Title

TRUNCATED DISTRIBUTIONS IN HYDROLOGIC ANALYSIS1

Permalink
https://escholarship.org/uc/item/0730k5kx

Journal
JAWRA Journal of the American Water Resources Association, 28(5)

ISSN
1093-474X

Authors
Loaiciga, Hugo A
Michaelsen, Joel
Hudak, Paul F

Publication Date
1992-10-01

DOI
10.1111/j.1752-1688.1992.tb03187.x

Peer reviewed

WATER RESOURCES BULLETIN
VOL. 28, NO. 5 AMERICAN WATER RESOURCES ASSOCIATION OCTOBER 1992

TRUNCATED DISTRIBUTIONS IN HYDROLOGIC ANALYSIS[1]

Hugo A. Loaiciga, Joel Michaelsen, and Paul F. Hudak[2]

ABSTRACT: Truncated samples arise in a variety of hydrologic situations in which certain values of the variables of interest are unobservable. Remotely sensed data, for example, are truncated below the resolution level of the sensor: all objects smaller than the resolution of the instrument are not detected and their characteristics cannot be recorded. Many other situations occur in hydrologic studies where a sampling procedure or method yields truncated samples. The main results on this work on truncated samples are: (1) a general expression for the probability distribution function of a truncated random variable; (2) a description of the effect of truncation on the distribution function of several important probability models and on their parameters; and (3) development of a parameter estimation methodology for parameter estimation from truncated samples.

From the theoretical results of this paper and the analysis of drought-related data, we have concluded that: (1) truncated sampling can be advantageous, especially when certain ranges of values are difficult or impossible to observe; (2) the developed method for truncated sample analysis leads to efficient and accurate parameter estimation and statistical inference on truncated data; and (3) the developed method for truncated sample analysis can also yield the parameters and the distribution of the entire population when only a subset of that population is observable.

(KEY TERMS: streamflow; truncation; sampling; water resources planning.)

INTRODUCTION

Censoring, Truncation, and Partial Series

There are many instances in which hydrologists work with subsets of complete data sets. When indexed by time, the complete data set is referred to as a complete series, whereas any subset of the complete series is called a partial series. Well known examples of partial series are those obtained from hydrologic records (rainfall, runoff) whereby only those values exceeding a certain threshold are kept in the data set, regardless of their time of occurrence. In general, however, the use of subsets of data in hydrology goes beyond the realm of time series analysis and includes spatial data or other data generated by a variety of sampling schemes. Analytical data on water quality, in which there exist minimum detection levels below which cannot be recorded, are classical examples of partial data sets (U.S. EPA, 1989). Remote sensing data, where the resolution of sampling instruments restricts observable phenomena to have a minimum size (Dubayah *et al.*,1990), is another example of an incomplete or partial data set. All of the cited examples of partial data sets present a common characteristic, namely, there is a censoring mechanism, intentional or not, by which certain observations are eliminated from the total sample.

In the statistical literature, partial data sets are known by different names, including truncated samples (Rao, 1989), censored samples (David, 1981; Miller, 1981; Kendall and Stuart, 1979), or incomplete samples (Rao, 1989). Strictly speaking, censoring purposely restricts data recording to some interval, whereas truncation is such that if the random variable of interest falls outside some interval, even its existence is unobserved. For example, if a microscope cannot detect bacteria below a certain diameter, the size distribution of the sampled bacteria becomes truncated due to this sampling limitation. On the other hand, a study of the lower tail of the distribution of streamflows might require using only those order statistics that fall below a certain threshold in a sample (Loaiciga and Marino, 1988). This would represent a case of censoring.

[1] Paper No. 92062 of the *Water Resources Bulletin*. **Discussions are open until June 1, 1993.**

[2] Respectively, Associate Professor, Department of Geography and Environmental Studies Program, University of California, Santa Barbara, CA 93106; Associate Professor and Chairman, Department of Geography, University of California, Santa Barbara, California 93106; and Assistant Professor, Department of Geography, University of North Texas, Denton, Texas 76203.

Censoring and truncated data have found a large number of applications across scientific disciplines, including hydrology. Censored data analysis in hydrologic studies has mainly focused on population parameter estimation based on a subset of the total data set (see, e.g., Gilliam and Helsel, 1986; Helsel and Cohn, 1988; Loaiciga and Marino, 1988).

There are three main types of censoring methods, the so-called types I, II, and III (Miller, 1981). In Type I, right-censoring, the recorded observations are equal to the variables of interest if the observations are less than (or equal to) a specified threshold. When the observations exceed the threshold, their recorded values are set equal to the threshold. Type I, left-censoring, occurs whenever the censored observations are those that fall below the threshold (see Equation 7 below). An example of Type I (right-) censoring would be the measurement of the times that it takes to complete a batch of experiments. If the completion times are less than a threshold, t_c, then they are recorded as measured. Otherwise, the times to completion may be considered too long and censored at the threshold t_c. The recorded sample of observed completion times and censored times (equal to t_c) represents a right-censored sample.

In Type II censoring, the censored sample is composed of either the first r order statistics of the total sample or, alternatively, of the r largest order statistics of the total sample. Loaiciga and Marino (1988) provided an example of Type II censoring, in which a set of order statistics was used to characterize the shape of the lower tail of streamflow distributions. Type III censoring is characterized by multiple, random, censoring thresholds. In the case of Type III right-censoring, the recorded observations are equal to the variables of interest if the observations are less than the thresholds. Otherwise, the recorded observations are set equal to the values taken by the (random) thresholds. For Type III left-censoring, the recorded observations equal the variables of interest only when they exceed the thresholds.

The methods presented in this work concern data sets obtained by truncation. In this situation, the variables of interest are observable, and recorded, only when their values are within a fixed interval. For example, if a sensor has a spatial resolution of ten meters, the distribution of the size of samples objects becomes truncated below the size of ten meters. If the distribution of the number of children in families is ascertained by surveying students in elementary schools, the distribution of the number of children becomes truncated below the size of one, since all families in the sample would have at least one child. Although the last example might be thought of as being the result of a faulty sampling plan, it may be advantageous in some cases to sample subsets of a population in consideration to limited time and resources. Thereafter, one can utilize statistical corrections on the truncated sample to make unbiased inferences on the sampled population as a whole. In the remainder of this paper, we will use the term truncated distributions rather than partial series to characterize data obtained by a truncation process. It has been discussed previously that the term partial series might encompass censored as well as truncated data sets.

The Distribution of Truncated Random Variables

Let us consider first the case of interval truncation. Suppose that a random variable X is truncated to the interval $(\theta,+\infty)$. Therefore, values below the truncation level θ are not observable. Concentrations of chemicals restricted by analytical precision to be measurable only above a minimum detection level is an example of interval truncation. Let us denote the truncated random variable by X_T. Notice then that the truncated random variable obeys the following probability law (P denotes probability):

$$P(X_T \geq x) = P(X \geq x \mid X \geq \theta) \tag{1}$$

where $x \geq \theta$ by definition of the truncated random variable. The conditional probability on the right-hand side of Equation (1) can be expanded to yield:

$$P(X_T \geq x) = P(X \geq x;\, X \geq \theta \;/\; P(X \geq \theta) \tag{2}$$

and, since $x \geq \theta$, Equation (2) is further simplified:

$$P(X_T \geq x) = P(X \geq x) \;/\; P(X \geq \theta) \tag{3}$$

The simplification on the right-hand side of Equation (3) follows from the fact that the event "$X \geq x$" is a subset of the event "$X \geq \theta$". Therefore, by the axioms of probability theory, the joint probability of these two events equals the probability of the "smaller" event. From Equation (3) the probability density function (pdf) of the truncated random variable becomes (the prime denotes differentiation with respect to x):

$$\begin{aligned} f_{X_T}(x) &= P'(X_T \geq x) \\ &= -\,P'(X_T \geq x) \\ &= P'(X \leq x) \;/\; P(X \geq \theta) \\ &= f_X(x) \;/\; P(X \geq \theta), \qquad x \geq \theta \end{aligned} \tag{4}$$

Equation (4) shows that the pdf of the truncated random variable is a scaled form of the pdf of the original variable, f_X. It is straightforward to show that the pdf of a right-truncated random variable (as opposed to the left-truncated case of Equation 4) is given by:

$$f_{X_T}(x) = f_X(x) \,/\, P(X \le \theta), \qquad x \le \theta \tag{5}$$

From either Equations (4) or (5) it is seen that the truncated pdf is positive and that

$$\int_R f_{X_T}(x)dx = 1 \tag{6}$$

where R denotes the restricted range of the truncated random variable. Therefore, the pdf of the truncated random variable, as expressed in either Equations (4) or (5), satisfies the conditions of a well-defined probability density function.

It is now possible to elucidate the fundamental difference between a censored variable, X_c, and a truncated variable. In the left-censoring, Type I, case, the censored variable is defined as follows:

$$X_c = \begin{cases} X \text{ if } X \ge \theta \\ \theta \text{ if } X < \theta \end{cases} \tag{7}$$

The censored random variable is characterized by the following probability distributions:

$$P(X_c \le x) = F_X(x), \qquad x \ge \theta \tag{8}$$

$$P(X_c \le x) = 0, \qquad x < \theta \tag{9}$$

where

$$F_X(x) = P(X \le x) \tag{10}$$

is the cumulative distribution function of the original random variable X. Equations (8) and (9) indicate that the left-censored random variable in Equation (7) has a discontinuous distribution function with a discontinuity equal to $F_X(\theta)$ at $x = \theta$. This is illustrated in Figure 1. The cumulative distribution function of the left-truncated variable follows immediately from Equation (4), and is shown in Figure 2. It is seen in Figure 2 that the distribution function of the (left-) truncated variable is continuous everywhere and is a scaled form of the cumulative distribution function of the random variable X. Incidentally, the expected values of the left-censored and left-truncated random variables can be shown to be:

$$E(X_C) = \int_\theta^\infty x f_x(x)dx + \theta F_X(\theta)$$

$$= \int_\theta^\infty [1 - F_x(x)]dx + \theta \tag{11}$$

$$E(X_T) = \int_\theta^\infty x f_{X_T}(x)dx$$

$$= \int_\theta^\infty [1 - F_X(x)]dx \,/\, P(X \ge \theta) + \theta \tag{12}$$

Equations (11) and (12) show the effect of censoring and truncation on the first moment of a probability distribution, respectively. For example, considering the exponential distribution (to be discussed in detail below), the truncated expected value is $(1/\lambda)+\theta$, compared with the unrestricted expected value of $1/\lambda$, where λ is the parameter of the exponential distribution. The censored expected value, on the other hand, is given by $(e^{-\lambda\theta}/\lambda)+\theta$. It is seen then, that although somewhat similar insofar as restricting the range of recorded observations, censoring and truncation are fundamentally different statistically, and they produce distinct derived distributions.

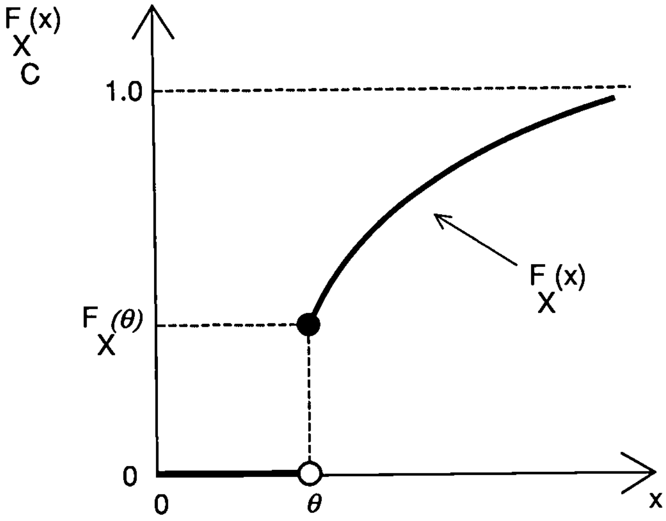

Figure 1. Distribution Function of a Left-Censored Random Variable.

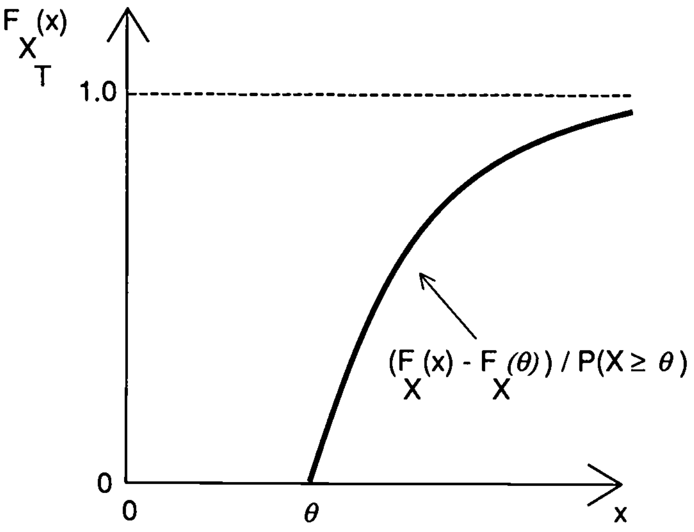

Figure 2. Distribution Function of a Left-Truncated Random Variable.

The distribution functions and expected values of left-censored and left-truncated random variables can be modified to describe the case of right-censoring or right-truncation without difficulty. In fact, the probability density function of a truncated random variable in the arbitrary set R (i.e., $X_T \varepsilon R$) is given by:

$$f_{X_T}(x) = f_X(x) / P(X \varepsilon R), \qquad x \varepsilon R \qquad (13)$$

Equation (13) indicates that truncation results in a scaling of the probability density function of the original random variable, and the scaling factor is equal to the inverse of the probability that the original variable is in the set R. Equation (13) is a generalization of the results previously obtained for interval-truncated random variables in Equations (4) and (5).

PARAMETER ESTIMATION AND DISTRIBUTION FITTING FOR TRUNCATED RANDOM VARIABLES

Truncated samples in hydrologic studies include partial series where only data that exceed certain threshold are recorded, or where the resolution of a sampling instrument restricts the range of the recorded observations. In some instances, limitations on budget, time, and geographic accessibility force a sampling scheme to truncate the random variable of interest. There is, therefore, fertile ground in hydrology for the application of the statistical methods for truncated sample analysis presented in this section.

Suppose that there are n observations of a truncated random variable, $x_1, x_2, \ldots, x_n$. Two questions arise: (1) is it possible to ascertain that these variables are indeed generated by a truncation process, and (2) can we infer the parameters of the parent population (i.e., described by an unrestricted random variable) from the truncated sample? The first question is of practical interest since, in many cases, either by experimental design or some other (accidental or intentional) mechanism, a sample may be truncated and it is necessary to answer statistical questions about the truncated random variable. Such questions could include quantile and moment estimation, or hypothesis testing on the truncated variable, amongst others. The issue of parameter estimation is also relevant. Recovering the parameters of the parent population from a truncated sample can lead to a full characterization of the parent population when it is only possible to observe the truncated random variable.

It has been shown in the previous section that truncation shifts and scales the truncated distribution function only, without altering its overall shape (see Figure 2). Therefore a truncated, skewed, lognormal distribution will remain skewed and retain the overall shape of the original distribution except for some scaling and shifting. Similarly, a truncated exponential distribution will retain its exponential decay, *albeit* with a different rate and a translation of its mean. This same pattern of shape preservation holds true for symmetric distributions such as in the normal (or Gaussian) case. Shape preservation is helpful in identifying the proper distribution model for a truncated random variable. By the simple plotting of data and basic descriptive analysis (e.g., histogram analysis) one can postulate a distribution model, fit it and then conduct a goodness-of-fit test for model verification. The preservation of the shape of the distribution function under truncation can be disadvantageous, however. When the existence of truncation is unknown to the data analyst, it can lead to fitting the wrong (i.e., unrestricted) probability model to the data, failing to recognize the underlying mechanism that generates them.

Besides the difficulties brought about by the specification of a suitable probability model when analyzing truncated data, truncation can introduce changes in the parameters that govern the truncated distribution relative to those parameters in the original distribution. Several probability models will be examined below to illustrate the effects of truncation on the resulting (truncated) distribution, and to show the modification of distribution parameters under truncation. The distribution models discussed below find frequent application in hydrologic and water resources planning studies.

The Uniform Distribution

This is one of the fundamental distributions in statistics (see Loaiciga, 1988, for an application in hydrological studies). Suppose that X is a random variable uniformly distributed in the interval (a,b). Its probability density function is:

$$f_X(x \mid a, b) = \frac{1}{b - a}. \qquad x \varepsilon (a, b) \qquad (14)$$

If observations of X are restricted to the interval $x \leq \theta$. then the truncated probability density function becomes:

$$f_{X_T}(x \mid a, \theta) = \frac{1}{\theta - a}, \qquad x \varepsilon (a, \theta) \qquad (15)$$

Equation (15) shows that the uniform distribution is preserved under (right) truncation, although one of its parameters, the upper bound of the distribution, is modified by truncation.

The Pareto Distribution

This distribution is useful in economic modeling, and, in particular, in water-resources economics (James and Lee, 1971). Assume that the random variable X is Pareto distributed with parameters α and γ. Its probability is given by the following expression:

$$f_X(x \mid \alpha, \gamma) = \gamma \alpha^{\gamma} x^{-(\gamma + 1)}, \qquad x \geq \alpha \qquad (16)$$

Suppose that the random variable X is truncated to the interval $X \geq \theta$. Then, the probability density function of the truncated random variable X_T becomes:

$$f_{X_T}(x \mid \theta, \gamma) = \gamma \theta^{\gamma} x^{-(\gamma + 1)}, \qquad x \geq \theta \qquad (17)$$

Therefore, the Pareto density function is preserved under (left) truncation. The parameter α is modified by the truncation process.

The Exponential Distribution

This distribution appears naturally in the study of waiting times and recurrent phenomena. The study of the inter-arrival time between such phenomena as floods or droughts relies in many cases on exponential models (Loaiciga *et al.*, 1992). Let X be an exponentially distributed random variable with the following probability density function (λ is the density's parameter):

$$f_X(x \mid \lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0 \qquad (18)$$

If the random variable X is truncated to the interval $X \geq \theta$, its truncated probability density becomes:

$$f_{X_T}(x \mid \lambda, \theta) = \lambda e^{-\alpha(x - \theta)}, \qquad x \geq \theta \qquad (19)$$

The exponential density is preserved, along with its parameter λ, although the argument in the exponential is decreased by an amount equal to the truncation threshold.

The Geometric Distribution

This is the analog of the exponential distribution for discrete random variables. The geometric distribution is better known in hydrology by its role in modeling the recurrence of annual events, such as streamflows (Loaiciga and Marino, 1991). Suppose that the random variable X has a geometric distribution with parameter p:

$$P(X = x \mid p) = (1 - p) p^{x - 1}, \qquad x = 1, 2, \ldots \qquad (20)$$

then, its left-truncated probability distribution ($X > \theta$) is given by the following expression:

$$P(X_T = x \mid p, \theta) = (1 - p) p^{x - \theta - 1}, \quad x = \theta + 1, \theta + 2, \ldots \qquad (21)$$

The truncated random variable still has a geometric distribution. Its range, however, is restricted, and the exponent in the distribution function is decreased by the amount θ.

The Gumbel Distribution

Introduced by Gumbel (1958), this distribution can be used to model extreme events, such as rainfall or floods. If X denotes a Gumbel-distributed random variable with parameters c and d, its density function is given by:

$$f_X(x \mid c, d) = cd \exp(-dx) \exp[-c \exp(-dx)], \quad -\infty < x < \infty \qquad (22)$$

then its right-truncated probability density function becomes (where $X \leq \theta$).

$$f_{X_T}(x \mid c, d, \theta) = cd \exp(-dx) \exp\{-c[\exp(-dx) - \exp(-d\theta)]\} \qquad -\infty < x \leq \theta \tag{23}$$

Unlike the previous distributions, a truncated Gumbel random variable does not have an exact Gumbel probability density function as shown in Equation (23).

Other important distributions in hydrology, such as the lognormal distribution, show a pattern similar to that exhibited by the Gumbel distribution under truncation: their random variables are restricted in their ranges and their truncated distributions do not conform exactly to the original distribution models. Except for a few (but important) distributions (e.g., the uniform, Pareto, exponential, geometric, and Gumbel), it is not possible, in general, to derive explicit, closed form, expressions for truncated distributions in terms of the original parameters and truncation thresholds. Numerical integration is required in some cases. Equation (13), however, provides a general rule for deriving probability density functions of truncated random variables.

Maximum Likelihood Estimation with Truncated Samples

For a truncated sample of n observations $x_1, x_2, \ldots x_n$, the likelihood function, L, follows from Equation (13):

$$L = \left[\pi_{i=1}^{n} f_X(x_i \mid \phi) \right] P(X \varepsilon R \mid \phi)^{-n} \tag{24}$$

In Equation (24), ϕ denotes the set of parameters governing the distribution of the unrestricted random variable X. It is also assumed in Equation (24) that the truncated random variable is restricted to the set R. Equation (24) can be readily modified to account for the special truncation cases involving left or right interval truncation (see Equations 4 and 5, respectively). In the method of maximum likelihood, one must find the parameters (ϕ) that maximize expression (24). In general, the maximization of (24) must be done numerically, although some distribution functions permit analytical, closed-form, solutions for the maximum likelihood estimators.

It is usually advantageous to work with the log-likelihood function, obtained by taking the logarithm of the likelihood function, when deriving maximum likelihood estimators. From Equation (24), the log-likelihood function for the truncated sample is given by the following expression:

$$\ln L = \sum_{i=1}^{n} \ln f_X(x_i \mid \phi) - n \ln P(X \varepsilon R \mid \phi) \tag{25}$$

The parameter set that maximizes the right-hand side of Equation (25) is the maximum likelihood estimator.

To illustrate the application of Equation (25) let us use the geometric distribution (see Equations 20 and 21). Given a sample of geometric random variables, $x_1, x_2, \ldots, x_n$, the maximum likelihood estimator of its parameter p is given by:

$$p^* = \frac{\overline{x} - 1}{\overline{x}} \tag{26}$$

where

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{27}$$

If the n observations represent a left-truncated sample from a geometric distribution (see Equation 21), then the maximum likelihood estimator of the parameter p is given by:

$$p^*_T = \frac{\overline{x}_T - \theta - 1}{\overline{x}_T - \theta} \tag{28}$$

where $\overline{x}_T$ is the arithmetic mean of the truncated sample. In the next section the geometric model is used to illustrate a method for model validation and parameter estimation from truncated samples.

CASE STUDY

Michaelsen *et al.* (1990), reconstructed total annual streamflow in the South Coast hydrologic area of Southern California from 1460 through 1966. (The South Coast hydrologic area of California comprises the total flow in all coastal drainages between the Ventura River and the U.S.-Mexico border.) The streamflow reconstruction was done by means of tree-ring analysis. The purpose of that study was to determine the characteristics of droughts; i.e., their duration, frequency, and severity, in the South Coast area using long streamflow series (Loaiciga *et al.*, 1990; Loaiciga *et al.,* 1992). Figure 3 shows the reconstructed streamflow time series (median = 747 KAF; mean = 902 KAF; standard deviation = 723 KAF; 1 KAF = 1,000 acre-feet; 1 acre-foot =1,233 m^3), exhibiting a conspicuously large value occurring in 1568 and possibly explained by unusually wet conditions created by the El Nino anomaly (Philander, 1990). Loaiciga *et al.* (1990), showed that, considering time scales of hundreds of years (i.e., 500 years), the South Coast hydrologic area's annual streamflow was stationary with a skewed, lognormal, distribution typical of semi-arid climates subject to extreme variability in annual precipitation. In other words, the distributions of the frequency, severity, and duration of droughts were found to be stable when examined statistically over the time scales considered in the study of Loaiciga *et al.* (1992).

Figure 4 shows a histogram of the duration (in years) of below-median streamflow runs for the South Coast hydrologic area. It is seen in Figure 4 that the distribution of the duration of runs of below-median, annual, streamflow has an exponential decay, suggesting a geometric distribution as a plausible model for the duration of runs. The geometric distribution (Equation 20) is applicable to discrete phenomena such as the duration of below-median streamflow runs. In the remainder of this section, the data in Figure 4 will be examined to determine: (1) whether or not the geometric model is a valid distribution for the duration of below-median runs; and (2) the effect of truncation on distribution fitting and population parameter estimation. The chi-squared test (Pearson, 1914) will be used to test the goodness-of-fit of the geometric model for unrestricted and truncated samples. Parameter estimation will be conducted by means of the maximum likelihood for both types of samples.

Goodness of Fit Test for the Geometric Model

The chi-squared test (see Rao, 1989) is by far the most popular goodness-of-fit test for discrete probability models. It can be used to test whether a sample is generated by a theoretical probability model. In this study we test whether the data in Figure 4 conforms to a geometric model. We examine two cases: (1) the distribution's parameter is known; and (2) the parameter is unknown and must be estimated.

The data in Figure 4 relate to below-median flow. Therefore, the theoretical value of the parameter in the geometric distribution is p = 0.5, and Equation (20) yields:

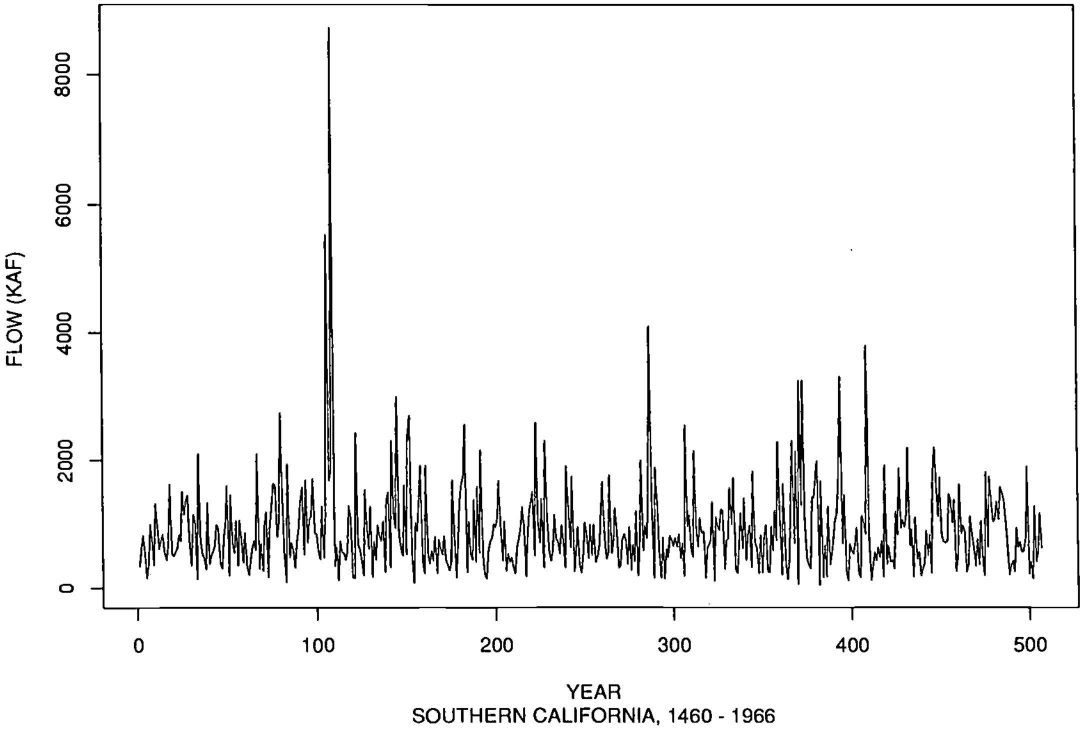

Figure 3. Reconstructed Streamflow Time Series for South Coast.

$$P(X = x) = 0.5^{x}, \qquad x = 1, 2, ... \tag{29}$$

where X represents the duration of below-median runs. On the other hand, one might consider that the true value of the median streamflow is unknown, and that the median streamflow is an estimate from the reconstructed streamflow time series. Under this scenario, the parameter of the geometric distribution, that purportedly generates the data in Figure 4, must be estimated. The maximum likelihood estimate of the parameter p is given by Equation (26). From the histogram in Figure 4, the mean duration of below-median runs is calculated to be $\bar{x} = 2.14$ years, and Equation (26) results in the maximum likelihood estimate of the geometric distribution's parameter being $p^* = 0.53$. Therefore, when its parameter is estimated, the geometric model becomes:

$$P(X = x) = 0.47(0.53)^{x-1}, \qquad x = 1, 2, 3, \cdots \tag{30}$$

Table 1 summarizes the results of the chi-squared test. Notice that at a significance level of 5 percent the geometric model satisfies the goodness-of-fit test for both cases of the distribution's parameter. The geometric model with the estimated parameter (p^* = 0.53) provided the better distribution fit with a P-value of approximately 0.70, against a P-value of about 0.57 for the geometric model with parameter equal to 0.50. Having established a suitable distribution model for the data in Figure 4, we examine next the effect of truncation in distribution fitting and parameter identification.

Model Fitting and Parameter Identification Under Truncation

Suppose that the data in Figure 4 are truncated in such a manner that only those runs of a length exceeding one year are observable, i.e., $X > 1$. This situation could arise, for example, if in the record keeping of droughts only those runs of at least a two-year duration were recorded. A single isolated event of below-median flow would not be counted as a drought. With the truncated record, the problem at hand becomes one of identifying the parameters of the distribution governing the population of droughts, as well as identifying the actual distribution governing the truncated data.

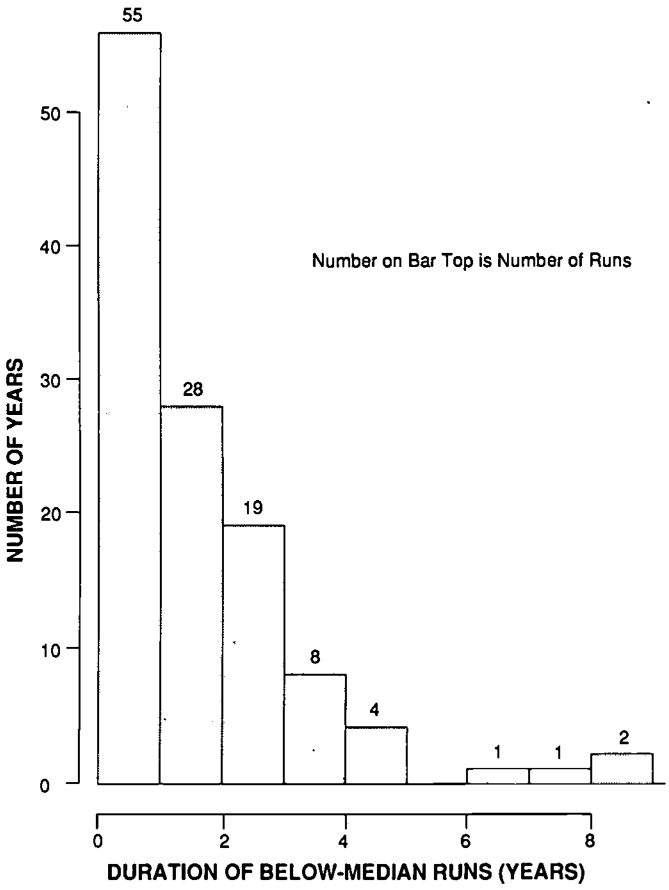

Figure 4. Histogram of the Duration of Below-Median Runs for California's South Coast Hydrologic Area.

The truncated geometric model of Equation (21), with threshold θ = 1, was fitted to the truncated sample of runs (i.e., runs larger than one year in Figure 4). The maximum likelihood estimator of the distribution parameter is $p_T^* = 0.53$, with $\bar{x}_T = 3.14$ and θ = 1 in Equation (28). Notice that the truncated maximum likelihood estimator of the geometric distribution's parameter is identical to the maximum likelihood estimator derived with the unrestricted data set (see Equation 30). Therefore, by accounting for the truncation of runs of length one, it is possible to estimate the parameter of the unrestricted distribution of runs having observed the truncated sample only. The truncated geometric model in Equation (21) becomes:

$$P(X_T = x) = 0.47(0.53)^{x-2}, \qquad x \geq 2 \tag{31}$$

If one attempts to fit the unrestricted geometric distribution in Equation (20) to the truncated sample, the maximum likelihood estimator of Equation (26) yields $p^* = 0.68$ (with $\bar{x} = 3.14$ in Equation (26)), quite

TABLE 1. Results of Goodness-of-Fit Test for Complete Data in Figure 4.

Duration of Below-Median Runs (years) (1)	Observed Number of Occurrences (from Figure 4) (2)		Model: p=0.50 (Equation 29) Expected Number of Occurrences* (3)	Model: p*=0.53 (Equation 30) Expected Number of Occurrences** (4)
1	55		118 (0.500) = 59.0	118 (0.470) = 55.5
2	28		118 (0.250) = 29.5	118 (0.249) = 29.4
3	19		118 (0.125) = 14.8	118 (0.132) = 15.6
4	8		118 (0.0625) = 7.38	118 (0.070) = 8.26
5	4		118 (0.0313) = 3.69	118 (0.0371) = 4.38
6	0		118 (0.0156) = 1.84	118 (0.0197) = 2.32
7	1		118 (0.00781) = 0.922	118 (0.0104) = 1.23
8	1	4***	118 (0.00391) = 0.461	118 (0.00552) = 0.661
9	2		118 (0.00195) = 0.230	118 (0.00293) = 0.346
$\overline{x} = 2.14$	118		117.8	117.7

*The chi-squared statistic is $D = \sum(\theta_i - E_i)^2 / E_i$, where θ_i represents the number of observed occurrences within each interval (column 2) and E_i is the expected number of occurrences within each interval (column 3). D = 6.65 which is less than the critical chi-squared value X^2 (0.05, 8) = 15.5, and the hypothesis of a geometric model is not rejected at a 5 percent significance level.

**The chi-squared statistic is D = 4.56 which is less than the critical chi-squared value X^2 (0.05, 7) = 14.1 and the hypothesis of a geometric model is not rejected at a 5 percent significance level.

***Values are lumped to improve the chi-squared test.

in contrast with the value of $p^* = 0.53$ derived previously (see Equation 30) with the whole sample of runs. The unrestricted geometric model becomes,

$$P(X = x) = 0.32(0.68)^{x-1}, \qquad x = 1, 2, \ldots \tag{32}$$

Table 2 shows the results of the goodness-of-fit test for the models in Equations (31) and (32) using the truncated sample. The truncated model of Equation (31) satisfies the goodness-of-fit test at a 5 percent significance level with a P-value of approximately 0.71, whereas the unrestricted model of Equation (32) did not pass the goodness-of-fit test at the 5 percent significance level, with a P-value of less than 10^{-5}. The results of Table 2 demonstrate that the truncated geometric model is the only feasible model for explaining the truncated data with $X > 1$. Any prediction or inference based on the truncated data must, therefore, be derived from the truncated model. Another important point is that we were able to calculate the parameter of the unrestricted population distribution based on the truncated sample alone ($p^*_T = p^* = 0.53$), and, by establishing the suitability of the truncated geometric model for data in which $X > 1$, we have also identified the unrestricted geometric distribution that generates the entire population of runs. This example illustrates that the methods for truncated sample analysis provide a means to recover the parameters and identify the distribution of the entire population, when one has access to only a subset of the population values.

SUMMARY AND CONCLUSIONS

In this paper we have (1) introduced the concept of truncated samples; (2) developed maximum likelihood estimation from truncated samples; and (3) developed (and applied) a method for probability distribution identification and population parameter estimation from truncated samples. Truncated samples have an apparent similarity to censored samples, the latter being well-known in hydrology and water resources applications. It has been demonstrated, however, that truncated samples are derived from a unique type of sampling process that leads to a specialized statistical inference methodology. Truncated samples arise in a variety of hydrological applications in which certain values of the variables of interest are unobservable. This paper also examined the effect of truncation on several important probability distributions, and, in particular, on the preservation of distribution under truncation as well as the modification of population parameters by that sampling process. Our theoretical results are general in that: (1) they apply to discrete and continuous random variables; and (2) they are applicable to any type of probability distribution. The most notable results of this research are embodied in: (1) Equation (13), that expresses the truncated probability density function; (2) Equations (11) and (12) for

TABLE 2. Results of Goodness-of-Fit Test for Truncated Data.

Duration of Below-Median Runs (years) (1)	Observed Number of Occurrences (from Figure 4) (2)		Model: p_T^*=0.53 (Equation 31) Expected Number of Occurrences* (3)	Model: p^*=0.68 (Equation 32) Expected Number of Occurrences** (4)
1	0		0	63 (0.320) = 20.2
2	28		63 (0.470) = 29.6	63 (0.218) = 13.7
3	19		63 (0.249) = 15.7	63 (0.148) = 9.32
4	8		63 (0.132) = 8.32	63 (0.101) = 6.36
5	4		63 (0.0700) = 4.40	63 (0.0684) = 4.31
6	0		63 (0.0371) = 2.34	63 (0.0465) = 2.93
7	1	4***	63 (0.0197) = 1.24	63 (0.0316) = 1.99
8	1		63 (0.0104) = 0.655	63 (0.0215) = 1.35
9	2		63 (0.00552) = 0.348	63 (0.0146) = 0.920
$\bar{x}_T$ = 3.14	63		62.6	61.1

*The chi-squared statistic is $D = \sum(\theta_i - E_i)^2 / E_i$, where θ_i represents the number of observed occurrences within each interval (column 2) and E_i is the expected number of occurrences within each interval (column 3). D = 4.55 which is less than the critical chi-squared value X^2 (0.05, 7) = 14.1, and the hypothesis of a geometric model is not rejected at a 5 percent significance level.

**The chi-squared statistic is D = 48.6 which is less than the critical chi-squared value X^2 (0.05, 7) = 14.1 and the hypothesis of a geometric model is rejected at a 5 percent significance level.

***Values are lumped to improve the chi-squared test.

the expected values of censored and truncated variables, respectively; (3) Equations (15), (17), (19), and (21) that show the effect of truncation on several important truncated distributions and their parameters; and (4) Equation (25), the log-likelihood function from a truncated sample.

Based on the theoretical results of this paper and on a case study involving a long hydrologic time series we have reached the following conclusions: (1) truncated sampling can be advantageous, especially when certain ranges of values are difficult or impossible to observe; (2) the method for truncated-sample analysis leads to efficient and accurate parameter estimation and statistical inference on truncated data; and (3) the method for truncated-sample analysis can also yield the parameters and the distribution of the entire population given that only a subset of that population is observable. These conclusions have significant value for experimental and sampling plan design. With the methods of this paper it is possible to access a portion of the population, possibly considering issues of cost, time, observability. and the like, and still be able to recover the parameters and distribution of the entire population from the restricted set of observations. The results of this research also revealed the importance of considering the effect that a sampling mechanism might have on statistical inference, and the types of corrections needed to arrive at a theoretically correct method of analysis. Possible new areas of related research, beyond the traditional censored sampling and the truncated sampling scheme developed herein, are situations in which, say, the likelihood of observing certain observations is proportional to the magnitude of the observations. This type of situation also requires special statistical methods of analysis for parameter estimation and probability distribution identification.

ACKNOWLEDGMENTS

This research project was supported by Grants W-752 and W-761 from the California Water Resources Center; Contract B-57105 from the California Department of Water Resources; and by Grant NSF-SES-88-109-17 from the National Science Foundation to the National Center for Geographic Information and Analysis.

LITERATURE CITED

David, H. A., 1981. Order Statistics (2nd Edition). John Wiley, New York, New York.

Dubayah, R., J. Dozier, and F. W. Davis, 1990. Topographic Distribution of Clear-Sky Radiation Over the Konza Prairie, Kansas. Water Resources Research 26(4):679-690.

Gilliam, R. J. and D. R. Helsel, 1986. Estimation of Distributional Parameters for Censored Trace Level Water Quality Data. 1. Estimation Techniques. Water Resources Research 22(2):135-146.

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

Helsel, D. R. and T. A. Cohn, 1988. Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. Water Resources Research 24(12):1997-2004.

James, L. D. and R. R. Lee, 1971. Economics of Water Resources Planning. McGraw-Hill Book Co., New York, New York.

Kendall, M. and A. Stuart, 1979. The Advanced Theory of Statistics. Oxford University Press, New York, New York.

Loaiciga, H. A. and M. A. Marino, 1988. Fitting Minima of Flows Via Maximum Likelihood. Journal of Water Resources Planning and Management. Am. Soc. Civ. Engrs. 114(1):78-90.

Loaiciga, H. A., 1988. On the Use of Chance-Constraints in Reservoir Design and Operation Modeling. Water Resources Research 24(11):1969-1975.

Loaiciga, H. A., J. Michaelsen, and L. Haston, 1990. Probability of Simultaneous Droughts in California and Colorado River Basins: A Study for Water Supply/Demand Analysis. EOS, Transactions American Geophysical Union 71(43):1309.

Loaiciga, H. A. and M. A. Marino, 1991. On the Recurrence Interval of Geophysical Events. Journal of Water Resources Planning Management. Am. Soc. Civil Engrs. 117(2):260-272.

Loaiciga, H. A., J. Michaelsen, S. Garver, L. Haston, and R. B. Leipnik, 1992. Droughts in River Basins of the United States. Geophysical Research Letters 19(20):2051-2054.

Michaelsen, J., H. A. Loaiciga, L. Haston, and S. Garver, 1990. Estimating Drought Probabilities in California Using Tree Rings. Report Contract B-57105 California Department of Water Resources, Department of Geography, University of California, Santa Barbara, California.

Miller, R. G., 1981. Survival Analysis. John Wiley and Sons, New York, New York.

Pearson, K., 1914. On the Probability That Two Independent Distributions of Frequency are Really Samples of the Same Population, With Special Reference to Recent Work on the Identity of Trypanosome Strains. Biometrika 10:85-154.

Philander, S. G., 1990. El Nino, la Nina, and the Southern Oscillation. Academic Press, Inc., San Diego, California.

Rao, C. R., 1989. Statistics and Truth. International Co-operative Publishing House, Burtonsville, Maryland.

U.S. EPA, 1989. Statistical Analysis of Ground Water Monitoring Data at RCRA Facilities. Washington, D.C.