FITTING PROCEDURES FOR THE LOG-BOUGHTON DISTRIBUTION

By Walter C. Boughton,[1] M. ASCE and Edward D. Shirley[2]

ABSTRACT: The paper describes a procedure for direct, optimal fitting of the log-Boughton frequency distribution to either complete sets or subsets of annual flood data. The analytical basis and derivation of the procedure are given. Annual floods from 65 yr of record on the 5,755-km^2 Santa Cruz River at Tucson, Arizona, are used to show the fitting of the distribution to a complete data set, and annual floods from 24 yr of record on the 149-km^2 Walnut Gulch watershed, in southeastern Arizona, are used to show the fitting of the distribution to a subset of data. Substantial differences between the log-Boughton and log-Pearson type 3 distributions occur in fitting to the Walnut Gulch data, due mainly to the large negative skew coefficient (−2.83) of this data set. The computer program which fits the log-Boughton distribution plots the data points on a probability paper which is automatically scaled to linearize the fitted distribution.

INTRODUCTION

The original description of the log-Boughton distribution for frequency analyses of annual floods (3) included trial-and-error procedures for fitting the distribution to a set of data. This paper describes a procedure for direct fitting of the distribution to either complete sets or subsets of data.

The original paper, mentioned previously, and an earlier report (2) describe the distribution and its development. A brief description is repeated here as a basis of explanation of the direct fitting procedure.

Using the notation of the Water Resources Council (5) guidelines, the base 10 logarithm of the discharge, Q, at a selected recurrence interval, T, is given by

$$\log Q = \bar{X} + KS \quad \ldots\ldots (1)$$

in which $\bar{X}$ = the mean of the logarithms of the annual floods; S = the standard deviation of the logarithms; and K = the frequency factor for the selected recurrence interval, T years.

The distribution is based on a relationship between frequency factor

[1]Reader, School of Australian Environmental Studies, Griffith Univ., Nathan, Brisbane, Queensland 4111, Australia.

[2]Mathematician, U.S. Dept. of Agr.-Agricultural Research Service, Southwest Rangeland Watershed Research Center, Tucson, Ariz. 85705.

Note.—Discussion open until September 1, 1983. To extend the closing date one month, a written request must be filed with the ASCE Manager of Technical and Professional Publications. The manuscript for this paper was submitted for review and possible publication on December 11, 1981. This paper is part of the Journal of Hydraulic Engineering, Vol. 109, No. 4, April, 1983. ©ASCE, ISSN 0733-9429/83/0004-0579/$01.00. Paper No. 17887.

K and the ln ln $[T/(T-1)]$ function of recurrence interval. For brevity of writing, we use

$$G = \ln \ln \left(\frac{T}{T-1}\right) \quad \ldots\ldots (2)$$

The distribution is based on the relationship

$$(K-A)(G-A) = C \quad \ldots\ldots (3)$$

in which A = a parameter of the distribution which determines its shape, similar to the skew coefficient of the log-Pearson type 3 distribution. C = a constant for a set of data.

When A and C are evaluated, the frequency factor corresponding to a required recurrence interval is given by

$$K = A + \left(\frac{C}{G-A}\right) \quad \ldots\ldots (4)$$

The distribution is fitted to the data so as to minimize the sum of squares of the errors of estimate, using a least squares linear regression. A new mean $\bar{X}^*$ and a new standard deviation S^* are calculated to replace the estimates of these parameters obtained from the data set.

Analytical Model for Fitting the Distribution

Given a data set of N annual floods, the logarithms of each flood value can be normalized by subtracting the mean and dividing by the standard deviation to give a set of actual frequency factors, K_i:

$$K_i = \frac{X_i - \bar{X}}{S} \quad (i = 1 \text{ to } N) \quad \ldots\ldots (5)$$

For any given value of A, there will be N estimates of the constant C, given by

$$(K_i - A)(G - A) = C_i \quad (i = 1 \text{ to } N) \quad \ldots\ldots (6)$$

The analytical model used to fix the shape of the distribution is to find the value of A which minimizes the mean squared error of C. This model also determines the value of C to be used in Eq. 4. The problem is to find values of A and C which minimize the function

$$f(A,C) = \sum_{i=1}^{N} [(K_i - A)(G - A) - C_i]^2 \quad \ldots\ldots (7)$$

Taking the partial derivative of Eq. 7 with respect to C, and setting $\partial f/\partial C = 0$, the optimum value of C is given by

$$C = \overline{KG} - (\overline{K + G})\, A + A^2 \quad \ldots\ldots (8)$$

Substituting the value for C given in Eq. 8 back into Eq. 7, and then taking the partial derivative of Eq. 7 with respect to A, $\partial f/\partial A = 0$ gives

$$A = \frac{\overline{KG\,(K+G)} - \overline{KG}\;\overline{(K+G)}}{[\overline{(K+G)^2}] - [\overline{(K+G)}]^2} \quad \ldots\ldots (9a)$$

$$= \frac{\text{cov}\,(KG, K + G)}{\text{var}\,[(K + G)]} \qquad (9b)$$

Details of the derivation of the equations are set out in Appendix I.

When A and C have been evaluated as shown, a new mean $\bar{X}^*$ and a new standard deviation S^* are determined so as to minimize the sum of squares of the errors of estimating the logarithms of floods in the data set.

The fitting procedure is to calculate values of K^* for each plotting position, using the optimum values of A and C with Eq. 4. For each flood in the data set, there are paired values of frequency factor K^* and base 10 logarithm of the flood X, i.e., there are N pairs of (K^*, X) values. The distribution is fitted as a least squares linear regression of X and K^*. The slope of the regression line is the new standard deviation S^*, and the intercept of the regression line, i.e., when $K^* = 0$, is the mean $\bar{X}^*$, which minimizes the sum of squares of errors.

The distribution is fitted to the data set using these values of $\bar{X}^*$ and S^* with frequency factors obtained from Eq. 4.

Calculation Procedure for Complete Data Set

The most common application is to fit the distribution to a complete data set, and this section describes the relevant procedure for this application. The following section deals with applications where some data values cannot be used in fitting the distributions, e.g., in years of no flow where the logarithms of zero cannot be evaluated.

For complete data sets, in which N is the number of data values:

1. Take the base 10 logarithm of each annual flood.
2. Calculate the mean $\bar{X}$ and standard deviation S of the logarithms.
3. Rank the logarithms in order of magnitude, the highest has rank $m = 1$; the lowest has a rank $m = N$.
4. For each flood, calculate a plotting position, i.e., the probability of exceeding P, and recurrence interval T. The Cunnane plotting position (4) is used for all results shown later in this paper.

$$P = \frac{m - 0.4}{N + 0.2}; \quad \text{in which} \quad m = \text{Rank Number}; \quad \text{and} \quad T = \frac{1}{P} \qquad (10)$$

5. For each flood, calculate

$$G_i = \ln \ln \left(\frac{T}{T - 1}\right); \quad (i = 1 \text{ to } N) \qquad (11)$$

6. For each flood calculate

$$K_i = \frac{\log X_i - \bar{X}}{S}; \quad (i = 1 \text{ to } N) \qquad (12)$$

7. Calculate

$$\overline{KG} = \frac{\Sigma K_i \cdot G_i}{N}; \quad (i = 1 \text{ to } N); \quad (\overline{K + G}) = \frac{\Sigma\,(K_i + G_i)}{N}; \quad (i = 1 \text{ to } N);$$

$$\overline{(K+G)^2} = \frac{\Sigma (K_i + G_i)^2}{N}; \quad (i = 1 \text{ to } N);$$

$$\overline{KG(K+G)} = \frac{\Sigma (K_i G_i)(K_i + G_i)}{N}; \quad (i = 1 \text{ to } N) \ldots\ldots\ldots\ldots\ldots\ldots (13)$$

$$8.\ A = \frac{\overline{KG(K+G)} - \overline{KG}(\overline{K+G})}{\overline{(K+G)^2} - [(\overline{K+G})]^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (14)$$

$$9.\ C = \overline{KG} - (\overline{K+G})\,A + A^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (15)$$

10. For each plotting position, calculate

$$K_i^* = A + \frac{C}{G_i - A}; \quad (i = 1 \text{ to } N) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (16)$$

using G_i from step 5, and A and C from steps 8 and 9.

11. Calculate the new standard deviation S^* and new mean $\bar{X}^*$ by fitting a linear regression as follows

$$S^* = \frac{\Sigma K_i^* X_i - \dfrac{\Sigma K_i^* \Sigma X_i}{N}}{\Sigma (K_i^*)^2 - \dfrac{(\Sigma K_i^*)^2}{N}}; \quad (i = 1 \text{ to } N) \quad \text{and} \quad \bar{X}^* = \bar{X} - \bar{K}^* S^* \ldots\ldots (17)$$

12. The distribution is then used to estimate the flood of recurrence interval T by

$$\log Q_T = \bar{X}^* + K_T S^* \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (18)$$

Calculation Procedures for Subsets of Data

There are many situations where it is impossible or undesirable to include all data values in the fitting of the distribution. In arid areas, years of no flow can occur, and it is impossible to incorporate zero flows when logarithms of the flow must be calculated. In other situations, it may be desirable to ignore one or more very low flows, which can unduly influence the lower value end of the distribution, in order to make the distribution fit more closely to the higher flood values.

In these cases, it is erroneous to ignore the unwanted data values and to treat the remaining values as a complete data set. The use of only nonzero values as a complete data set will give estimates of nonzero floods for all future years, which is erroneous if years of zero flows have already been observed in the period of record.

Existing practices for coping with these situations include adding arbitrary amounts to all data values or separating zero and nonzero floods into separate distributions. These arbitrary practices are not necessary. The solution given by Eqs. 8 and 9 is optimal for fitting the log-Boughton distribution to subsets of data as well as to complete data sets. Let N = total number of data values; d = number of data values to be omitted from the fitting of the distribution; n = number of data values to be included in the fitting of the distribution; thus, $n = N - d$.

The calculation procedure for fitting the distribution to the subset of n values is:

1. Rank the complete set of N values in order of magnitude; in the highest rank $m = 1$.
2. For each flood in the complete set, calculate a plotting position i.e., the probability of exceedence P and recurrence interval T. For instance, using the Cunnane plotting position formula

$$P = \frac{m - 0.4}{N + 0.2} \quad \text{(19)}$$

The use of N, the total number of data values in the complete data set, in calculating the probabilities of exceedence is emphasized to minimize misunderstandings.

3. Discard the data values which are not to be used in fitting the distribution, leaving the subset of n data values.
4. Take the base 10 logarithm of each of the n annual floods.
5. Calculate the mean $\bar{X}$ and standard deviation S of the logarithms of the subset.
6. For each flood in the subset, calculate

$$G_i = \ln \ln \left(\frac{T}{T-1}\right); \quad (i = 1 \text{ to } N) \quad \text{(20)}$$

7. For each flood in the subset, calculate

$$K_i = \frac{\log X_i - \bar{X}}{S}; \quad (i = 1 \text{ to } N) \quad \text{(21)}$$

8. Calculate $\overline{KG} = \frac{\Sigma K_i G_i}{n}; \quad (i = 1 \text{ to } N);$

$$\overline{(K+G)} = \frac{\Sigma (K_i + G_i)}{n}; \quad (i = 1 \text{ to } N); \quad \overline{(K+G)^2} = \frac{\Sigma (K_i + G_i)^2}{n};$$

$$(i = 1 \text{ to } N); \quad \overline{KG\,(K+G)} = \frac{\Sigma K_i G_i (K_i + G_i)}{n}; \quad (i - 1 \text{ to } N) \quad \text{(22)}$$

9. $$A = \frac{\overline{KG(K+G)} - \overline{KG}\,\overline{(K+G)}}{\overline{(K+G)^2} - [\overline{(K+G)}]^2} \quad \text{(23)}$$

10. $$C = \overline{KG} - \overline{(K+G)}\,A + A^2 \quad \text{(24)}$$

11. Continue as in steps 10–12 for the complete data set.

Computer Program

The calculation procedures just described, together with a plotting routine for automatically drawing the data points on probability paper, have been programmed in FORTRAN on the PDP 11/34 computer at the Southwest Rangeland Watershed Research Center in Tucson. The computer is equipped with a Techtronix graphics terminal and hardcopy

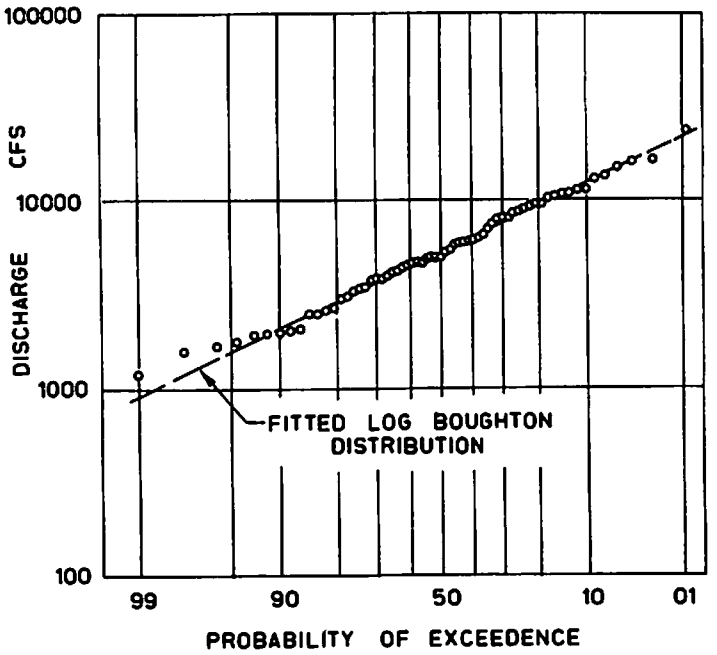

FIG. 1.—Santa Cruz River at Tucson, Arizona: Log-Boughton Distribution Fitted to Complete Data Set (1915–79)

attachment. A copy of the program will be made available free on request to the Director, Southwest Rangeland Watershed Research Center, 442 East Seventh Street, Tucson, Ariz. 85705. (Note: Mention of the equipment by name is given for completeness of information and does not imply endorsement by the U.S. Department of Agriculture.)

Examples of Fitting the Distribution

Fitting to a Complete Data Set.—The annual floods from 65 yr of record, 1915–1979, on the Santa Cruz River at Tucson in southern Arizona (1), are used to illustrate the fitting of the distribution to a complete data set. The basic data and calculated values of $\bar{X}^*$, S^*, C, and A are shown in Table 1.

Figure 1 shows the data points and the fitted distribution on a probability paper which has been drawn to linearize the distribution for $A = 4.5$. Because the distribution has more than two parameters, no single probability paper will linearize the distribution for all values of A. However, the computer program used at the Southwest Rangeland Watershed Research Center to fit the distribution to a set of data includes a plotting subroutine which automatically draws the probability paper to suit the fitted value of A and plots the data points on this paper.

The fitted distribution was used to calculate flood magnitudes of 2, 10, and 100 yr recurrence intervals. The log-Pearson type 3 distribution was also fitted to the data set (skew coefficient = −0.0889), and floods of the same recurrence intervals were calculated using the distribution. The results were very similar, as shown by the comparison in Table 2. However, results from data sets which have much larger negative skew coefficients can differ greatly, as shown later in this paper.

TABLE 1.—Santa Cruz River at Tucson, Arizona Station No. 09482500

Annual peak discharges, in cubic feet per second (1)	0 (2)	1 (3)	2 (4)	3 (5)	4 (6)	5 (7)	6 (8)	7 (9)	8 (10)	9 (11)
191	—	—	—	—	—	15,000	5,000	7,500	4,900	4,700
192	1,950	4,000	2,000	1,900	2,050	3,400	11,400	1,950	1,600	10,400
193	1,770	9,200	4,200	6,100	6,000	10,300	5,400	3,280	9,000	8,000
194	11,300	2,490	1,670	4,510	6,530	10,800	4,260	2,960	3,860	3,800
195	9,490	5,020	3,820	5,900	9,570	10,900	2,610	3,050	6,350	4,420
196	6,140	16,600	4,980	4,670	13,000	1,190	5,500	5,860	16,100	8,710
197	8,530	8,000	3,470	4,710	7,930	2,480	7,100	2,660	23,700	13,500

Note: 1 cu ft/sec = 0.028 m^3/sec. $\bar{X}^*$ = 3.7203; S^* = 0.2755; A = 4.4976; C = 21.5990. Drainage area = 2,222 sq mile (5,755 km^2).

Fitting to a Subset of Data.—The 57.66 sq mile Walnut Gulch watershed, in southeastern Arizona, has been gaged by the U.S. Department of Agriculture-Agricultural Research Service (USDA-ARS) Southwest Rangeland Watershed Research Center since 1957. The flow is ephemeral, and transmission losses can reduce the runoff peak discharge as flow traverses the dry alluvial stream channel in moving through the catchment to the gaging station (6). Smaller floods can be affected by such losses to an extent that they become separated from the general distribution of larger floods.

Annual floods from 24 yr of record, 1957–1980, on station 63.001, Flume 1 on the Walnut Gulch watershed, are shown in Table 3. In 1979, there was very little runoff at the outlet of this watershed, resulting in a maximum runoff rate of only 0.6 cu ft/sec (0.168 m^3/s). Another low rate of 54 cu ft/sec (1.5 m^3/s) in 1960 is notably lower than the other annual floods in the data set. Table 4 shows the resulting of fitting the distribution to the complete data set, and then to several subsets, omitting 1, 2, 3, 4, and 5 lowest data values in turn.

When the distribution is fitted to the complete data set, the effect of the two lowest annual floods is to give the data set a negative skew, resulting in low values of A and C and unduly large estimates of the 2- and 10-yr recurrence interval floods. As the lowest flood values are progressively omitted from the fitting of the distribution, the mean $\bar{X}^*$, increases, and the standard deviation S^* decreases, as is expected to occur.

TABLE 2.—Santa Cruz River—Comparison of Flood Estimates from Two Distributions, in Cubic Feet per Second

Distribution (1)	Recurrence Interval, in Years		
	2 (2)	10 (3)	100 (4)
Log-Boughton	5,450	12,000	20,500
Log-Pearson type 3	5,280	12,200	23,500

Note: 1 cu ft/sec = 0.28 m^3/s.

TABLE 3.—Walnut Gulch Flume 1 Station No. 63.001

Annual peak discharges, in cubic feet per second (1)	0 (2)	1 (3)	2 (4)	3 (5)	4 (6)	5 (7)	6 (8)	7 (9)	8 (10)	9 (11)
195								11,253	3,388	2,767
196	54	3,928	851	2,709	4,288	841	1,574	4,680	808	1,679
197	710	3,615	6,057	2,978	639	2,071	1,365	2,850	1,229	0.6
198	360									

Note: 1 cu ft/sec = 0.028 m^3/s. Drainage area = 57.66 sq mile (149.3 km^2).

The skewness of the data set reaches its most positive value when three data values are omitted, and A and C reach their maximum values before declining as further data are omitted.

The estimated floods of 2-, 10- and 100-yr recurrence intervals do not vary much when the 2, 3, 4 or 5 lowest values are omitted in turn from the fitting of the distribution. This suggests that only the two lowest values are outliers from the rest of the data set. The ability to fit the distribution readily to several subsets of data is a valuable tool in checking the effects of outliers on the fit of the distribution.

TABLE 4.—Walnut Gulch Flume 1—Results of Fitting the Distribution to Complete Data Set and to Several Subsets of Data

Number of values omitted (1)	Parameter Values				Calculated Flood, in Cubic Feet per Second of Specified Recurrence Interval, in Years		
	$\bar{X}^*$ (2)	S^* (3)	A (4)	C (5)	2 (6)	10 (7)	100 (8)
None (complete data set)	3.2259	0.3057	1.4597	2.0885	2,100	3,160	3,690
Lowest 1 value omitted	3.2170	0.5316	2.3530	6.0535	1,930	5,870	10,100
Lowest 2 values omitted	3.2877	0.4182	3.9648	17.5031	1,800	5,860	12,300
Lowest 3 values omitted	3.3222	0.3952	4.3032	20.8686	1,800	5,810	12,500
Lowest 4 values omitted	3.3455	0.3972	4.1317	19.6064	1,800	5,840	12,400
Lowest 5 values omitted	3.3690	0.3982	3.9155	17.9839	1,800	5,840	12,200

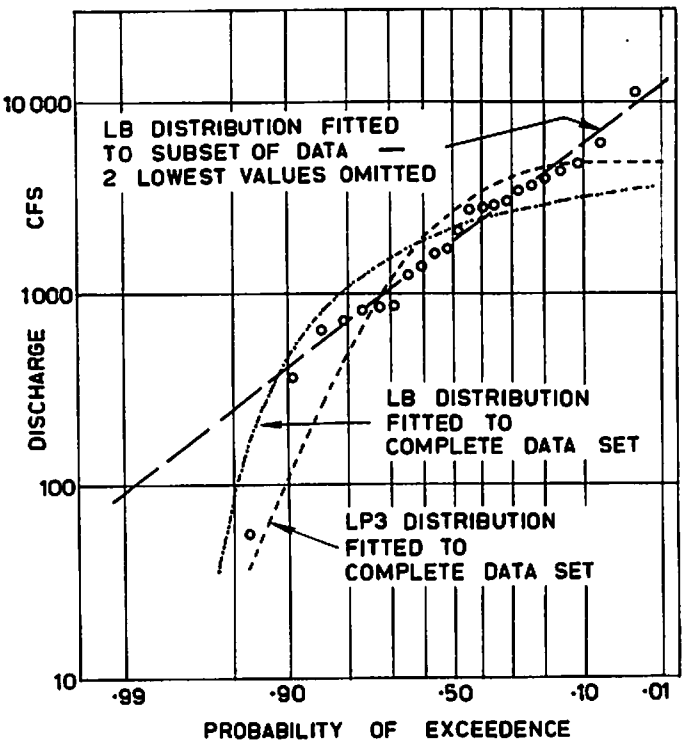

FIG. 2.—Flume 1 at Walnut Gulch, Tombstone, Arizona: Distribution Fitted to Subset of Data

Figure 2 shows the log-Boughton distribution fitted to the subset of data with the two lowest values omitted. For comparison, the figure also shows the log-Boughton and the log-Pearson type 3 distributions fitted to the complete data set. The effect of the low flood values in distorting the skew coefficient to a large negative value of −2.8367 is obvious when the distributions are fitted to the complete data set.

Because of the simplicity of the procedure for fitting the log-Boughton distribution to subsets of data, it is possible to examine quickly the results from both subsets and complete sets. The quality of judgments which are made by practicing professionals is dependent upon the quality of information on which the judgments are based. The ease with which the distribution can be fitted to subsets of data helps increase the amount of information on which estimates of flood frequencies are made. The other major benefit is that records with years of zero flow can be fitted directly without making arbitrary adjustments, such as the adding of a constant to all values.

Conclusions

The analysis of the log-Boughton distribution described in the paper has produced an optimal solution for direct fitting of the distribution to a set of data. This simplifies and improves on the trial-and-error fitting procedures used before.

The fitting procedure can be used for subsets as well as complete sets of data. Records containing years of zero flow cannot be included in any fitting of a distribution to the logarithms of data. The method reported

in the paper provides a means for optimal fitting of the distribution to a subset of data as well as to complete sets.

Acknowledgment

This paper was prepared while the first author was on study leave from Griffith University, Brisbane, Australia. The opportunity and assistance given by the University for travel to the U.S. is acknowledged. The U.S. Department of Agriculture made facilities and data available at the Southwest Rangeland Watershed Research Center to enable the study to be undertaken.

Appendix I.—Derivation of the Fitting Equations

Given $(K - A)(G - A) = C \quad [N(K, G) \text{ points}]$ ..... (25)

the chosen objective is to minimize the variance of C. This can be stated as finding values of A and C which minimize the function

$$f(A, C) = \sum_{i=1}^{N} [(K_i - A)(G - A) - C_i]^2 = \Sigma\,[KG - (K + G)\,A + A^2 - C]^2. \quad (26)$$

The minimum value of the function occurs when

$$\frac{\partial f}{\partial C} = \frac{\partial f}{\partial A} = 0 \quad \ldots\ldots (27)$$

Taking the partial derivative of the function with respect to C

$$\frac{\partial f}{\partial C} = \Sigma\, 2\,[KG - (K + G)\,A + A^2 - C][-1] = 0;$$

$$\Sigma\,(KG) - [\Sigma(K + G)]\,A + NA^2 - NC = 0 \quad \ldots\ldots (28)$$

Dividing by $N \quad \overline{KG} - (\overline{K + G})\,A + A^2 - C = 0;$

$$C = \overline{KG} - (\overline{K + G})\,A + A^2 \quad \ldots\ldots (29)$$

Taking the partial derivative of the function with respect to A

$$\frac{\partial f}{\partial A} = \Sigma\, 2\,[KG - (K + G)\,A + A^2 - C][-(K + G) + 2A] = 0 \quad \ldots\ldots (30)$$

Expanding and dividing by $2N$

$$\overline{-KG\,(K + G)} + \overline{(K + G)^2}\,A - (\overline{K + G})\,A^2$$
$$+ (\overline{K + G})\,C + 2\,(\overline{KG})\,A - 2\,(\overline{K + G})\,A^2 + 2A^3 - 2AC = 0 \quad \ldots\ldots (31)$$

Substitution Eq. 29 for C gives

$$A = \frac{\overline{KG(K + G)} - \overline{KG}\,(\overline{K + G})}{\overline{(K + G)^2} - [(\overline{K + G})]^2} = \frac{\text{cov}\,(KG, K + G)}{\text{var}\,(K + G)} \quad \ldots\ldots (32)$$

APPENDIX II.—REFERENCES

1. Anderson, T. W., and White, N. D., "Statistical Summaries of Arizona Streamflow Data," U.S. Geological Survey, Tucson, Ariz., Jan., 1979, 420 pp.
2. Boughton, W. C., "A Study of Queensland Floods," *Symposium on the Frequency of Floods in Queensland*, Institution of Engineers, Australia, Queensland Division, Brisbane, Australia, 1975.
3. Boughton, W. C., "A Frequency Distribution for Annual Floods," *Water Resources Research*, Vol. 16, No. 2, Apr., 1980, pp. 347–354.
4. Cunnane, C., "Unbiased Plotting Positions—A Review," *Journal of Hydrology*, Vol. 37, No. 3, 1978, pp. 205–222.
5. "Guidelines for Determining Flood Flow Frequency," Water Resources Council, Bulletin 17, Hydrology Committee, Washington, D.C.
6. Reich, B. M., and Renard, K. G., "Application of Advances in Flood Frequency Analysis," *Water Resources Bulletin*, Vol. 17, No. 1, Feb., 1981, pp. 67–74.

APPENDIX III.—NOTATION

The following symbols are used in this paper:

A = fitted parameter in LB distribution;
C = constant for a data set, used in fitting LB distribution;
G = ln ln $T/(T - 1)$;
K = frequency factor derived from data;
K^* = frequency factor calculated from LB distribution;
N = number of data in complete set;
P = plotting position (probability of exceedence);
Q = flood magnitude;
S = standard deviation of logarithms of flood magnitudes;
S^* = slope of linear regression;
T = recurrence interval;
X = logarithm of flood magnitude;
$\bar{X}$ = mean of logarithms of flood magnitudes;
$\bar{X}^*$ = intercept of linear regression;
d = number of data values omitted when fitting the distribution;
n = number of data values included when fitting the distribution; and
m = rank number of flood.