WATER RESOURCES RESEARCH, VOL. 29, NO. 2, PAGES 531–534, FEBRUARY 1993

Testing for Nonzero Skew in Maximum Discharge Runoff Data

ROBERT WHITLEY

Department of Mathematics, University of California, Irvine

T. V. HROMADKA II

Boyle Engineering, Newport Beach, California

Water Resource Council Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) recommends finding T-year flood values by fitting stream gauge data with a log Pearson III distribution. A problem arising in using this method is that, as is well known, the estimator for the skew of this distribution is highly variable and the computed T-year values are sensitive to the value of skew used. The variability in estimating the skew is studied in this note by computing operating characteristic curves which display the probability that a sample from a site with a nonzero skew will actually fall in a specific interval. In rough summary it is not usually possible to reject, at a high level of significance the hypothesis that the skew at a site is zero by using only the data available for that site.

INTRODUCTION

Water Resource Council Bulletin 17B [*Interagency Advisory Committee on Water Data*, 1982] recommends the use of a log Pearson III distribution, fit to yearly maximum discharge data, for the prediction of T-year events. Other methods have been proposed (see, for example, the discussion by *Cohon et al.* [1988]), and an important area of research is in obtaining more accurate methods for estimating extreme floods. However, to quote *McCuen* [1979, p. 269]

> While accuracy of flood estimates is very important, consistency is also an important consideration. Consistency requires a uniform estimation procedure, and the U.S. Water Resources Council issued Bulletin 17 and a revision Bulletin 17a for use in defining flood damage potential at sites where a stream gage was located.

Because of the authority of the U.S. Water Resource Council, the procedure developed in Bulletin 17B is used extensively.

Estimates for the skew coefficient are required when using the log Pearson III distribution, and it has been known for a long time that such estimates are quite variable. There are several procedures suggested in Bulletin 17B to reduce this variability, which include the use of generalized skew coefficients from a map, estimates using regression equations, weighted estimates, and estimates using "nearby similar sites," in addition to just using the station skew. Since all these improved estimators for skew combine individual site skews, although sometimes indirectly as when a map of regional skews is constructed, the variability of the skew at a single site remains an important factor in the accuracy with which the station skew is known. For example, when estimates from several sites are combined to form an estimate for station skew, the underlying assumption is that these sites all have the same skew and part of the evidence for that assumption is the individual site skew estimates themselves.

Paper number 92WR02538.
0043-1397/93/92WR-02538$02.00

Historically, the variability of the skew was at first overestimated; *Slade* [1936] seemed to imply that a series of at least 140 terms was necessary when estimating skew. The point of *Matalas and Benson* [1968] was to indicate this was not the case. In this paper a formula was given for the standard deviation of the distribution of the usual estimator for the skew, under the assumption that the skew was actually zero. Even though the exact distribution of the estimator for skew remained unknown, the value of its standard deviation allowed one to crudely estimate the variation among sampled skews from a distribution with zero skew. These estimates could be also used to get a rough idea of the possible variation involved when estimating a nonzero skew.

In the 24 years since the publication of *Matalas and Benson*'s [1968] article, numerous papers have appeared which deal with the problem of estimating the parameters in a Log Pearson III distribution [see *Arora and Singh*, 1989; *Bobee*, 1973; *Bobee and Robitaille*, 1975, 1977; *Chowdhury and Stedinger*, 1991; *Hardison*, 1974; *Hoshi and Burges*, 1981; *Hu*, 1987; *Kite*, 1975; *Lall and Beard*, 1982; *Lettenmaier and Burges*, 1980; *McCuen*, 1979; *Nozdryn-Plotnicki and Watt*, 1979; *Phien and Hsu*, 1985; *Stedinger*, 1980, 1983; *Tasker*, 1978; *Tasker and Stedinger*, 1986; *Tung and Mays*, 1981; *Wallis et al.*, 1974; *Whitley and Hromadka*, 1986*a*, 1986*b*; 1987, and references therein]. One way in which progress has been made is that, even though the exact distribution for the estimator for the skew remains intractable, the ability to simulate distributions on a computer has increased dramatically.

Graphs of the distribution function for the estimator for the skew, obtained by simulation, are given by *Wallis et al.* [1974]. Such empirical distribution functions are also obtained in this note where they are used in computing intervals for testing the hypothesis of zero skew. Even though the distribution of the estimator for the skew can be simulated, the distribution obtained depends upon the value of the skew of the distribution from which the sampling occurs. Thus it seems that obtaining a test interval for the estimator requires a knowledge of the unknown value of the skew which is exactly the parameter one is trying to estimate. In this note a standard statistical technique is applied to this problem:

compute an interval for the skew estimator, for testing for zero skew, and then compute the operating characteristic φ (φ is one minus the "power" of the test) of using this interval as a test. By looking at the curves so constructed one can see with what confidence one can distinguish between different values of skew.

The example presented is the case of testing to see whether the skew is zero, but any nonzero value of skew could be handled in the same way. The case of zero skew is most important for two reasons. One, the "average" United States skew seems to be approximately zero: the data from the 2972 gauging stations, each having a record length of 25 years or more, used to construct the skew map of Bulletin 17B "... suggest that without additional information, such as the map estimate of skew or a computed station skew, the best estimate of skew would be −0.013" [*McCuen*, 1979, p. 271]. Two, Appendix 14 of Bulletin 17B summarizes research showing that among the distributions tested, the log Pearson III and the lognormal, which is the log Pearson III with zero skew, were the two best predictors of large floods among the group of distributions tested and both were equally effective.

The Estimator for Skew

Water Resource Council Bulletin 17B recommends the use of a log Pearson III distribution, fit to yearly maximum discharge data, for the prediction of T-year events. The logarithm of the yearly peak discharge then has a density function of the form:

$$f(x) = [1/|a|\Gamma(b)][(x - c)/a]^{b-1} \exp - [(x - c)/a] \qquad (1)$$

where, in the case of positive a, the density is given by the expression (1) for $x > c$ and is zero for $x < c$, while in the case of negative a the density is given by (1) for $x < c$ and is zero for $x > c$. Computing the mean μ, standard deviation σ, and skew γ from equation (1) shows that

$$\sigma^2 = a^2 b, \quad \gamma^2 = 4/b, \quad \mu = c + ab \qquad (2)$$

where a has the same sign as γ.

In the case of zero skew, which is the limiting case where the positive parameter b tends to infinity to obtain $\gamma = 0$, it can be shown that the density in equation (1) tends to the density for the normal distribution. In this way the lognormal distribution is the special case of the log Pearson III distribution in which the skew is zero.

It is further recommended in Bulletin 17B that the parameters a, b, and c be estimated by using equation (2) and the usual moment estimators for μ, σ, and γ, with the moment estimator for γ scaled to make it less biased. Specifically,

$$\hat{\mu} = \sum_{i=1}^{n} x_i/n$$

$$\hat{\sigma} = \{n/n - 1\}^{1/2}\left[\sum_{i=1}^{n} x_i^2/n - \hat{\mu}^2\right]^{1/2}$$

$$\hat{\gamma} = \{[n(n-1)]^{1/2}/(n-2)\}\left[\sum_{i=1}^{n} x_i^3/n - 3\hat{\mu}\hat{\sigma} - \hat{\mu}^3\right] \Big/ \hat{\sigma}^3 \qquad (3)$$

These recommendations raise statistical problems which remain open (for a brief summary, see *Chowdhury and Stedinger* [1991, and references therein]). For example, the accurate computation of confidence intervals for the T-year flood has been solved by *Stedinger* [1983] and *Whitley and Hromadka* [1986*a*, 1986*b*, 1987] in the case where μ and σ are estimated but γ is known. The case where μ, σ, and γ are all estimated by equation (3) has been considered, but not completely settled, by *Chowdhury and Stedinger* [1991] and R. Whitley and T. Hromadka (Confidence intervals for T-year floods with estimated skew coefficient, in preparation, 1992). If the skew is zero and therefore the logarithms of the data normally distributed, the computation of confidence intervals is an easy application of the known noncentral t distribution [e.g., *Stedinger*, 1980].

The situation with confidence intervals is typical of the problems raised by the use of the log Pearson III distribution. As a general rule, most statistical calculations are easy in the case of zero skew, and complicated, sometimes impossibly so, in the case of nonzero skew. Since the statistical analysis of the maximal discharges at a site is more difficult in the case of nonzero skew, there is a practical reason to test the hypothesis that the site skew is zero. If it turns out that the assumption of zero skew is consistent with the data, then the use of zero skew allows a much deeper statistical analysis of the data.

Operating Characteristics

The usual way to test whether the unknown skew parameter is zero, is to choose a level of significance $1 - p$, $0 < p < 1$, and compute a number t_p so that the probability that $\hat{\gamma}$ is, in absolute value, less than t_p, given that $\gamma = 0$, is p. In symbols

$$P(|\hat{\gamma}| < t_p | \gamma = 0) = p \qquad (4)$$

Note that the value t_p depends on the number m of points in the sample size and on the estimator (3) used to estimate γ. For example, in order to construct the curves of Figures 1 and 2, it was necessary to compute, by a simulation, the values $t_p = 0.77$ for $p = 0.90$, $m = 20$ points; $t_p = 0.59$ for $p = 0.80$, $m = 20$ points; $t_p = 0.53$ for $p = 0.90$, $m = 50$ points; and $t_p = 0.40$, for $p = 0.80$, $m = 50$ points. For a specific instance, if $\hat{\gamma}$ is the estimator given by (3) for a 50-point sample from a log Pearson III distribution with zero skew, then if repeated samples of size 50 were taken from the distribution, the computed skew estimates would in the long run fall in the interval [−0.53, +0.53] 90% of the time.

A good way to understand the discriminating ability of this test is to compute its operating characteristic φ, which is the function defined by *Breiman* [1973, p. 131]:

$$\varphi(x) = P(|\hat{\gamma}| < t_p | \gamma = x) \qquad (5)$$

For a sample size of m points, for each possible true but unknown value x of the skew, this curve gives the probability $\varphi(x)$ that the sampled estimate $\hat{\gamma}$ of the skew will fall into the interval $(-t_p, t_p)$ and therefore that we will mistakenly not reject the hypothesis that the skew is zero.

If X is a random variable representing the logarithm of a log Pearson III random variable with skew γ, then $-X$ is the logarithm of a log Pearson III random variable with skew $-\gamma$; likewise, replacing x_i by $-x_i$ in equation (3) changes the sign of $\hat{\gamma}$. Thus the operating characteristic satisfies $\varphi(-x) = \varphi(x)$ and therefore only needs to be calculated for $x \geq 0$.

Operating characteristic curves are given below for 90%

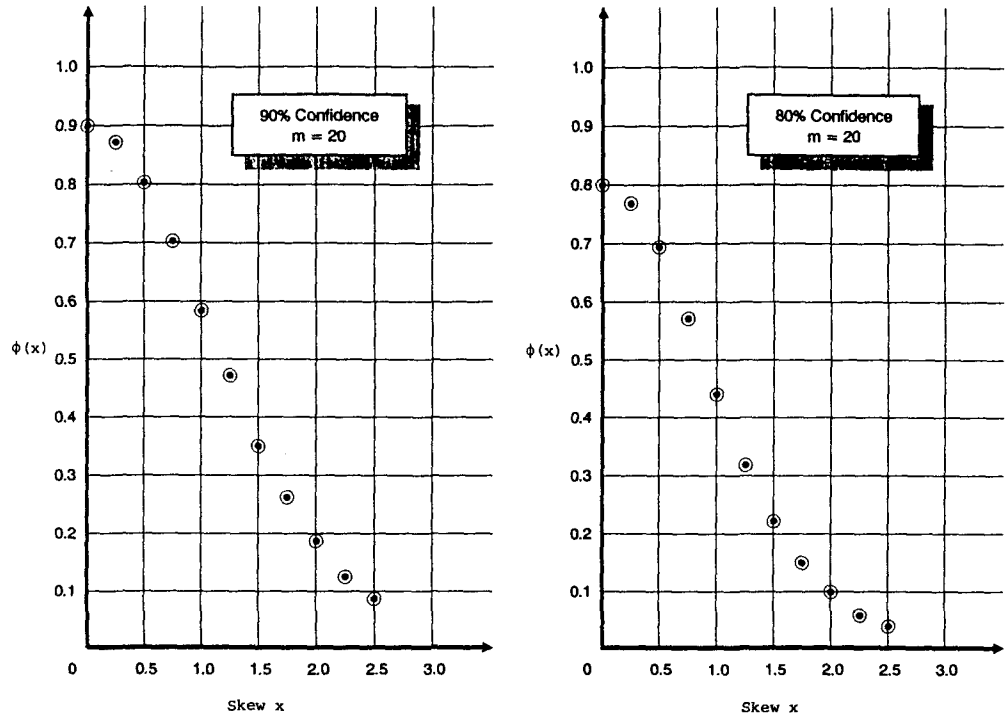

Fig. 1. Operating characteristic $P(|\hat{\gamma}| < t_p|\gamma = x)$, 20 data points.

and 80% test intervals and $m = 20$ and $m = 50$ data points. These curves were obtained from simulated empirical distribution functions for $\hat{\gamma}$ using, for each point on the curves, 30,000 values of $\hat{\gamma}$, based on, respectively, 600,000 or 1,500,000 log Pearson III random variables generated by methods described by *Devroye* [1986]. The data sets for $m = 20$ and $m = 50$ were generated independently.

An examination of these operating characteristic curves show that for the number of years of hydrological data usually available for a single site and the generally observed range of values of skew it is unlikely that one can reject, with a test of moderately high significance the hypothesis that the site has zero skew. For example, with as many as 50 data points and a modest 90% test interval for $\hat{\gamma}$, if the true skew is

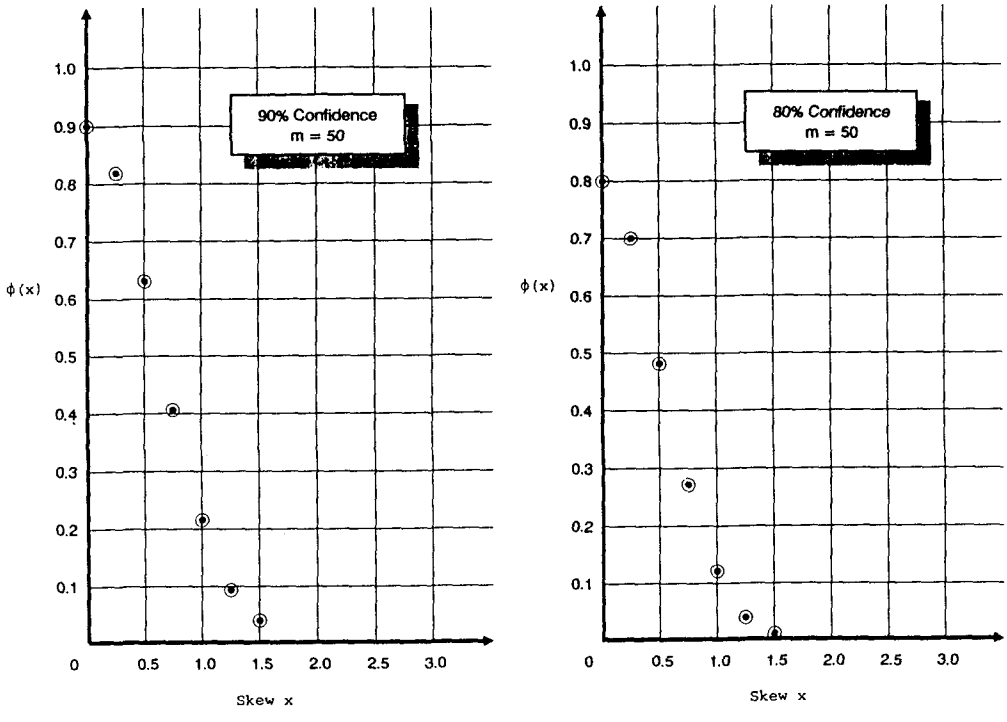

Fig. 2. Operating characteristic $P(|\hat{\gamma}| < t_p|\gamma = x)$, 50 data points.

0.5, the estimator $\hat{\gamma}$ will fall in the zero skew 90% test interval about 63% of the time; when that occurs the skew is mistakenly considered to be zero (i.e., the null hypothesis that the skew is zero cannot be rejected at the 90% significance level).

It is possible, indeed standard, to combine several data sets in an attempt to increase the accuracy of the estimate for the skew. To do this raises the question of which stations to include; the underlying assumption being that those combined stations have discharges coming from distributions with the same skew. The consideration of whether or not that is true brings us back to the problem of understanding the variation in sample skews for single stations. The same problem occurs in making a map of regional skews beginning from a set of station skews. This is not to say that none of this should be done but that the problem of the extent of the variability of site estimates of skew dos not vanish when one regionalizes the data.

As mentioned above, Appendix 14 of Bulletin 17B, Flood Flow Frequency Techniques, summarizes research showing that among the distributions tested, the log Pearson III distribution and the log normal distribution were the two best predictors of large floods among the group of distributions tested and both were approximately equally effective. Because there is no strong reason to prefer log Pearson III over lognormal, and because, as we have seen, in most situations we cannot distinguish between the two distributions with convincing significance, the principle of economical explanation of phenomena should apply. One form of this principle, called Ockham's Razor after William of Ockham (c. 1285–1389), is that "Entities are not to be multiplied beyond necessity," i.e., between two equally effective explanations choose the simpler. The strongest application of this principle would be to use zero skew, thereby using a distribution determined by two parameters rather than three parameters, unless there is a statistically compelling reason to adopt a nonzero skew. Whatever the estimated value of skew used, zero or nonzero, the statistical uncertainty of the value of the skew should be reflected as an uncertainty in the calculated T-year flood value.

CONCLUSIONS

The operating characteristic for a statistical test for zero skew can be computed by simulations. These curves show that it is usually not possible to reject the hypothesis that the skew at any single site is zero, for any convincing significance level, when using only data from that one station.

Combining or regionalizing data will lead to a more accurate estimate of skew if one can decide which stations should be regarded as belonging to the "region"; in making these decisions it will be necessary to consider the statistics of the variation in estimated skews at each single site.

REFERENCES

Arora, N., and V. Singh, A comparative evaluation of the estimators of the log Pearson type (LP) 3 distribution, *J. Hydrol.*, *105*, 19–37, 1989.

Bobee, B., Sample error of T-year events computed by fitting a Pearson type 3 distribution, *Water Resour. Res.*, *9*, 1264–1270, 1973.

Bobee, B., and R. Robitaille, Correction of bias in the estimation of the coefficient of skewness, *Water Resour. Res.*, *11*, 851–854, 1975.

Bobee, B., and R. Robitaille, The use of the Pearson type 3 and log-Pearson type 3 distributions revisited, *Water Resour. Res.*, *13*, 427–443, 1977.

Breiman, L., *Statistics: With a View Toward Applications*, Houghton Mifflin, Boston, Mass., 1973.

Chowdhury, J., and J. Stedinger, Confidence interval for design floods with estimated skew coefficient, *J. of Hydraul. Eng.*, *117*, 811–831, 1991.

Cohon, J., et al., *Estimating Probabilities of Extreme Floods*, National Academy Press, Washington, D. C., 1988.

Devroye, L., *Non-Uniform Random Variable Generation*, Springer, New York, 1986.

Hardison, C., Generalized skew coefficients of annual floods in the United States and their application, *Water Resour. Res.*, *10*, 745–752, 1974.

Hoshi, I., and S. Burges, Sampling properties of parameter estimates for the log Pearson type 3 distribution, using moments in real space, *J. Hydrol.*, *53*, 305–316, 1981.

Hu, S., Determination of confidence intervals for design floods, *J. Hydrol.*, *96*, 201–213, 1987.

Interagency Advisory Committee on Water Data, Guidelines for determining flood flow frequency, *Water Resour. Counc. Bull. 17B*, Hydrol. Subcomm. Off. of Water Data Coord., U.S. Geol. Surv., Reston, Va., 1982.

Kite, G., Confidence intervals for design events, *Water Resour. Res.*, *11*, 48–53, 1975.

Lall, U., and L. Beard, Estimation of Pearson type 3 moments, *Water Resour. Res.*, *18*, 1563–1569, 1982.

Lettenmaier, D., and S. Burges, Correction for bias in estimation of the standard deviation and coefficient of skewness of the log Pearson 3 distribution, *Water Resour. Res.*, *16*, 762–766, 1980.

Matalas, N., and M. Benson, Note on the standard error of the coefficient of skewness, *Water Resour. Res.*, *4*, 204–205, 1968.

McCuen, R., Map skew???, *J. Water Resour. Plann. Manage. Div. Am. Soc. Civ. Eng.*, *105*(2), 269–277, 1979.

Nozdryn-Plotnicki, M., and W. Watt, Assessment of fitting techniques for the log Pearson type 3 distribution using Monte Carlo simulation, *Water Resour. Res.*, *15*, 714–718, 1979.

Phien, H., and L. Hsu, Variance of the T-year event in the log-Pearson type 3 distribution, *J. Hydrol.*, *77*, 141–158, 1985.

Slade, J., The reliability of statistical methods in the determination of flood frequencies, *Water-Supply Pap. 771*, 421–432, U.S. Geol. Surv., Washington, D. C., 1936.

Stedinger, J., Fitting log normal distributions to hydrologic data, *Water Resour. Res.*, *16*, 481–490, 1980.

Stedinger, J., Confidence intervals for design events, *J. Hydraul. Eng.*, *109*, 13–27, 1983.

Tasker, G., Flood frequency analysis with a generalized skew coefficient, *Water Resour. Res.*, *14*, 373–376, 1978.

Tasker, G., and J. Stedinger, Regional skew with weighted LS regression, *J. Water Resour. Plann. Manage. Div. Am. Soc. Civ. Eng.*, *112*, 709–722, 1986.

Tung, Y., and L. Mays, Reducing hydrologic parameter uncertainty, *J. Water Resour. Plann. Manage. Div. Am. Soc. Civ. Eng.*, *107*, 245–262, 1981.

Wallis, J., N. Matalas, and J. Slack, Just a moment!, *Water Resour. Res.*, *10*, 211–219, 1974.

Whitley, R., and T. Hromadka II, Computing confidence intervals for floods, I, *Microsoftware Eng.*, *2*(3), 138–150, 1986*a*.

Whitley, R., and T. Hromadka II, Computing confidence intervals for floods, II, *Microsoftware Eng.*, *2*(3), 151–158, 1986*b*.

Whitley, R., and T. Hromadka II, Estimating 100-year flood confidence intervals, *Adv. Water. Resour.*, *10*, 225–227, 1987.

T. V. Hromadka II, Boyle Engineering, 1501 Quail Street, Newport Beach, CA 92658.

R. Whitley, Department of Mathematics, University of California, Irvine, CA 92717.

(Received May 8, 1992;
revised October 11, 1992;
accepted October 22, 1992.)