 95% confidence on a single t-test with 14 degrees of freedom is 1.761. If it seemed that all 10 pollutant measures might be significant, the lower confidence limit or 5% significance limit of the maximum of the 10 t-tests would be used to test the hypothesis that all the pollutants were significant. More complicated situations arise if only some of the t-values are significant; this is best handled by a multivariate test

procedure.

If tables of the cumulative Binomial distribution are available (Beyer, 1966, Table III.2, or National Bureau of Standards, 1950) equation 10.2 can be solved without the need to transform into an incomplete beta function. For the example above, enter the tables in Beyer at n=5 and x'=2, then look across the line for the value of p that has a table entry of 0.05. Interpolation between table enteries of 0.0226 for p=0.05 and 0.07326 for p=0.10 is necessary. A linear interpolation yields p=0.07326. Using tables of the Gaussian distribution to find z such that 1 - F(z) = 0.07326 yields z=1.452. The table in Beyer gives the summation of binomial terms from r to n. A more commonly available form of such tables (Odeh et. al., 1977, Table 24; or Conover, 1971, Table 3) gives the summation from zero to r. This type of table is often found in nonparametric statistics textbooks. Such tables can be used with the relationship that the probability summed over all values if j must equal unity, thus

$$(10.11) \qquad \sum_{j=r}^{n}\binom{n}{j}F^{j}(1-F)^{n-j} = 1 - \sum_{j=0}^{r-1}\binom{n}{j}F^{j}(1-F)^{n-j} \quad .$$

If a computer is available, the value of F(x) in I(F(x),r,n-r+1) can be calculated using an algorithm for the inverse of the incomplete beta function (Majumder and Bhattacharjee, 1973; update by Cran et. al., 1977).

These examples are both an introduction to those aspects of order statistics that are most applicable to the study of extreme values from small samples, and also background for the next section which addresses the same problem in a different way.

10.3 SIMULTANEOUS STATISTICAL INFERENCE

In Section 10.2 the concept of finding one probability statement for a group of statistical tests is introduced, rather than considering each test separately. If 20 t-tests are performed at the 95% confidence level, on the average, one of the results will be in error. The basic purpose of simultaneous inference is that, by treating the tests as a group, a probability statement can be made that is simultaneously valid for all members of the group. For 20 t-tests, one is able to say that all 20 are not significant with 95% confidence rather than that each of the 20 is not significant at 95% confidence. In the latter case there is high probability (64% to be exact) that at least one such conclusion in a group of 20 is in error. A philosophical point arises here, how big should the group be? The general opinion among statisticians is that a group should include all statistical tests performed on a single data set. An extensive study of simultaneous inference is given by Miller (1966, 1977).

Order statistics and simultaneous inference approach the problem of testing groups of hypotheses in different ways. The approaches are equivalent and the choice of approach depends upon convenience and the specifics of the problem. Order statistics picks the maximum (or minimum) of the group and derives a probability statement about the maximum (or minimum) as a function of sample size. If the maximum satisfies an order statistic hypothesis of not exceeding its expected value, then it necessarily follows that all values of the test statistic smaller than the maximum also satisfy the hypothesis. Simultaneous inference approaches the same problem by altering the probability test applied to each statistic of the group so that an overall probability statement can be made about the group as a whole if all of the individual statistics satisfy the altered test.

The most familiar application of simultaneous inference is in analysis of variance. Whenever the analysis concludes that there is a significant

difference between means, the next step is to examine the relationships between those means to find the source of the significance. The tests for these relationships are familiar under the names F-projection or Scheffe test, multiple range test or Duncan test, least significant differences or Fisher test, and Studentized range or Tukey test.

The Bonferront inequality, one of many available methods from simultaneous inference, is presented in this chapter. Miller (1977) observes: 'I have become even more impressed with the tightness of the bound...', in his discussion of studies of the Bonferroni inequality over the years 1966 to 1976.

10.4 Bonferroni Statistics.

The Bonferroni method is a simple adjustment of probability levels that can be applied to any statistical hypothesis test or confidence limit procedure to produce results that are valid for a group of statistics. If A(i) is an event that can be assigned a probability and if there are n such events in a group, the Bonferroni inequality states that

(10.12) The probability of the intersection of n events A(i) is greater than or equal to 1.0 minus the sum of the compliments of the probabilities of the individual events.

The term 'intersection' is the mathematically precise way of saying that all events are considered as a group. This inequality is a simple extension of Boole's inequality:

$$(10.13) \qquad P(A \text{ or } B) \leq P(A) + P(B) \quad .$$

The Bonferroni statistic is obtained by applying the Bonferroni inequality to any group of statistical tests. The stated error level, for

example an error level or alpha of 5%, is divided by the number of tests in the group to obtain a new error level, call this b. The sum of n b's represents the sum of probabilities on the right hand side of equation 10.13. The inequality of this equation then allows this sum to be related to the intersection of the events. For example, suppose the second example of section 10.2 is reanalyzed, in which 10 pollutants in a single water sample are analyzed using t-test. A 95% significance level corresponds to an error level of 0.05. With 10 tests in the group the b level is 0.05/10 or 0.005. The Bonferroni statistic tells one to test each of the pollutants at the 99.5% significance level; if all 10 tests show no significance individually at 99.5% significance, then, with at least 95% significance, all 10 pollutants are simultaneously not significant. The t-test critical value for 99.5% significance and 14 degrees of freedom is 2.98. (The corresponding 95% critical value is 1.76.) The critical value found in section 10.2 was 2.07. The difference between 2.98 and 2.07 is an expression of the inequality within the Bonferroni statistic: it gives an upper bound. The advantages of the Bonferroni statistic are its great simplicity, special tables are not required, and its applicability to all statistical situations.

Care should be used when finding the new critical values from a table of probabilities. Many tables do not contain very small error levels such as those that result from dividing by the number of tests in the group. Also, tables differ widely in how they express error levels; one or two sided, using error levels or significance levels (tail or central areas). In the example, a one-sided t-test and a one-sided table is used. If a two-sided test is desired from one-sided tables, the Bonferroni statistic divides the error level by 2*n.

Dunn (1959) showed that the Bonferroni statistic is applicable to correlated as well as to independent statistical tests. The order statistic results presented in section 10.2 are also valid for correlated statistics. In order to prove this, it is necessary to use the

Union-Intersection principle (Morrison, 1967, section 4.2).

10.5 COMBINING INDEPENDENT PROBABILITIES

Two simple techniques are available to combine independent probability tests into an overall probability for a group of statistical tests; the sum of Chi-square values and Fisher's method.

When a number of independent tests of significance are applied it sometimes happens that although none can be individually claimed as significant, the aggregate gives the impression that on the whole, the probabilities are lower, or higher, than would be obtained by chance alone. Two theorems found in statistics textbooks are usefull in deriving an aggregate significance statement about a group of independent statistical tests.

THEOREM 1

The sum of independent Chi-square values is also a Chi-square value with degrees of freedom equal to the sum of degrees of freedom of the individual values.

Perhaps a series of Chi-square contingency table analyses are performed and all are just a few percent too low to be judged significant. Many tests that are close to significance is unlikely to be an aggregate event that is due to chance alone. Theorem 1 gives the theoretical basis to sum all the Chi-square values to obtain an aggregate test of significance. The second theorem allows this technique to be extended to tests that obtain significance levels from a Gaussian distribution.

THEOREM 2

The square of a standardized (subtract mean and divide by standard deviation) Gaussian-distributed value is a Chi-square distributed value with one degree of freedom.

Suppose a series of Mann-Whitney Rank Sums tests (the Wilcoxon Rank Sums test is algebraically equivalent) are performed and an aggregate situation similar to the one mentioned for the contingency table tests is found. If individual sample sizes are large, the significance of each Rank Sums test is determined by calculating a value that has approximately a Gaussian distribution. Then the individual significance levels can be closely approximated using a table of the Gaussian distribution. The aggregate significance can be obtained by squaring and summing these values, and comparing the sum to a Chi-square table using degrees of freedom equal to the number of items summed. These two theorems are useful only when Chi-square or Gaussian values are available. They cannot handle other distributions or combinations of several distributions.

Fisher's Combination of Probabilities Test of Significance (Fisher, 1970, section 21.1) can be used to test the significance of an aggregate of any group of significance levels from any distribution or even several different distributions. Let P(I) represent the significance (or probability) resulting from the Ith test in an aggregate of size n. Fisher found that minus twice the natural logarithm of P(I) has a Chi-square distribution with 2 degrees of freedom. It follows then from Theorem 1 that:

$$(10.14) \qquad X(2n) = -2\sum_{I=1}^{n} \ln(P(I))$$

has a Chi-square distribution with 2n degrees of freedom. The aggregate probability may then be found by comparing X(2n) to a Chi-square table.

For an example of these two methods of combining independent probabilities, suppose a group of 5 Student's t-tests each showing 90% confidence. The corresponding probability is 0.10, and -2*ln(0.10) = 4.605. The sum of 5 such identical values is 23.026. For 10 degrees of

freedom, a Chi-square table shows that the aggregate significance is 98.9%. Intuitively 5 tests at 90% significance suggest an aggregate significance even though individually the significances do not meet the usual accepted criteria of exceeding 95% (the 95% criteria for significance was proposed by R. A. Fisher in 1926). Fisher's method shows that the aggregate is highly significant. This example also shows that if instead of a group of 5 t-tests, one has a group of 5 Chi-square tests each with 2 degrees of freedom and values of 4.605 (90% significance), the aggregate would have 98.9% significance. Likewise, a group of 5 z-tests, each with a value of 1.645 (90% significance) would be significant. Squared and summed, these give a Chi-square value of 13.530 with 5 degrees of freedom. A Chi-square table shows this aggregate to have 98.1% significance.

10.5.1 Maximum Chi Square

In the discussion of Section 10.4 it was assumed that the probabilities or distributions being combined are independent. For combining independent or dependent Chi-square values, the Union - Intersection principle of Roy (1953) leads to the following result.

Theorm 3

The maximum of p one degree of freedom Chi-square values has a Chi-square distribution with p degrees of freedom.

Thus, if one analizes 10 correlated 2-by-2 contingency tables and finds that the maximum Chi-square value is less than the 95% deviate of a 10 degree of freedom Chi-square distribution, then it may be concluded that simultaneously all 10 contingency tables show no significance at the 95% confidence level.

It can be shown, using characteristic functions and Equation 10.3, that the maximum Chi-square principle holds for independent Chi-square values with any degrees of freedom. The maximum of a group of independent

Chi-square values has a Chi-square distribution with degrees of freedom equal to the sum of the degrees of freedom of the group.

10.6 SUMMARY

A necessary assumption for all the techniques presented in previous chapters is that the extreme values be obtained from a large sample. This chapter has presented alternatives that can be used when this assumption cannot be accepted. Extremes of small groups of statistical tests and of small samples are important because they are a common situation. Three complementary statistical methods were presented.

Order statistics show how to derive the exact distribution of the extreme of a small sample if the distributions of the elements of the sample are known. Often the algebra of this derivation is intractable. As an alternative, order statistics offers a way, through the use of binomial sums and the incomplete beta function, to compute probabilities without finding the expression for the distribution of an extreme.

Simultaneous statistical inference yields methods for obtaining an aggregate probability statement about a series or group of statistical tests. The most common use of these methods is to identify the source of significance within an analysis of variance problem. The simplest of such statistical tools, the Bonferroni inequality, is presented. This inequality gives an upper bound for aggregate confidence statements and hypothesis testing. It has the advantages of being simple to implement, and being valid with correlated data or correlated statistical tests.

In the special situation in which the elements of the group of tests are statistically independent, some theorems from probability theory can be used to derive an aggregate probability statement about the group as a whole. These techniques are based upon the properties of the Chi-square distribution, and upon a relationship, discovered by R. A. Fisher, between

the Chi-square distribution and the natural loga-ithm of a probability.

10.7 REFERENCES

Beyer,W. H., 1966, (Editor) Handbook of Tables for Probability and Statistics, The Chemical Rubber Publishing Co., (Section III.10).

Conover, W. J., 1971, Practical Nonparametric Statistics, John Wiley and Sons. (This book is a typical Nonparametric Statistics textbook.)

Cran, G. W., Martin, K. J., and Thomas, G. E., 1977, Remark AS R19 and Algorithm AS109, 'A Remark on Algorithms AS63: The Incomplete Beta Integral, AS64: Inverse of the Incomplete Beta Function Ratio', Applied Statistics, Vol. 26, No. 1, pp 111 - 114.

David, H. A., 1970, Order Statistics, John Wiley and Sons, Inc. (2nd edition, 1981.)

Dunn, O. J., 1959, 'Confidence Intervals for the Means of Dependent Normally Distributed Variables', J. Amer. Statistical Assoc., Vol. 54, pp 613 - 621.

Fisher, R. A., 1926, 'The Arrangement of Field Experiments', J. of Ministry of Agriculture, Vol. XXXIII, pp 503 - 513. Reprinted in: R. A. Fisher, 1950, Contributions to Mathematical Statistics, John Wiley and Sons.

Fisher, R. A., 1970, Statistical Methods for Research Workers, 14th Edition, Oliver and Boyd, Edinburgh.

Odeh, R. E., Owen, D. B., Birnbaum, Z. W., and Fisher, L., 1977, Pocket Book of Statistical Tables, Marcel Decker, Inc. (Table 24).

Majumder, K. L., and Bhattacharjee, G. P., 1973, Algorithm AS64 'Inverse of the Incomplete Beta Function Ratio', Applied Statistics, Vol. 22, No. 3, pp 411 - 414.

Miller, R. G., 1966, Simultaneous Statistical Inference, McGraw-Hill.

Miller, R. G., 1977, 'Developments In Multiple Comparisons 1966-1976', J. Amer. Statistical Assoc., Vol. 72, No. 360, pp 779 - 788.

Morrison, D. F., 1967, Multivariate Statistical Methods, McGraw-Hill.

National Bureau of Standards, 1950, 'Tables of the Binomial Probability Distribution', Applied Mathematics Series, No. 6.

Roy, S. N., 1953, 'On a Heuristic Method of Test Construction and Its Use In Multivariate Analysis', Ann. Math. Stat., Vol. 24, pp 220 - 237.

Wilk, S. S., 1944, Mathematical Statistics, Princeton Univ. Press.

CHAPTER 11
MULTIVARIATE EXTREMES

11.1 INTRODUCTION

The currently available statistical theory of multivariate Extreme Value distributions considers each component of a (standardized) multivariate observation separately. The maximum (or minimum) of each of n components of the vector is determined and the joint distribution of these maxima is studied (Galambos, 1978). It is important to note that this theory does not consider a single multivariate (vector) measurement that is the extreme, but rather it abstracts pieces from many measurements and considers these pieces jointly. For example, rather than considering the most toxic mixture of chemicals, the most toxic concentration of each chemical is considered, ignoring the synergistic effects of the mixture.

Consider the univariate unit (mean = 1.0) Exponential distribution $F(x) = P(X<x) = 1-\exp(-x)$. The simplest of many possible two-dimensional analogs is

$$F(x,y) = 1 - \exp(-x) - \exp(-y) + 1/(\exp(x) + \exp(y) - 1) .$$

The small sample distribution of the maximum of this distribution, the multivariate analog of equation 10.3, for a sample size of n is

$$(F(x,y))^n$$

and the asymptotic distribution on large sample size, the multivariate analog of equation 7.4, is

$$H(x,y)=\exp(-\exp(-x))*\exp(-\exp(-y))*\exp(1/(\exp(x)+\exp(y))) .$$

Galambos (1978) gives the theory for deriving such equations for any kind of distribution. This example shows that the limit distribution H is not determined by the univariate marginal distributions. However, the marginal distributions of H are univariate Extreme Value distributions.

This mathematical theory considers multivariate observations componentwise. The probability that the maxima of all components will occur in the same observation is small and is a decreasing function of sample size. There is no special theory of multivariate extreme values for the case in which one is interested in the particular multivariate observation from a sample that is extreme when all components are considered simultaneously. For example, one might be interested in identifying the smoggiest day of the year from daily averages of carbon monoxide, nitrous oxides, sulfur oxides, and hydrocarbons. There is no reason to expect that the maxima of these component chemicals in a year of data would occur together, nor that the maxima of any component would occur on the day that was perceived by humans to have the maximum smog concentration.

Even though no special statistical theory exists to determine the simultaneous extreme, a few useful techniques are available that allow one to statistically analyze such data. These techniques are derivatives of statistical concepts used in multivariate statistics. The basis of these techniques is to transform the multivariate data into an equivalent univariate number, and then apply the methods presented in previous chapters. Each of the transformation techniques presented in this chapter emphasizes different data characteristics. The transformations are not equivalent, and will not yield the same conclusions. The choice of a transformation depends upon the purpose of the study. Since these emphases are manifold, particularly for the cluster analysis techniques, details are not discussed here. Complete discussions are available in textbooks and in the statistics literature.

Radioisotope concentrations represent a special class of multivariate measurements, because all the isotopes in a sample can be measured in the same cumulative units. If soil samples are measured for uranium, thorium, lead and radium concentrations, all in picocuries per gram, the total radioactivity is the sum of the picocuries per gram of each component. The sum of the radioactivity could also be expressed in rems or rads, but not in micrograms of isotope per gram of soil. Air pollutants cannot not be summed because there is now no measure of concentration for toxic gasses that is analogous to the way rem's quantify biological activity, or to the way the Curie measures nuclear disintegrations.

11.2 DISTANCE MEASURES

If the covariance matrix of the components of the multivariate measurements is known or can be estimated, the Mahalanobis distance (generalized distance, Euclidian distance, l-2 norm) can be calculated. Let $\underline{Y(i)}$ be a vector observation in a sample of size n, i = 1,2,...n. Each $\underline{Y(i)}$ has several components, such as measurements of isotope concentrations of different elements in a single sample. Let $\underline{V}$ be the covariance matrix of the components of $\underline{Y}$. Then the univariate generalized distance of $\underline{Y(i)}$ from the origin is the square root d(i) of D(i) where:

$$D(i) = \underline{Y(i)}^T \underline{V}^{-1} \underline{Y(i)} \quad . \tag{11.1}$$

The n values of d(i) or of D(i) can be treated as a univariate sample. If n is large, the maximum of the n values is approximately a sample from an extreme value distribution. If some of the elements of the $\underline{Y(i)}$ are considered more important, a weighting matrix can be included in the distance calculation; this is discussed in textbooks on multivariate statistics. This distance measure cannot be thought of as analogous to any particular element of $\underline{Y(i)}$; it must be considered as an abstraction that includes all elements.

As an example, suppose an overall measure of maximum air pollution is desired. Daily averages are available in parts per million of ozone, nitrous oxides, sulfur oxides, and hydrocarbons. These are the four components of each $\underline{Y(i)}$ vector, and there is one such vector for each day of a year ($i = 1,2,...365$). The covariance matrix of the elements can be estimated from the 365 $\underline{Y(i)}$ vectors. Then 365 generalized distances can be computed and the largest is the maximum overall pollution for the year. If this procedure is repeated for N years, the N maximum values of the distance can be used to find an extreme value distribution since n, the sample size within each year, is large. For a given distance, there is no unique set of values of the components.

The generalized distance usually will not yield a maximum that corresponds to the maximum of any one of the individual pollutants. Nor will it necessarily yield a maximum that corresponds to the greatest pollution perceived by the residents of the area.

In this example only gasseous pollutants that are typically measured in parts per million are included: this was intentional. Significant artifacts can be introduced by a naive choice of components and units. Many multivariate statistical test are not invariant to changes in scale and origin of the measurements.

For a small sample of size n from a multivariate Gaussian distribution, the distribution of the D(i) has a Hotelling's distribution if the covariances are estimated from the data, and a Chi-square distribution if the covariances are known apriori. Hypotheses can be tested using the methods of simultaneous inference outlined in Chapter 10.

The generalized distance becomes computationally unstable if the components of the $\underline{Y(i)}$ vector are highly correlated. In such a case, it is advisable to use a generalized matrix inversion algorithm to invert the covariance matrix. This yields a minimum distance measure. Specifically,

if the covariance matrix is singular, it is known from matrix algebra that no unique inverse exists. This is the same as asserting that an infinite number of inverses do exist (Boullion and Odell, 1971). In this case, a different value of the generalized distance will result for each of the infinite number of possible inverses. However if the generalized matrix inverse is used, the resulting generalized distance will be the minimum of all the possible values. Using a generalized matrix inversion algorithm is a good way of avoiding the computational problems of nearly singular matrices.

Another possible multivariate to univariate transformation is to use a probability value from the multivariate distribution function of the data. The multivariate observation (x,y,z,...) is replaced by the probability that values smaller than those observed would occur, that is p = P(X<x,Y<y,Z<z,...). Since the range of values of p is zero to unity, the extreme of the p values must be of the Weibull family of distributions. When determining the value of p, the multivariate distribution of the data must be known (or hypothesized) and it must account for the covariances between the components of the measurements. The only multivariate distribution that is well established for more than two components is the Multivariate Gaussian.

11.3 ORTHOGONAL ROTATIONS

A multivariate to univariate transformation may also be obtained from an orthogonal rotation of the multivariate axes followed by a choice of one of the resulting projections. The typical way of performing such calculations is to use either factor analysis or principal components analysis. These are described in multivariate statistics textbooks. Principle components studies the variance of the multivariate data elements, while factor analysis studies the correlations of these elements. Both procedures yield a series of linear combinations of the data values that are ranked in importance by the amount of information from the data

that is explained by each linear combination. For extreme value analysis, the first principle component or the first factor can be used as a univariate generalization of the data. Using only one factor or component does not use all the information available in the data. However, the purpose of these procedures is to simplify the data structure, and these techniques do this by dividing the information into parts and ignoring the less significant parts. There is no requirement that the first component or factor be chosen for extreme value analysis. In the air pollutant example, perhaps a second or third factor or component would be better associated with the severity of pollution as perceived by humans. The currently available statistical analysis computer program packages contain good algorithms for obtaining principle components and factors. Thus, it is easy to use (and missuse) these procedures for the multivariate to univariate transformation needed for extreme value analysis.

For computational purposes in this example, only one of the 4 species, Iris setosa is considered in detail.

Fisher gives the covariance matrix of this species:

TABLE 11.1
Iris setosa Covariance Matrix

6.0882	4.8616	0.8014	0.5062
	7.0408	0.5732	0.4556
(Symmetric)		1.4778	0.2974
			0.5442

The inverse of this matrix is:

TABLE 11.2
Inverse of Covariance Matrix

0.38860	-.25316	-.09184	-.09747
	0.31777	0.02268	-.04294

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
0.1	5	*****
0.2	29	*****************************
0.3	7	*******
0.4	7	*******
0.5	1	*
0.6	1	*

FIGURE 11.4
Petal Width

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
1.90	1	*
1.95	3	***
2.00	2	**
2.05	3	***
2.10	2	**
2.15	3	***
2.20	13	*************
2.25	5	*****
2.30	7	*******
2.35	5	*****
2.40	3	***
2.45	2	**
2.50	1	*

FIGURE 11.5
Mahalanobis Distance

the extremes of each component into a single measure of extreme value.

11.5 CLUSTER ANALYSIS TECHNIQUES

Cluster analysis offers the most general methods of measuring distances between multivariate observations. These methods are only recently available in textbooks (Everitt, 1980). However, statistical journals contain ample information on the wide variety of clustering techniques available. It is important to emphasize that this variety results from a diversity in the type of information authors are attempting to expose from within their data. Clustering algorithms range from parametric to nonparametric in genesis, and originate from a variety of scientific fields.

Analyzing extreme values of cluster distances is a powerful tool because the variety of distance measures available offers a choice appropriate to the purpose of the analysis at hand. The theories of cluster analysis are well developed statistically, but the clustering algorithm must be chosen carefully in order to assure an appropriate measure for the problem being analized.

11.6 EXAMPLE

A classic data set of the statistical literature, R. A. Fisher's (1936) Iris data, is used as an example. It has an analogy to pollutant data. This Iris data had a key role in the development of Discriminant Analysis (Fisher, 1936, 1938), a technique that uses Mahalanobis distance measure. Fisher (1936) gave tables of 50 observations on each of 4 species of Iris. Each observation consisted of 4 components: septal width, septal length, petal width, and petal length. As an analogy, suppose that the 4 species are 4 pollution measurement stations and that the 4 components are concentrations of ozone, nitrous oxides, sulfur oxides, and hydrocarbons. For each station, 50 hourly averages of the four pollutants are available.

For computational purposes in this example, only one of the 4 species, Iris setosa is considered in detail.

Fisher gives the covariance matrix of this species:

TABLE 11.1
Iris setosa Covariance Matrix

6.0882	4.8616	0.8014	0.5062
	7.0408	0.5732	0.4556
(Symmetric)		1.4778	0.2974
			0.5442

The inverse of this matrix is:

TABLE 11.2
Inverse of Covariance Matrix

0.38860	-.25316	-.09184	-.09747
	0.31777	0.02268	-.04294
(Symmetric)		0.79135	-.36620
			2.16420

The data (too voluminous to present here) is summarized in the following histograms of the values of the 4 components and of the Mahalanobis distances.

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
4.2	0	
4.4	4	****
4.6	5	*****
4.8	7	*******
5.0	12	************
5.2	11	***********
5.4	6	******
5.6	2	**
5.8	3	***

FIGURE 11.1
Septal Length

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
2.2	0	
2.4	1	*
2.6	0	
2.8	0	
3.0	7	*******
3.2	9	*********
3.4	11	**********
3.6	9	*********
3.8	7	*******
4.0	3	***
4.2	2	**
4.4	1	*

FIGURE 11.2
Septal Width

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
1.0	1	*
1.1	1	*
1.2	2	**
1.3	7	*******
1.4	13	**********
1.5	13	**********
1.6	7	*******
1.7	4	****
1.8	0	
1.9	2	**

FIGURE 11.3
Petal Length

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
0.1	5	*****
0.2	29	*****************************
0.3	7	*******
0.4	7	*******
0.5	1	*
0.6	1	*

FIGURE 11.4
Petal Width

MIDDLE OF INTERVAL	NUMBER OF OBSERVATIONS	
1.90	1	*
1.95	3	***
2.00	2	**
2.05	3	***
2.10	2	**
2.15	3	***
2.20	13	*************
2.25	5	*****
2.30	7	*******
2.35	5	*****
2.40	3	***
2.45	2	**
2.50	1	*

FIGURE 11.5
Mahalanobis Distance

TABLE 11.3
Partial Listing of Iris Data

ROW	SL	SW	PL	PW	DISTANCE
1	5.1	3.5	1.4	0.20	2.23489
2	4.9	3.0	1.4	0.20	2.19540
3	4.7	3.2	1.3	0.20	2.05929
4	4.6	3.1	1.5	0.20	2.09104
5	5.0	3.6	1.4	0.20	2.18819
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
19	5.7	3.8	1.7	0.30	2.52440
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
50	5.0	3.3	1.4	0.20	2.20816

Using equation 11.1 to compute generalized distances gives the numbers summarized in Figure 11.5. Table 11.3 gives a few of the data values. The maximum of the 50 distances was found to be at data row 19. Since all the measurements are in the same units (millimeters) the maximum Mahalanobis distance is 2.524 millimeters. In general the measurement units of the elements of a multivariate measurement will not be identical, then the distance is only a mathematical number. Scanning this data set reveals that the maximum distance does not occur at the maximum of any of the individual components of the multivariate data values. A correlation coefficient goodness-of-fit test for the Gaussian distribution, discussed in Appendix B of Chapter 7, was performed on the computed distances. This test concluded that the distances are reasonably described by a Gaussian distribution. Thus, the maximum value will be in the exponential family of extreme values. If similar computations are performed for the other 3 Iris species of Fishers article, the resultant 4 univariate extreme distance measures can be used to estimate the location and scale parameters of a reduced Extreme Value distribution.

It is tempting to try to invert this problem; given an extreme pollution measurement and a covariance matrix, what is the range of individual pollutant values that could produce the given extreme pollution index? Although it is mathematically possible to compute such ranges, the ranges are highly correlated and at least one range can always take on infinite values. Thus, such calculations should be avoided unless a great deal of additional information is available about limits of such ranges.

11.7 SUMMARY

This chapter has outlined the statistical techniques that can be used for extreme value analysis of multivariate data. All the techniques suggested are derived from multivariate statistical procedures, ranging from classic discriminant analysis to modern cluster analysis algorithms. The common feature of these is to transform the multivariate observations into a univariate quantity which can be analyzed using the extreme value techniques presented in previous chapters.

11.8 REFERENCES

Boullion, T. L., and Odell, P. L., 1971, Generalized Inverse Matrices, Wiley Interscience.

Everitt, B., 1980, Cluster Analysis, John Wiley - Halstead Press.

Fisher, R. A., 1936, 'The Use of Multivariate Measurements in Taxonomic Problems', Annals of Eugenics, Vol. 7, Part 2, pp 179 - 188.

Fisher, R. A., 1938, 'The Statistical Utilization of Multiple Measurements', Annals of Eugenics, Vol. 8, Part 4, pp 376 - 386.

NOTE
Both the Fisher papers are reprinted in:
Fisher, R. A., 1950, Contributions to Mathematical Statistics, John Wiley and Sons.

Galambos, J., 1978, The Asymptotic Theory of Extreme Order Statistics, John Wiley and Sons, (Chapter 5).

CHAPTER 12
THE WEIBULL DISTRIBUTION

12.1 INTRODUCTION

The Weibull distribution has been found experimentally to describe the reliability of mechanical systems such as the minimum breaking strength of steel beams, or the minimum operating time between failures of an assembly line. To a lesser extent, the Weibull has been used to study biological phenomena such as the response to stress. For example, Peto et. al. (1972) describe age-specific cancer induction rates with a Weibull distribution. This distribution is named after Walodi Weibull, who in 1939 derived it in an analysis of breaking strength. It had been derived in 1928 by Fisher and Tippet as the third asymptotic distribution of extreme values.

In this chapter the notation of previous chapters is changed to agree with the literature on the Weibull distribution. This distribution is used to study smallest extremes. The variable measured is typically time to failure or load that causes failure. The measurements are ordered from smallest, x[1], to largest, x[n].

An important distinction between the Extreme Value and Weibull distributions is that the Weibull allows a lower bound below which the probability is zero that an event, such as failure, will occur (this bound can be zero). No matter how much carcinogen an animal is exposed to, there is a minimum time necessary for a tumor to kill the animal. This is distinct from receiving so much carcinogen that the animal is killed by the direct toxicity of the carcinogen itself. Upper bound situations can be analyzed by changing the sign of the data values.

The Weibull distribution has received little attention in environmental and biological research. Morbidity from air pollutants is directly analogous to the mortality from carcinogens reported by Peto. The effects of most pollutants are generally assumed to have a threshold. The lower bound parameter of the Weibull distribution is a measure of such a threshold. The medical use of drugs involves a threshold dose above which toxicity becomes more important than therapeutic effect. The Weibull lower bound can measure the dose at which toxicity is expected in the most sensitive member of a population.

When statistically modeling the effects on a population of exposure to a pollutant, one should consider the confound distribution of an Extreme Value distribution of maximum exposure with a Weibull distribution of response to minimum insult. Confound distributions occur when the response being studied depends upon the sequential or simultaneous actions of two or more statistical processes. In the pollutant example, the response of humans to a fixed and known pollutant exposure is a statistical phenomena describing variability between individuals. The exposure an individual receives is also a statistical phenomena that varies with such things as time of day, weather conditions, and location of persons during the day. Thus the morbidity within the population is a function of two statistical processes, the exposure and the response. This combination of two distributions is called a confound distribution, and mathematical methods exist for deriving a single statistical distribution if the two distribution functions for exposure and effect are known. The people within the population are exposed depending upon both where they are located and the time of day. The subpopulation that receives the maximum exposure is to be considered. The magnitude of this maximum exposure might be described by an Extreme Value distribution and the minimum exposure that will cause a response might be described by a Weibull distribution. The combination of these two statistical distributions can be expressed as the confound distribution of response and exposure. It does not follow that the most sensitive person, nor the person maximally exposed, will show the

maximum response.

Most pollutants have natural or background levels. The actual exposure of humans, animals and plants is the sum of the natural background levels plus what man adds. For example, in addition to the hydrocarbons added to air pollution by man, there are the natural terpines emitted by trees and brush. The study of the effects of man-made pollution should adjust for the background levels of the pollutant in order to study the response in excess of that caused by the background. In such a situation it might be reasonable to allow a negative value of the Weibull lower bound and assume that the portion of the Weibull distribution that falls between the lower bound and zero measures the proportion of the total response that is due to the background levels of the pollutant.

12.2 THE WEIBULL DISTRIBUTION AND DENSITY

The mathematics of the Weibull distribution are presented first as a two parameter function, and then with the additional condition of a threshold.

Assume a variate x in the range $0 \leq x \leq +\infty$ which depends upon two parameters, b and c; b is called the characteristic life parameter, and c is called the shape parameter. The distribution function is

$$F(x) = 1 - \exp(-(x/b)^c) \quad , \tag{12.1}$$

and the density function is

$$f(x) = (cx^{c-1}/b^c)\exp(-(x/b)^c) \quad . \tag{12.2}$$

Let G() signify a Gamma function. The mean of the Weibull density is

$$\text{mean} = b*G((c+1)/c) \quad , \tag{12.3}$$

the variance is

$$v = b^2(G((c+2)/c) - (G((c+1)/c))^2) \quad , \tag{12.4}$$

and the mode is

$$\text{mode} = b(1 - 1/c)^{1/c} \quad \text{for } c>1; \quad = 0 \quad \text{for } c\leq 1 \quad . \tag{12.5}$$

By the method of maximum likelihood, estimates of b and c are the solutions of the simultaneous equations

$$\hat{b} = \left((1/n)*\sum_{i=1}^{n} x_i^{\hat{c}}\right)^{1/\hat{c}} \text{ and} \tag{12.6}$$

$$\hat{c} = n/((1/\hat{b})^{\hat{c}}*\sum_{i=1}^{n} x_i^{\hat{c}}*\ln(x_i) - \sum_{i=1}^{n} \ln(x_i)) \quad .$$

These equations must be solved iteratively.

The Weibull is also related to two other statistical distributions. The Weibull with the shape parameter fixed at a value of unity, c = 1.0, is an Exponential distribution with a mean of b. If the characteristic life parameter is fixed at a value of two, b = 2.0, a Raleigh distribution with parameter c is obtained.

If a threshold parameter, u, is included in the Weibull distribution function, the variate x in equations 12.1 and 12.2 is replaced by x - u. The distribution function is then

$$F(x) = 1 - \exp(-((x-u)/b)^c) \quad , \tag{12.7}$$

and the density function is

$$f(x) = c/b * ((x-u)/b)^{c-1}*\exp(-((x-u)/b)^c) \quad \text{for } x \geq u; \quad =0 \quad \text{for } x < u \tag{12.8}$$

with the restrictions that the parameters b and c are greater than zero.

12.3 PARAMETER ESTIMATION

Substantial literature exists on parameter estimation for the Weibull distribution. A sampling of this will be presented here. The literature falls into three classes: 1) simple estimators, 2) estimation from censored samples, and 3) construction of tolerance and confidence limits. Through the use of the logarithmic relationship between the Weibull and Extreme Value distributions, discussed in Chapter 3, all the statistics outlined in this section may also be applied to estimation for the Extreme Value distribution. In fact, many estimators used for the Weibull are based upon this transformation.

The generalized maximum likelihood procedure discussed in Section 7.3 can be used for the Weibull distribution by substituting either equation 12.2 (two parameter Weibull) or 12.7 (three parameter Weibull) for h(x) in the equations of Section 7.3. This procedure yields all the statistical properties of Maximum Likelihood Estimators, but a general purpose scientific computer capable of executing a functional maximization algorithm is necessary.

12.3.1 Simple Estimators.

The method of moments, using functions of the mean and variance of the data, cannot be used directly for parameter estimation for the Weibull distribution. Equations 12.3 and 12.4 show that this method would require the inverse of a Gamma function be used in simultaneous equations. However, for the <u>two-parameter</u> Weibull, the logarithmic transformation and equations 7.15 and 7.16 yield estimates of the Weibull characteristic life and shape parameters. Thus, the method of moments can be used by analyzing the logarithms of the data as an Extreme Value distribution, then performing an inverse transformation on the parameters:

(12.9) Weibull scale parameter = exp(Extreme Value location parameter)

(12.10) Weibull shape parameter = 1.0/(Extreme Value scale parameter) .

When using equations 12.9 and 12.10, equation 7.15 is changed from a form for the largest extreme to a form appropriate for the smallest extreme. This is accomplished by a change of sign so that equation 7.15 becomes mean = mode - scale parameter * Euler's constant. Equation 7.16 for the standard deviation remains unchanged.

The method of moments is frequently used, but the statistical considerations of bias, efficiency, and sufficiency are not yet fully studied. Since moments are easy to compute with a hand calculator, it is a practical Weibull estimation procedure. This method may also be used for the three parameter Weibull when the location parameter is known. In this case, the known location parameter value is subtracted from all the data values and then analysis proceeds as for a two parameter distribution.

All the techniques of Chapter 9 may be used with the logarithmic transformation of the data and the corresponding inverse transformation of the parameters when the location parameter is known. Gumbel's regression estimators for censored samples, discussed in section 9.2.1, is a simple estimation technique for censored samples with a Weibull distribution.

A refinement is available to compensate for the bias introduced by simple parameter estimators. Engelhardt and Bain (1974), and Engelhardt (1975) discuss such compensation for both complete and censored samples. These papers also contain a good review of the literature, Recent references may be found in Bain and Engelhardt (1981).

12.3.2 Statistical Inference from Censored Weibull Samples.

A variety of computing methods are available for censored samples. The few methods outlined here were chosen because they are available in current statistical journals.

A characteristic of most Weibull parameter estimation techniques for censored samples is that special tables of 'unbiasing factors' are needed. These are given in the literature. Billmann, Antle, and Bain (1972) offer a method and give tables of unbiasing parameters for the two-parameter case. Lemon (1975) gives the corresponding information for the three-parameter Weibull. Cohen (1975) proposes a three parameter technique that does not require special tables.

12.3.3 Confidence and Tolerance Limits for the Weibull Distribution.

Confidence limits are upper and lower bounds determined so that the interval between these limits will include the true value of the parameter with the specified confidence. Sometimes it is desirable to obtain an interval which will cover a fixed portion of the distribution with a specified confidence. Such intervals are called tolerance intervals, and the end points of such intervals are called tolerance limits. These intervals can be applied to either the distribution of parameter estimates, or to the distribution of the data. When used on the data distribution, they are sometimes called prediction intervals.

The methods reviewed in this section depend on simulation results and on the relation between the Weibull and Extreme Value distributions. Lawless (1975) gives a method for estimating quantiles or tolerance bounds for a variable with a Weibull or an Extreme Value distribution. His method is applicable for censored data. Mann and Fertig (1977) give a method using quantiles of the data for estimating confidence bounds of parameter estimates and tolerance or prediction intervals for the measured variable.

Their method is an extension with bias correction of the work of Hassanein discussed in Section 9.3. Fertig, Meyer, and Mann (1980) discuss methods for obtaining a prediction interval with a pre-specified probability of containing a future observation. This paper also uses tables of bias correction factors, and is a good review of the literature up to 1980. Bain and Engelhardt (1981) find good approximations to the distributions of the parameters of the Weibull distribution and use these to construct approximate confidence intervals for the parameter estimates, and tolerance limits on the data values.

12.4 LIFE TESTING

Life testing is a part of statistics called stochastic processes, and is a special case of order statistics that can be used in many of the problems for which the Weibull distribution is used. A full discussion of life testing theory can be found in stochastic processes textbooks (for example, Parzen, 1962, section 4.3). Much of the statistical theory of life testing has been published by Epstein (1953, 1960a, 1960b).

Life testing statistics are not completely analogous to using the Weibull distribution, but many kinds of problems can be analyzed either way. Life testing typically describes the time to failure of a known (typically small) number of items (appliances, machines, death of animals, etc.) subjected to a constant stress. Although life testing is usually used to describe the average failure time of the population from which the items are sampled, it contains all the statistical tools needed to study the first failure; thus it can be used to study the statistical properties of the smallest extreme value. The Weibull distribution is more general in scope; it can be used to describe the failures in a changing environment that cause the first failure, as well as the time to failure under conditions of constant environment. Statistically, it is usually preferable to use the Weibull for changing environmental conditions and life testing for studies of time to failure. For example, suppose a small

town builds a wooden bridge over their stream. If the mayor is interested in estimating how many months before the bridge needs repair (assuming no unusual conditions), he should use life testing statistics. If, instead, he is interested in estimating how big a truck can use the bridge before it falls, he should use Weibull statistics.

In the simplest form of life testing failures are regarded as events having a Poisson distribution with mean time to failure of 1/g. It then follows that the time to the first failure in a group of n items has an Exponential distribution with a mean of g/n. The times of successive failures are independent and Exponentially distributed. If T is the observed time of the first failure, then 2nT/g is Chi-square distributed with 2 degrees of freedom. An unbiased estimate of g is nT. From these assertions, a 100a% confidence interval for the mean life g may be stated after the first failure in a group of size n is observed to be

$$\frac{2nT}{X^2(2,a/2)} \leq g \leq \frac{2nT}{X^2(2,1-a/2)} \qquad (12.11)$$

where $X^2(2,a/2)$ is the 100a%/2 value found in a table of the Chi-square distribution with 2 degrees of freedom. For a 95% confidence interval on g, with a=0.05, $X^2(2,0.025)=7.378$, and $X^2(2,0.975)=0.0506$.

For example, suppose a manufacturer places 10 units of a new kind of ozone measurement device around a city, and observes the first breakdown after 5.3 weeks of operation. At that time, his best estimate of the mean time to failure of each of this type of unit is 53 weeks, with a 95% confidence interval of 14 to 2094 weeks (37 years). This is not a very useful confidence interval, but one should not expect that some other approach would give better results. The manufacturer also wishes to estimate how many repair men he needs to maintain 100 units. Using the fact that the time between failures is exponentially distributed with mean of g/100 (assuming each unit is repaired as soon as it fails so that the

sample size, n, remains constant), the manufacturer can expect an average of just over two failures per week, and with 95% confidence successive failures will occur between 0.048 week (just over 8 hours) and 6.96 weeks (2.5% and 97.5% tolerance limits of an Exponential distribution with a mean of 53/100). These limits are obtained by solving the Exponential density function, $F(x) = 1 - \exp(-t/m)$, for t when F(x) is set to the desired probability limit and the mean, m, is known.

12.5 SUMMARY

This chapter presents an introduction to the Weibull or Fisher Type 3 Extreme Value distribution. This distribution is used in the study of reliability and in materials failure studies. The density and distribution functions are presented along with formulas for several estimable statistics. The Weibull distribution allows the option of including a third parameter, in addition to scale and shape parameters, which represents a threshold below which the probability of an effect or a measured response is zero. Simple parameter estimators are given, and it is noted that such estimators usually depend upon the logarithmic relationship between the Weibull and Extreme Value distributions. The literature on parameter estimation is reviewed and papers are cited that propose unbiased and efficient estimators for parameter values, confidence intervals, and tolerance limits, and that can be used with censored samples. Finally, the use of life testing statistics for extreme value problems is discussed.

12.6 REFERENCES

Bain, L. J., and Engelhardt, M., 1981, 'Simple Approximate Distribution Results for Confidence and Tolerance Limits for the Weibull Distribution Based on Maximum Likelihood Estimators', Technometrics, Vol. 23, No. 1, pp 15 - 20.

Billmann, B. R., Antle, C. E., and Bain, L. J., 1972, 'Statistical Inference from Censored Weibull Samples', Technometrics, Vol. 14, No. 4, pp 831 - 840.

Cohen, A. C., 1975, 'Multi-censored Sampling in the Three Parameter Weibull Distribution', Technometrics, Vol. 17, No. 3, pp 347 - 351.

Engelhardt, M., and Bain, L. J., 1974, 'Some Results on Point Estimation for the Two-Parameter Weibull or Extreme-Value Distribution', Technometrics, Vol. 16, No. 1, pp 49 - 56.

Engelhardt, M., 1975, 'On Simple Estimation of the Parameters of the Weibull or Extreme Value Distribution', Technometrics, Vol. 17, No. 3, pp 369 - 374.

Epstein, B., and Sobel, M., 1953, 'Life Testing', J. Amer. Statistical Assoc., Vol. 48, pp 486 - 502.

Epstein, B., 1960a, 'Statistical Life Test Acceptance Procedures',

Technometrics, Vol. 2, No. 4, pp 435 - 446.

Epstein, B., 1960b, 'Estimation From Life Test Data', Technometrics, Vol. 2, No. 4, pp 447 - 454.

Fertig, K. W., Meyer, M. E., and Mann, N. R., 1980, 'On Constructing Prediction Intervals for Samples From a Weibull or Extreme Value Distribution', Technometrics, Vol. 22, No. 4, pp 567 - 573.

Lawless, J. F., 1975, 'Construction of Tolerance Bounds for the Extreme Value and Weibull Distributions', Technometrics, Vol. 17, No. 2, pp 255 - 261.

Lemon, G. H., 1975, 'Maximum Likelihood Estimation for the Three Parameter Weibull Distribution Based on Censored Samples', Technometrics, Vol. 17, No. 2, pp 247 - 254.

Mann, N. R., and Fertig, K. W., 1977, 'Efficient Unbiased Quantile Estimators for Moderate-Size Complete Samples from Extreme Value and Weibull Distributions; Confidence Bounds and Tolerance Prediction Intervals', Technometrics, Vol. 19, No. 1, pp 87 - 93.

Parzen, E., 1962, Stochastic Processes, Holden-Day.

Peto, R., Lee, P. N., and Page, W. S., 1972, 'Statistical Analysis of the Bioassay of Continuous Carcinogens', Br. J. Cancer, Vol. 26, pp 258 - 261.

CHAPTER 13
MISCELLANEOUS TOPICS

13.1 INTRODUCTION

This chapter contains a variety of unrelated topics that do not logically fit into other chapters and whose discussions are too short to constitute separate chapters.

13.2 RECORD TIMES

A sequence of record times is obtained by sequentially examining a data list, or sequentially collecting data, and extracting a sublist of records. A record is the maxima (or minima) of the data so far examined or collected. Previous records are not discarded when a new record is found, thus the sublist consist of a sequence of increasingly better records and as the sublist gets longer additions to it become less frequent. The key element of this concept is that the size of the groups of data values between record values is not fixed, but rather is a sequence of increasing random variables. This variable group size results in the distribution of record values being approximately Gaussian rather than one of the extreme value distributions. Extremes are derived from fixed sample sizes, records are derived from a sequence of increasingly larger and variable sample sizes.

Let A[n] and B[n] be the location and scale parameters respectively for examinimg n data values, and let N(n) be the number of records extracted from the n data values. Designate the N record values as R[1], R[2],...R[N], and for any arbitrary value of n designate the overall record as R[N(n)]. One might expect that the quantity

$$(R[N(n)] - A[n])/B[n]$$

would converge in probability to a reduced Extreme Value distribution. This quantity actually converges to a distribution closely related to the Gaussian.

Thus, if a data set of n values is sequentially examined and N records, R, are extracted, the statistical distribution law of the R's will look Gaussian rather than like an extreme value distribution. This is because the R's are not the maxima (or minima) of equal sized subsamples, and also the R's are sequentially correlated. A rigerous mathematical development of this concept may be found in Galambos (1978, Sections 6.3 and 6.4).

13.3 MIXTURES OF EXTREME VALUE DISTRIBUTIONS

Sometimes a data set is composed of samples from two or more populations mixed together. This situation can be caused by such things as a change in measurement conditions, or by collecting data from a nonhomogeneous population. When data is collected over a long time period, such as meterological or air pollution data, the location of measurement stations can change and instruments are often upgraded, introducing bias and a change of variance. The data set of all weights of new employees of a company is a mixture of two biological phenomena because there are both male and female employees.

In many data analysis situations the mixture is caused by much less obvious conditions than the examples of the previous paragraph, and data is not collected on some auxiliary variables necessary to identify the components of the mixture. When such auxiliary variables are available or can be found in other records, the data can be separated into subsets for statistical analysis or the auxiliary variable can be used as a covariate.

Often it is possible to recognize a mixture on a probability plot, where mixtures appear as a segmentation of the data into clusters or segments of lines. Once a mixture is recognized, or suspected, it is possible to define a mixed distribution function and use maximum likelihood methods (Section 7.3) to find estimates for the parameters of the component distributions of the mixture.

A mixed density function is defined as a weighted sum of component densities where the weights describe the proportion of each density in the total.

$$(13.1) \qquad f_M(x) = p_1 f_1(\theta_1,x) + p_2 f_2(\theta_2,x) + \ldots . p_n f_n(\theta_n,x)$$

where x = the data variable,
$f_M(\ )$ = the mixed density function,
$f_I(\ )$ = the Ith component density,
θ_I = parameters of the Ith component,
p_I = the mixing weight,
Restriction: $p_1 + p_2 + \ldots . p_n = 1.0$.

The mixing weight is the probability that a data value comes from the Ith density. For data values $x[j]$, $j = 1,2,\ldots J$, maximum likelihood parameter estimates are found by maximizing the function L defined in Equation 13.2.

$$(13.2) \qquad L = \sum_{j=1}^{J} f_M(x[j]) \quad .$$

The use of a generalized functional minimization algorithm, such as Simplex described in Section 7.3, can be used to find the parameter values that maximize L or l, the logarithm of L.

Changery (1982) describes a mixture of extreme value distributions for wind speed. He identifies storms of two types, tropical and nontropical. The annual extreme wind speeds of tropical storms appear to follow a

Weibull distribution, and the extreme wind speeds of nontropical storms appear to follow an Extreme Value distribution. A mixture of tropical and nontropical storms is typical for weather stations located in Florida and along the gulf coast. A list of annual maximum wind speeds from such stations is a sample from a mixture of extreme value distributions. Changery actually identified tropical storms by reviewing historical weather maps for the days on which the annual maxima occurred (that is, he used auxillary variables). He then separated the data into two subsets, for each subset he used appropriate parameter estimation techniques, then used a mixed distribution function to compute wind speeds versus return period. In this discussion the more general likelihood function approach will be used.

Table 13.1 gives 30 years of annual extreme wind speeds for Jacksonville, Florida, the storm type is that determined by Changery. The wind speeds are corrected to be miles per hour at 10 meters height.

Table 13.1
Extreme wind speeds, Jacksonville, FL
T indicates a tropical storm.

YEAR	MPH	TYPE	YEAR	MPH	TYPE
1950	65	T	1965	52	T
1951	38	T	1966	44	T
1952	51		1967	69	
1953	47		1968	47	T
1954	42		1969	53	
1955	42		1970	40	
1956	44		1971	51	
1957	42		1972	48	
1958	38		1973	53	
1959	34		1974	48	
1960	42	T	1975	68	
1961	44		1976	46	
1962	49		1977	36	T
1963	56		1978	43	
1964	74	T	1979	37	

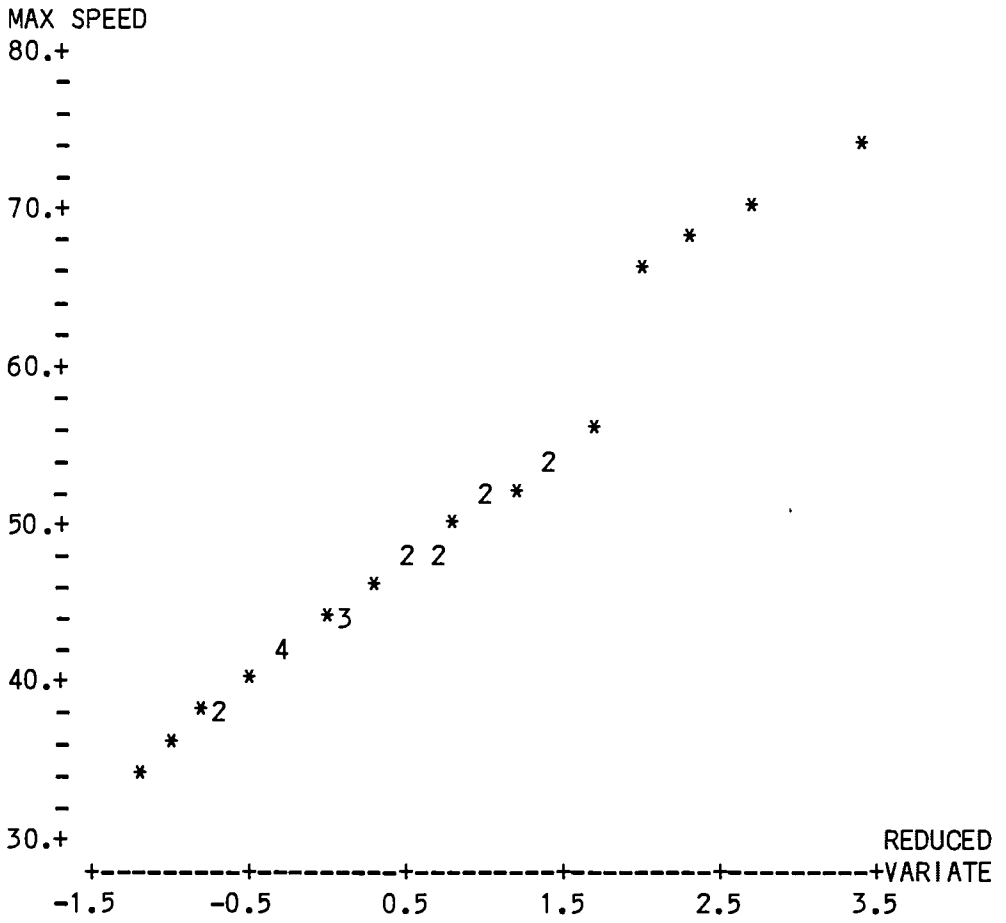


Figure 13.1
Extreme Value Probability Plot
of Jacksonville, FL Data

Figure 13.1 is an extreme value probability plot of the data in Table 13.1. The numbers indicate that more than one data value falls on the same plotting position. This figure clearly shows the segmented line characteristic of a mixed distribution, the four largest values are displaced to higher wind speeds than would be obtained by extrapolating from the 26 lower values. A comparison of Figure 13.1 and Table 13.1 shows that the tropical storms identified by Changery are not the second population suggested by the figure, in fact the tropical storm data is rather evenly mixed among the nontropical storm data. Thus, there appears

to be more than one basis for separating populations from this storm data. In order to illustrate the use of the maximum likelihood estimation technique using a mixed density function, the remainder of this discussion will ignore the classification given in Table 13.1 and instead use the population structure suggested in Figure 13.1.

An examination of Figure 13.1 suggest that the two segments are well represented by straight lines. This in turn suggest that both segments are samples from different Extreme Value distributions. Additional plotting, not shown here, using logarithms indicated that a Weibull or Cauchy distribution are not good choices for either segment. Admittedly, four data points are scant information for such a choice of distributions, but that's all the information there is in the upper data segment.

The Extreme Value density function is given in Equation 7.9. Using the definition of a mixed density given in Equation 13.1, the density of a mixture of two Extreme Value distributions is the five parameter density given in Equation 13.2.

$$(13.2) \quad f_M(x[i]) = p*h_1(x[i]) + (1-p)*h_2(x[i])$$

$$h_1(x[i]) = \exp(-y_1[i] - \exp(-y_1[i]))/a_1$$

$$y_1[i] = (x[i] - u_1)/a_1$$

$$h_2(x[i]) = \exp(-y_2[i] - \exp(-y_2[i]))/a_2$$

$$y_2[i] = (x[i] - u_2)/a_2$$

The five parameters are the two modes, the two scale parameters, and the probability that a data value is a member of the first population.

In order to use a generalized iterative function maximization algorithm to find the maximum likelihood estimates of the parameters of Equation 13.2, initial estimates of all parameter values are required. An

initial estimate of p is simply the proportion of the data points that appear to be in the lower population in Figure 13.1: $p = 26/30 = 0.867$. Initial estimates for the modes and scale parameters can be obtained from regressions on subsets of the data, or by the method of moments applied to the subsets. The method of moments defined in Equations 7.15a and 7.16a is used here. The 4 data points in the upper segment have a mean of 69.00 and a standard deviation of 3.74, yielding an estimated mode of 67.32 and scale parameter of 2.92. The data points of the lower segment have a mean of 44.89 and a standard deviation of 5.81. The corresponding estimate of the mode is 42.27 and the estimate of the scale parameter is 4.53.

The SIMPLEX algorithm was used to maximize the logarithm of the mixed density defined in Equation 13.2 using the data in Table 13.1, and the initial estimates described in the previous paragraph. The results are given in Table 13.2.

TABLE 13.2
Maximum Likelihood Estimates using a Mixed Extreme Value Density and Jacksonville, FL Data

Parameter	Estimate	Standard Error
p	0.885	0.092
u_1	42.29	1.70
u_2	67.63	2.37
a_1	5.59	1.32
a_2	2.62	1.64

The same technique was used to find the maximum likelihood parameter estimates using a single Extreme Value density function. The results are given in Table 13.3.

TABLE 13.3
Maximum Likelihood Estimates Using a Single Density,
Jacksonville, FL data

Parameter	Estimate	Standard Error
u	43.74	1.962
a	7.230	1.495

These two sets of parameter estimates can be evaluated by comparing the empirical distribution of the data to the distribution models on probability plots. For this, the densities h(x) in Equation 13.2 are replaced by the corresponding distributions H(x), and the equation is solved for each data value x[i]. The results for the mixed distribution are shown in Figure 13.2. This figure suggest a good fit to the data was achieved.

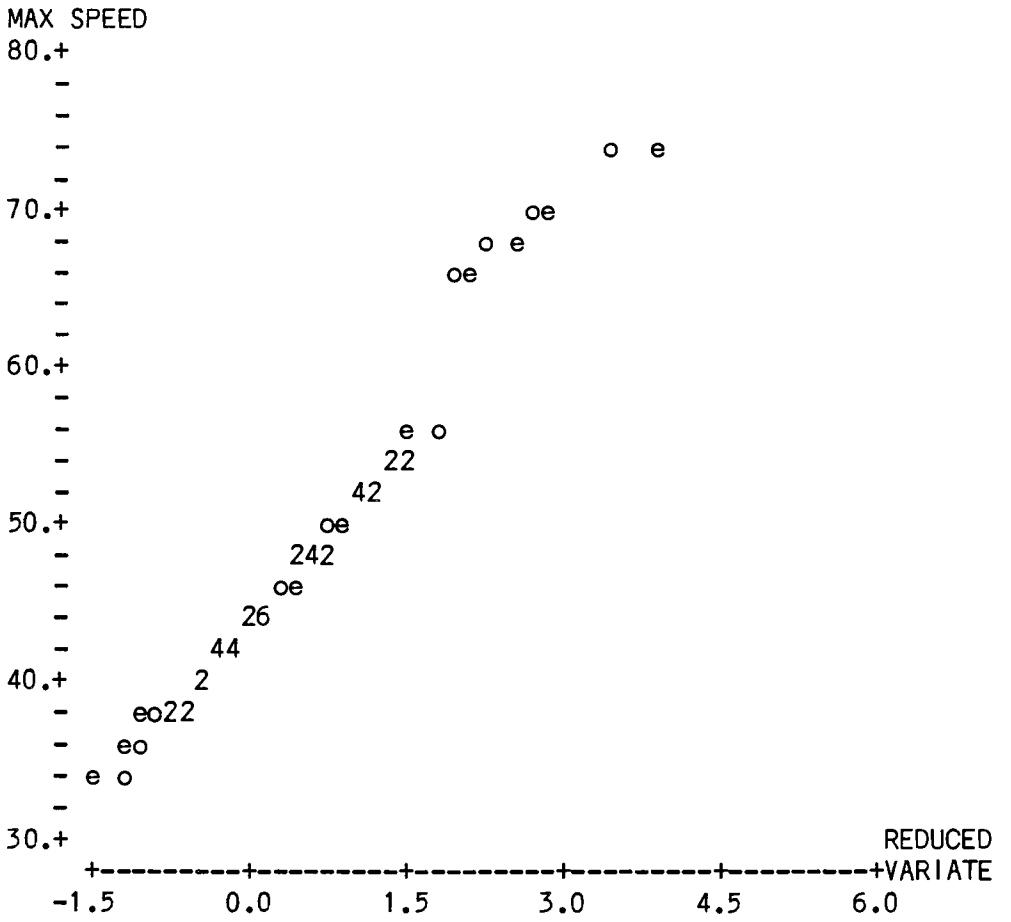


FIGURE 13.2
Observed (o) and Expected (e) Probabilities
Extreme Value Probability Plot
Mixed Density Model

The corresponding plot for the single density model is given in Figure 13.3. A visual comparison of Figures 13.2 and 13.3 shows that a single density does not model the four highest wind speeds as well as the mixed density model does.

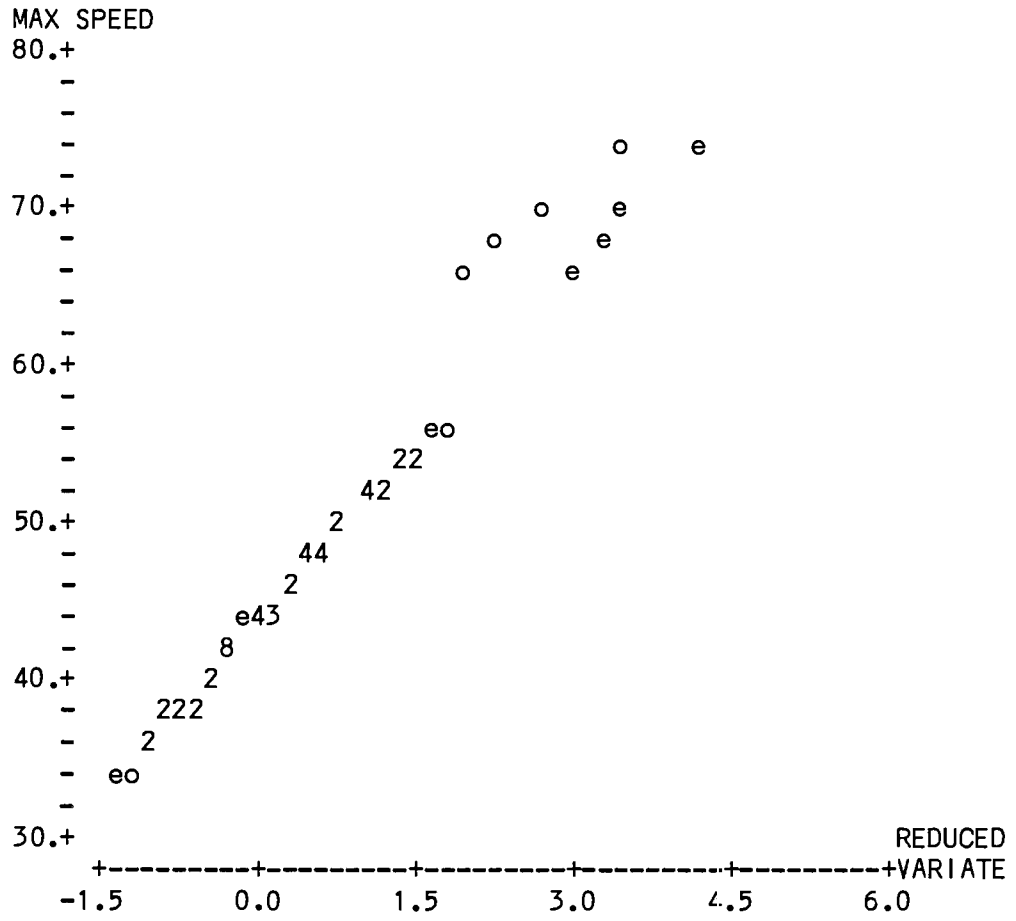


FIGURE 13.3
Observed (o) and Expected (e) Probabilities
Extreme Value Probability Plot
Single Density Model

Consideration of the return periods shows the consequencees of choosing the wrong density model. Return periods are defined and discussed in Section 7.2, Equation 7.8 is used except that H(y) is replaced with the single or mixed distribution model using the appropriate maximum likelihood parameter estimates. The return period using the single density model can be solved analytically, however the mixed censity model requires an iterative solution to find the value of x corresponding to a given value of T(x). The results are listed in Table 13.3.

TABLE 13.3
Return Period Versus Wind Speed, Jacksonville, FL

Prob = probability of the given wind speed occurring in any single year (1-F(x))

T(x) = return period (years)

$x^{(1)}$ = single density model wind speed (MPH)

$x^{(2)}$ = mixture of densities model wind speed

Prob	T(x)	$x^{(1)}$	$x^{(2)}$
0.50	2	46	45
0.20	5	55	55
0.10	10	60	67
0.05	20	65	70
0.04	25	70	70
0.02	50	72	73
0.01	100	77	75
.005	200	82	77
.002	500	89	80
.001	1000	94	83

Table 13.3 shows that the expected wind speeds do not differ much between the single density and the mixed density models for return periods less than 100 years. For return periods of 100 years or more, the single density model predicts higher wind speeds than does the mixed density model. The displacement of the upper segment in Figure 13.2 to higher wind speeds suggest that the mixed model should yield higher speeds than the simple model for the return periods over 100 years. The scale parameter of the upper segment of the mixture density model is less than half the scale parameter of the single density model so that extrapolation beyond the figure yields lower wind speeds for the mixed model.

The correlation coefficient goodness of fit test, discussed in Appendix 7-B is based upon a theoretical consideration of a single population distribution. There is no reason to assume it is applicable to a mixture of distributions situation.

13.4 BIOASSAY AND EXTREME VALUES

The Extreme Value distribution can be used as a dose-response model in bioassay. This use is not an extreme value situation because all data is used rather than just the extremes, The shape of the Extreme Value distribution is used as another possible bioassay model. A better terminology is 'the double exponential bioassay model'. The most commonly used bioassay models are the probit (Gaussian) and logistic distributions. Any of the special features of the Extreme Value distribution that result from using extreme data values, such as extrapolation to larger sample sizes (Section 7.4.3), are not applicable to bioassay.

In bioassay work a plot or regression is made of a nonlinear function of a biological response versus dose or a function of dose (such as the logarithm of dose). If the right functions are chosen, the resulting plot will be a straight line. Statistical inferences and hypothesis test can then be derived from the line. Suppose groups of animals are exposed to a sequence of increasing concentrations of a pollutant. Let d[i] represent the dose for the ith group, and let p[i] represent the proportion of subjects or animals responding in the ith group (p[i] = number if responders in ith group/total exposed in group). The response can be any yes-no type measurement such as sick or well, dead or alive, active or inactive and so on. The data points (d[i],p[i]) are transformed into values (x[i],y[i]) where

$$y[i] = \ln(-\ln(1 - p[i])) ,$$

$$x[i] = f(d[i]) .$$

A plot of the x's and y's is a dose-response curve for the double exponential bioassay model. This situation differs from 'ordinary' bioassay only in the substitution of the inverse of the Extreme Value distribution for the inverse of the Gaussian or logistic distributions in the computation of the y's.

The simulated data given in Table 13.4 show the computations used in a typical bioassay analysis. The natural logarithms are used as the data transformation to derive the x's. Suppose 5 groups of laboratory animals are exposed to increasing doses of an air pollutant in an environmental chamber for one hour each day and the weight of each animal at the beginning and end of the experiment is recorded. A response is defined to be a loss of weight during the experiment. The proportion of animals responding is given in the p column of Table 13.4, the dose d is the concentration of pollutant in parts per million in the chamber. For comparison with 'ordinary' bioassay, the normal deviates are included in the INVNORM column, these are calculated as the inverse Gaussian distribution function of the p's.

TABLE 13.4
Bioassay Data

i	d[i]	p[i]	ln(d)	y[i]	INVNORM
1	7.	0.05	2.	-3.	-1.7
2	20.	0.13	3.	-2.	-1.1
3	55.	0.31	4.	-1.	-0.5
4	148.	0.63	5.	0.	0.3
5	403.	0.93	6.	1.	1.5

Figure 13.4 plots the raw data in two ways, the proportion responding versus dose and versus the natural logarithm of the dose.

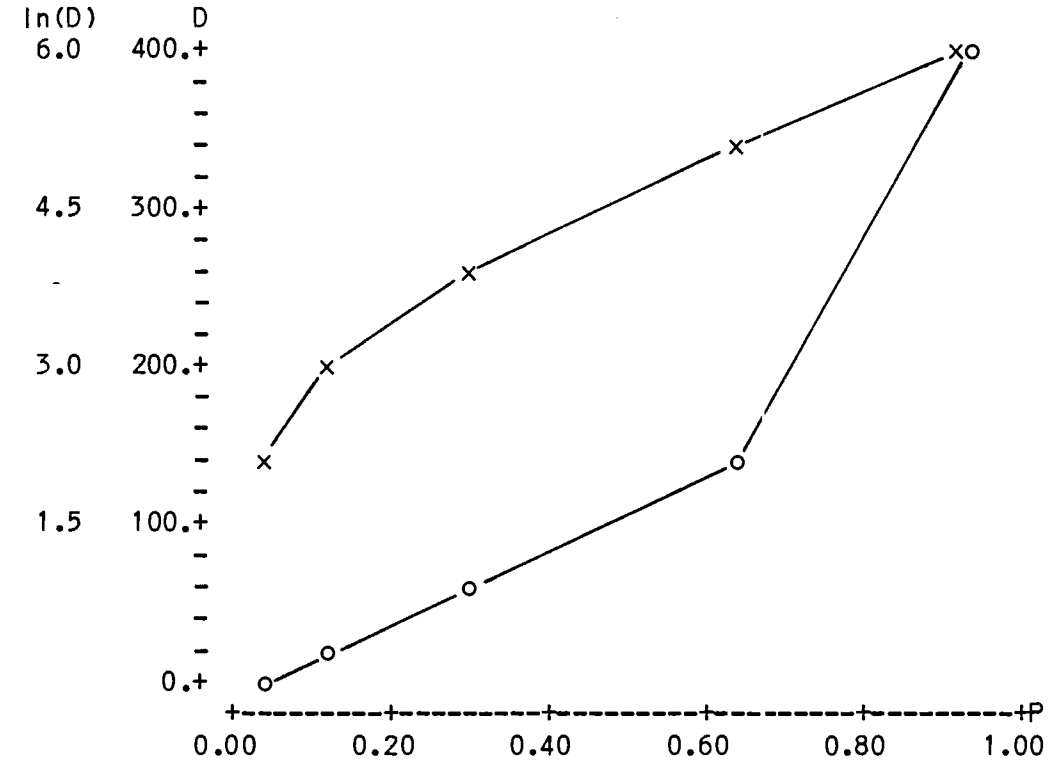


FIGURE 13.4
Log Dose (x) versus Proportion Responding and
Dose (o) versus proportion Responding

Neither of these curves are close to a straight line, suggesting that a transformation of the response should be considered. Figure 13.5 plots the inverse Gaussian transformation of the proportion of responders versus dose and natural logarithm of the dose.

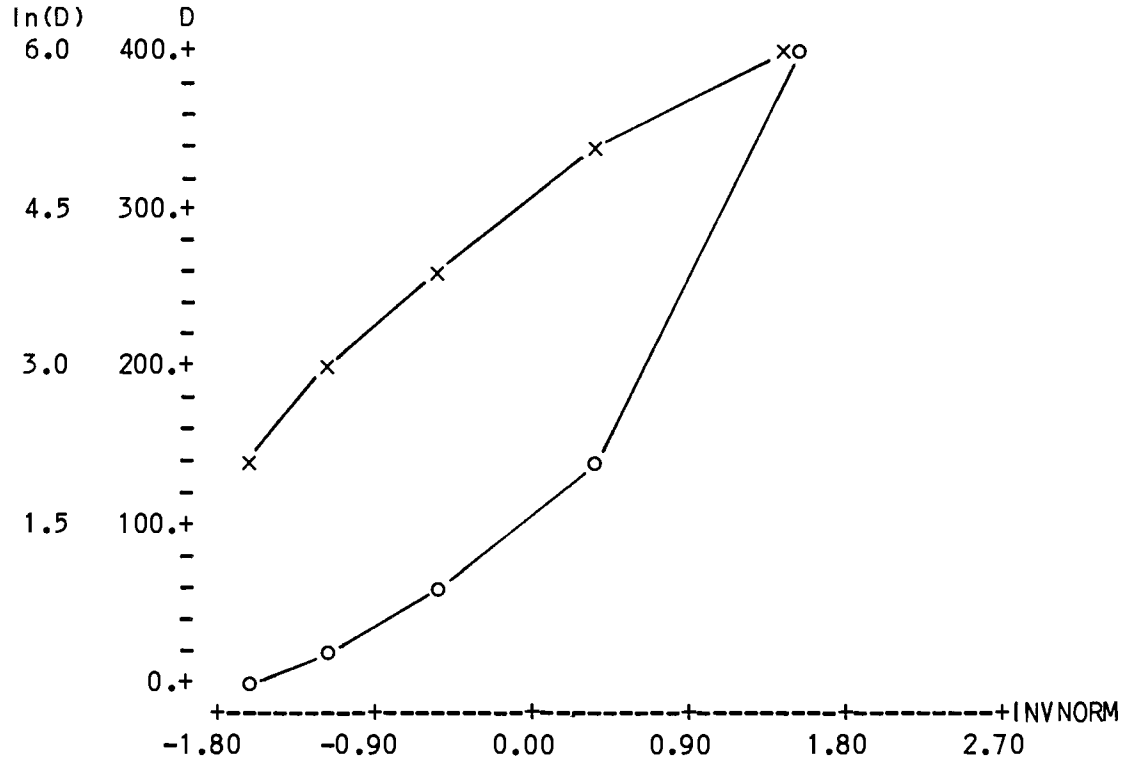


FIGURE 13.5
Bioassay Plot Using Gaussian Distribution
x = ln(Dose),
o = Dose

The dose-response plots of Figure 13.5 also are not close to a straight line. The upward curve on an inverse Gaussian plot suggest a double exponential model. Figure 13.6 is the corresponding plot for the double exponential model, that is a double logarithmic transformation of the response.

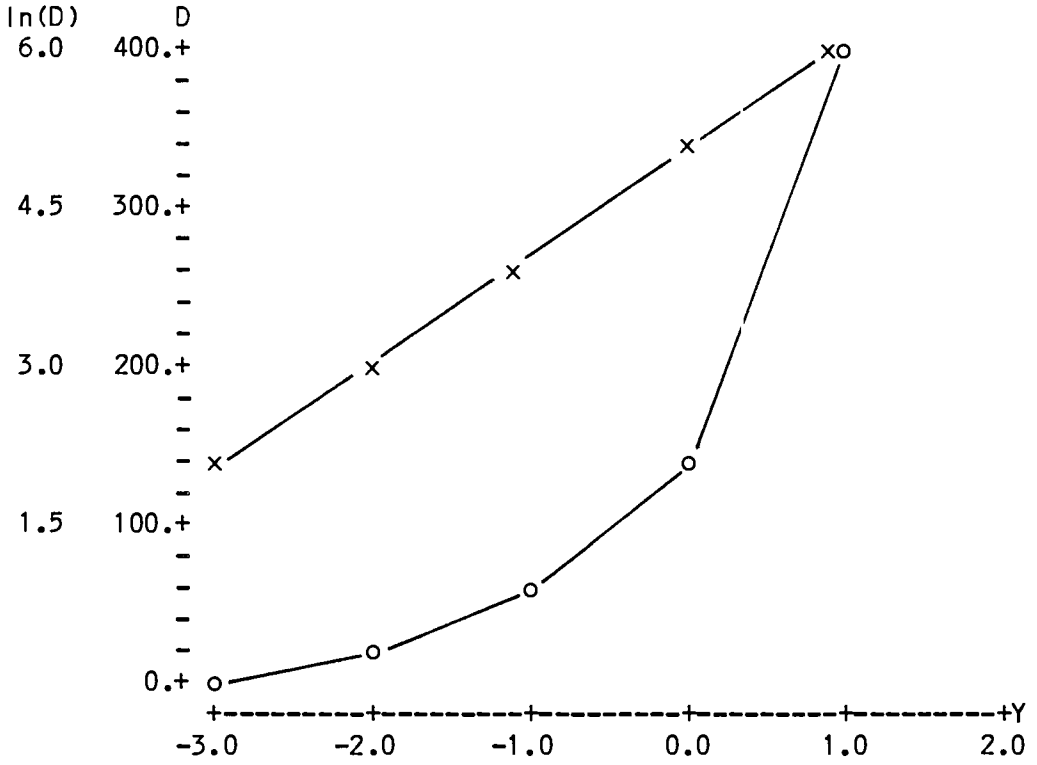


FIGURE 13.6
Bioassay Plot Using Double Exponential Model
x = ln(Dose),
o = Dose

The double exponential response model along with a logarithmic transformation of the dose describes the data very well, and the other combinations of response and dose transformations do not result in a reasonably straight line on the bioassay plots.

Fitting data to bioassay lines, comparison of models, and other statistical uses of bioassay models is not a simple regression problem. A review of a bioassay text, such as Finney (1971), indicates that an iteratively reweighted regression is required to get unbiased estimates of regression parameters. The weights are a function of the expected value of

the proportion responding, using the observed proportion yields biased estimates. The likelihood function approach to parameter estimation (Section 7.3) can also be used with bioassay models. The likelihood function approach, which is mathematically equivalent to the iteratively reweighted regression approach, is presented only in recent bioassay text because the necessary computer hardware and software was not available in the 1930's and 1940's when the theory of bioassay was being developed.

13.5 REFERENCES

Changery, M. J., 1982, 'Historical Extreme Winds for the United States - Atlantic and Gulf of Mexico Coastlines', Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission, NUREG/CR-2639.

Finney, D. J., 1971, Probit Analysis (3rd Ed.), Cambridge University Press.

Galambos, J., 1978, The Asymptotic Theory of Extreme Order Statistics, John Wiley and Sons.

INDEX

PNL-4690
UC-32

DISTRIBUTION

No. of
Copies

OFFSITE

Dr. Robert L. Watters
Office of Health and
Environmental Research
Washington, DC 20545

27 DOE Technical Information Center

ONSITE

DOE Richland Operations Office

H. E. Ransom

31 Pacific Northwest Laboratory

DB Carr
RO Gilbert
DL Hall
RR Kinnison (18)
JA Mahaffey
WL Nicholson
AR Olsen
Publishing Coordination (2)
Technical Information (5)