Section 8-III

STATISTICAL AND PROBABILITY ANALYSIS OF HYDROLOGIC DATA

PART III. ANALYSIS OF VARIANCE, COVARIANCE, AND TIME SERIES

D. R. DAWDY *and* **N. C. MATALAS,** *Hydraulic Engineers, U.S. Geological Survey.*

I. Analysis of Variance and Covariance 8-69
A. Introduction 8-69
1. Definition 8-69
2. Chi-square Distribution 8-69
3. *F* Distribution 8-70
B. Analysis-of-variance Models 8-72
1. One-way Classification 8-72
2. Two-way Classification 8-73
3. Linearity of Regression 8-74
C. Analysis-of-covariance Models 8-75
1. One-way Classification 8-75
2. Study of Regression Effect 8-76
II. Analysis of Time Series 8-78
A. Introduction 8-78
1. Definition of Time Series 8-78
2. Characteristics of Time Series 8-79
3. Properties of the Nonrandom Element 8-79
B. Trend Analyses 8-81
1. Use of Moving Averages 8-81
2. Slutzky-Yule Effect 8-82
C. Tests for Serial Dependence 8-82
1. Parametric Test of Significance 8-82
2. Nonparametric Tests of Significance 8-83
D. Generating Processes 8-84
1. Definition 8-84
2. Moving-average Process 8-84
3. Sum-of-harmonics Process 8-84
4. Autoregression Process 8-85
5. Correlograms 8-85

E. Effect of Serial Correlation 8-85
1. Estimation of the Variance 8-85
2. Correlation and Regression Analyses 8-86
III. References 8-89

I. ANALYSIS OF VARIANCE AND COVARIANCE

A. Introduction

1. Definition. Analysis of variance is a statistical technique which tests mean values by a partitioning of the total variance of a sample into component parts, each of which can be assigned to a particular cause. Thus, in a study of point rainfall, several rain gages may be used and several storms measured by each. Within the total variance of all measured values of precipitation, there is a portion of the variation which is due to the variation of the mean values recorded for each gage and there is a portion which is the result of the variation of individual values about these mean values. A comparison of the two variances can aid in determining whether any measured difference in average rainfall is due to chance. The partitioning of the variance in an analysis of variance is determined by the test to be made.

The partitioning of the variance can include the covariance of the variable being studied with another independent variable. Thus the difference in mean values can be studied after they have been corrected for the effect of the correlated independent variable.

Before discussing the analysis of variance, it is well to consider two probability distributions which play an important role in the analysis. These are the chi-square distribution and the F distribution.

2. Chi-square Distribution. The distribution of the variance is fundamental to many tests of statistical inference. First, consider the distribution of the sum of squares of a variable. Let $x_1, x_2, \ldots, x_n$ be normally and independently distributed variables, each with mean 0 and variance 1. Then

$$\chi^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \tag{8-III-1}$$

is called *chi-square* and has the probability density function

$$p(\chi^2) = \frac{1}{2^{\frac{1}{2}\nu}\Gamma\left(\frac{\nu}{2}\right)} (\chi^2)^{\frac{1}{2}(\nu-2)} \exp\left(-\frac{1}{2}\chi^2\right) \tag{8-III-2}$$

where ν is called the *number of degrees of freedom*, and Γ represents the gamma function. Figure 8-III-1 shows this distribution, which is tabulated in most standard statistics books for $\nu \leq 30$ (i.e., Hoel [1]). For larger values of ν, the quantity $(2\chi^2)^{\frac{1}{2}} - (2\nu - 1)^{\frac{1}{2}}$ is approximately normally distributed with mean 0 and variance 1. Since, from Eq. (8-III-1), it can be shown that ns^2/σ^2 is distributed like χ^2 with $\nu = n - 1$ degrees of freedom, where s^2 and σ^2 are the sample and population variances, respectively, then the sample variance of the χ_i's is distributed as χ^2 with $n - 1$ degrees of freedom [2].

Equation (8-III-2) is the basis for a test of whether or not a sample variance is significantly different from a presumed population variance. If, for instance, a regionalized flood-frequency curve gives a variance for annual floods of σ_0^2, and if n floods for a given station which was not used to determine σ_0^2, and is therefore independent of it, have a variance of s^2, then the ratio of the sample variance to the "true," or regional, variance can be tested as $s^2/\sigma_0^2 = \chi^2/n$. The critical value of χ^2

for rejecting the null hypothesis $H_0: s^2 = \sigma_0^2$ is taken from tables of the chi-square distribution at a chosen confidence level with $\nu = n - 1$ degrees of freedom.

Often it is desirable to test the hypothesis that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$. Consider that a record of annual precipitation is available for n years and that during this period of time the location of the rain gage has been moved k times. Thus the homogeneity of the rainfall record is questioned. In order to test the hypothesis of the homogeneity of the variance, the rainfall record is divided into $k + 1$ parts, each being that part of the record during which time the rain gage was at a particular location. Let n_i denote the number of years in the ith segment of the record and s_i^2 the variance of the rainfall data in the ith segment, where $i = 1, 2, \ldots, k + 1$. Bartlett [3] has

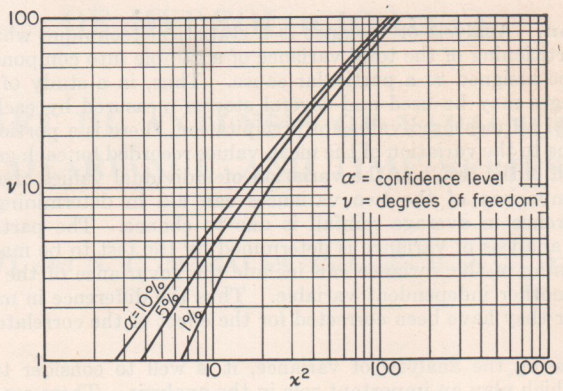

FIG. 8-III-1. Critical values of χ^2.

shown that the hypothesis of homogeneous variance can be tested by means of the chi-square distribution, where χ^2 for k degrees of freedom is defined as

$$\chi_k^2 = \frac{2.3026}{c}\left[\nu \log\left(\frac{1}{\nu}\sum_{i=1}^{k+1}\nu_i s_i^2\right) - \sum_{i=1}^{k+1}\nu_i \log s_i^2\right] \tag{8-III-3}$$

where

$$c = 1 + \frac{1}{3k}\left(\sum_{i=1}^{k+1}\frac{1}{\nu_i} - \frac{1}{\nu}\right) \tag{8-III-4}$$

and

$$\nu = \sum_{i=1}^{k+1}\nu_i = \sum_{i=1}^{k+1}(n_i - 1) = n - (k + 1) \tag{8-III-5}$$

If the value of χ^2 computed by Eq. (8-III-3) exceeds the tabulated value of χ^2 for k degrees of freedom at a chosen confidence level, then the hypothesis of homogeneity of the variance is rejected. It should be noted that the value of c is greater than unity and tends to unity as the number of degrees of freedom increases. Hence, the value of c may be set equal to unity in Eq. (8-III-3) unless there is some doubt about the significance of χ_k^2. An insignificant value of χ_k^2 cannot be made significant by using the actual value of c computed by Eq. (8-III-4).

3. *F* Distribution. Assume that x_i ($i = 1, \ldots, n_1$) and y_j ($j = 1, \ldots, n_2$) are two independent random samples and that s_1^2 and s_2^2 are unbiased estimates of the variance for each sample. The question arises whether or not the two samples have been drawn from the same normal population having variance σ^2. Thus it is necessary to test the hypothesis $\sigma_1^2 = \sigma_2^2$.

The basis for testing the hypothesis $\sigma_1^2 = \sigma_2^2$ is the ratio of the two sample variances, which is defined as

$$F = \frac{s_1^2}{s_2^2} = \frac{\nu_1 s_1^2/\nu_1\sigma^2}{\nu_2 s_2^2/\nu_2\sigma^2} \tag{8-III-6}$$

whereby

$$\frac{\nu_1 F}{\nu_2} = \frac{\nu_1 s_1^2/\sigma^2}{\nu_2 s_2^2/\sigma^2} \tag{8-III-7}$$

From the previous discussion of the chi-square distribution, it is seen that the numerator and denominator of the right-hand side of Eq. (8-III-7) are distributed inde-

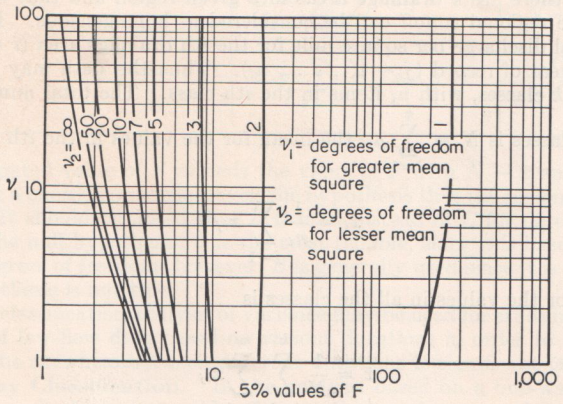

FIG. 8-III-2. Critical values of F at the 5 per cent level.

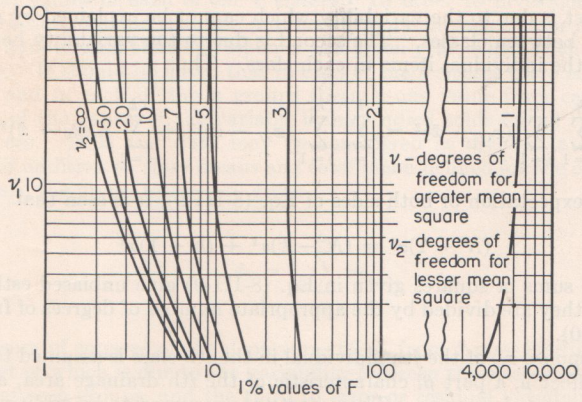

FIG. 8-III-3. Critical values of F at the 1 per cent level.

pendently as χ^2, with ν_1 and ν_2 degrees of freedom, respectively. It can be shown that the probability density function of F is

$$p(F) = \frac{\nu_1^{\frac{1}{2}\nu_1}\nu_2^{\frac{1}{2}\nu_2}F^{\frac{1}{2}(\nu_1-2)}}{B(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2)(\nu_1 F + \nu_2)^{\frac{1}{2}(\nu_1+\nu_2)}} \tag{8-III-8}$$

where B denotes the beta function. Figures 8-III-2 and 8-III-3 show values of F corresponding to cumulative probabilities of 0.01 and 0.05 for various values of ν_1 and ν_2. These values, which are available in tabular form (i.e., Hoel [1]), facilitate the

use of the F distribution for making statistical inferences when analysis-of-variance techniques are used.

B. Analysis-of-variance Models

1. One-way Classification. Quite often it is necessary to determine whether or not an abrupt change in the mean value of some hydrologic statistic, e.g., measured precipitation or streamflow, has occurred or whether an appreciable difference in rainfall or runoff exists among several experimental watersheds. This test can be made through an analysis of variance.

Assume that there are k drainage areas in a given region and that it is desired to determine if the regional runoff can be considered as homogeneous. Let x_{ij} denote the mean annual discharge per square mile for the ith drainage area ($i = 1, \ldots, k$) during the jth year of record ($j = 1, \ldots, n_i$). Thus the data may be considered as divided into k classes, with n_i items in the ith class. The total number of values within all the classes is $N = \sum_{i=1}^{k} n_i$, the mean for the values in the ith class is

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

and the mean for the values in all the classes is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

The total sum of squares of the departures of x_{ij} from $\bar{x}$ may be divided into two parts. The first is due to the variability which cannot be explained by differences of regional effects between classes. The second is due to the variability between classes averaged over the individual items in each class. Thus

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 \qquad \text{(8-III-9)}$$

By taking the expectation of both sides of Eq. (8-III-9), it is seen that

$$(N - 1)\sigma^2 = (N - k)\sigma^2 + (k - 1)\sigma^2 \qquad \text{(8-III-10)}$$

Thus the three sums of squares given in Eq. (8-III-9) give unbiased estimates of the variance when they are divided by the appropriate number of degrees of freedom given in Eq. (8-III-10).

The total response x_{ij} of the jth individual in the ith class is assumed to be made up of an overall effect μ, a part β_i characteristic of the ith drainage area, and a part ϵ_{ij} which can be regarded as error. These parts are assumed to be additive, so that

$$x_{ij} = \mu' + \beta_i' + \epsilon_{ij} \qquad \text{(8-III-11)}$$

where μ' is adjusted ($\mu' = \mu + \bar{\beta}_i$) so that $\sum_{i=1}^{k} \beta_i' = 0$. It is also assumed that each ϵ_{ij} is an independent random normal variate with expectation 0 and variance σ^2, independent not only of the other ϵ's, but also of the β's.

The hypothesis which is to be tested is that the β's are all zero, in which case the regional runoff may be considered as homogeneous. The testing of this hypothesis may be summarized in an analysis-of-variance table, such as Table 8-III-1.

Table 8-III-1. Test of Hypothesis for One-way Classification

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Between class means.....	$k-1$	$A=\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2$	$\frac{A}{k-1}$	$\frac{A(N-k)}{B(k-1)}$
Within classes.....	$N-k$	$B=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$	$\frac{B}{N-k}$	
Total.....	$N-1$	$C=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$	$\frac{C}{N-1}$	

If the computed value of F exceeds the value of F with $N-k$ and $k-1$ degrees of freedom for a chosen confidence level, the hypothesis that the region is homogeneous is rejected. It should be noted that B is used instead of C for determining F. On the basis of the null hypothesis, C is the most reliable, since it is based on the largest number of degrees of freedom; however, B is generally used since it is valid even when the null hypothesis is not true.

A one-way-classification analysis of variance may be used for studying the coefficient of skewness of low-flow data based on various durations in order to determine if the variation of the skewness between different durations is significant [4].

2. Two-way Classification. In the analysis based on a one-way classification, the values in each class are considered as replicates of one another, subject only to random variation. However, there may be a possible significant variation between the individual values in each of the k classes. In order to investigate this possible source of variation, assume that the number of items in each class is a constant equal to n. Each item in each class corresponds to a given year, where each year is referred to as a group. It should be noted that in each of the k classes there is one value from each group, and in each of the n groups there is one value from each class. This arrangement of the values of the variable being studied holds for all two-way-classification analyses. Thus the data may be considered as divided into k classes and n groups. In addition to class means and total mean defined for the one-way-classification analysis, the group means are given by

$$\bar{x}_j=\frac{1}{k}\sum_{i=1}^{k} x_{ij}$$

The total sum of squares of the departures of x_{ij} from $\bar{x}$ may be divided into three parts, the first of which is due to the variability between the classes, the second to the variability between groups, and the third to error or residual variability. Thus

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2=\sum_{i=1}^{k} n(\bar{x}_i-\bar{x})^2+\sum_{j=1}^{n} k(\bar{x}_j-\bar{x})^2+\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{x})^2 \quad \text{(8-III-12)}$$

By taking the expectation of both sides of Eq. (8-III-12), it is seen that

$$(nk-1)\sigma^2=(k-1)\sigma^2+(n-1)\sigma^2+(nk-n-k+1)\sigma^2 \quad \text{(8-III-13)}$$

whereby the four sums of squares given in Eq. (8-III-12) provide unbiased estimates of the variance when they are divided by their appropriate number of degrees of freedom given in Eq. (8-III-13).

The mathematical model is now

$$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{8-III-14}$$

where μ is the overall effect, α_i is the characteristic of the ith drainage area, β_j is the characteristic of the jth year of record, and ϵ_{ij} is the error. The overall effect μ is considered to be adjusted so that $\sum_{i=1}^{k} \alpha_i = 0$ and $\sum_{j=1}^{n} \beta_j = 0$. It is assumed that the ϵ_{ij}'s are all normally and independently distributed about zero with the same variance σ^2.

The hypothesis which is to be tested is that all the α_i's and β_j's are zero, in which case the region may be considered as homogeneous with respect to space and time. The testing of this hypothesis is summarized in Table 8-III-2.

Table 8-III-2. Test of Hypothesis for Two-way Classification

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Between classes...	$k - 1$	$A = \sum_{i=1}^{k} n(\bar{x}_i - \bar{x})^2$	$\frac{A}{k-1}$	$\frac{A(n-1)}{C}$
Between groups...	$n - 1$	$B = \sum_{j=1}^{n} k(\bar{x}_j - \bar{x})^2$	$\frac{B}{n-1}$	$\frac{B(k-1)}{C}$
Error.....	$(k-1)(n-1)$	$C = \sum_{i=1}^{k}\sum_{j=1}^{n} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$	$\frac{C}{(k-1)(n-1)}$	
Total.....	$nk - 1$	$D = \sum_{i=1}^{k}\sum_{j=1}^{n} (x_{ij} - \bar{x})^2$	$\frac{D}{nk-1}$	

If the first value of F is found to exceed the value of F with $(k-1)(n-1)$ and $k-1$ degrees of freedom, then the hypothesis is rejected that all the α's are zero, and if the second value of F exceeds the value of F with $(k-1)(n-1)$ and $n-1$ degrees of freedom, then the hypothesis is rejected that all the β's are zero. If both computed values of F are found to be significant, then the entire hypothesis of homogeneity is rejected.

3. Linearity of Regression. Many hydrologic studies are based on regression analyses where generally linear functions are fitted to the data. However, in some studies, a linear function may be questionable and it may be necessary to test for nonlinearity of the regression line. This test may be made by an appropriate analysis of variance.

The data should be partitioned into k arrays according to the values of the independent variable. The range of values of the independent variable within each array should be narrow enough so that the range in the values of the dependent variable in each array approximates the spread of the values of the dependent variable about the regression line within each array. Let Y_i be the estimate of the dependent variable from the regression line for the mean value of the independent variable in the ith array, let y_{ij} be the jth value ($j = 1, \ldots, n_i$) of the dependent variable in the ith array, let $\bar{y}_i$ be the mean of the values of the dependent variable in the ith array, and let $\bar{y}$ be the mean for all the values of the dependent variable.

The sum of squares of $(y_{ij} - \bar{y})^2$, which is proportional to the variance of the dependent variable, may be divided into three parts. The first part is due to the regression function itself; the second part is due to the deviations of the means from the regression line; and the third part is due to the variation within the arrays. Thus

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} n_i(Y_i - \bar{y})^2 + \sum_{i=1}^{k} n_i(\bar{y}_i - Y_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \quad \text{(8-III-15)}$$

By taking the expectation of Eq. (8-III-15), it is seen that

$$(N - 1)\sigma^2 = \sigma^2 + (k - 2)\sigma^2 + (N - k)\sigma^2 \quad \text{(8-III-16)}$$

where $N = \sum_{i=1}^{k} n_i$. Thus the four sums of squares given in Eq. (8-III-15) provide unbiased estimates of the variance when they are divided by their appropriate number of degrees of freedom given in Eq. (8-III-16). The test for linearity is summarized in Table 8-III-3.

Table 8-III-3. Test for Linearity

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Linear regression	1	$A = \sum_{i=1}^{k} n_i(Y_i - \bar{y})^2$	$\frac{A}{1}$	
Deviation of means from regression line	$k - 2$	$B = \sum_{i=1}^{k} n_i(\bar{y}_i - Y_i)^2$	$\frac{B}{k-2}$	$\frac{B(N-k)}{C(k-2)}$
Within arrays	$N - k$	$C = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$\frac{C}{N-k}$	
Total	$N - 1$	$D = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$\frac{D}{N-1}$	

On the assumption of linearity of regression, the sum of squares denoted by B is due to sampling errors and the estimate of σ^2 obtained from B should not be greater than that derived from the sum of squares within arrays, which is denoted by C. If the computed value of F is found to be significant, then the hypothesis of linearity of regression is rejected.

C. Analysis-of-covariance Models

1. One-way Classification. At times it is desirable to test the significance of the difference in mean values of a variable after these means have been corrected for the effect of another correlated variable. Thus a study of the effect of deforestation or reforestation upon streamflow must first eliminate the effect of varying precipitation through the test period. If the data in this example were classified by years, the covariance of streamflow with precipitation might be removed and the remainder would be variance between years.

The covariance first is partitioned into

$$\sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y}) = \sum_i \sum_j (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i) + \sum_i n_i(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) \quad \text{(8-III-17)}$$

where the first term on the right represents covariance within classes (years), and the second term that between classes (years). All three terms give an estimate of the covariance, assuming there is no difference between years. The analysis of covariance, as shown in Table 8-III-4, is intended to test the null hypothesis that there is no difference in covariance between years and in the population as a whole. Thus our null hypothesis would be that reforestation had no effect on streamflow. This type of analysis can also be used for studying the significance of changes in the slopes of double-mass curves [5].

Table 8-III-4. Analysis of Covariance

Source of covariance	Degrees of freedom	Sum of cross products	F
Between classes (years)	$k - 1$	$A = \sum_i n_i(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})$	
Within classes (years)	$N - k$	$B = \sum_i \sum_j (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)$	$(k - 1)B/(N - k)A$
Total	$N - 1$	$C = \sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y})$	

A more exact method of testing the same hypothesis would be to determine the significance of the coefficient of regression of streamflow with precipitation within years. If this is significant, then the yearly means can be corrected for the regression effect, and the differences between the corrected yearly means then are tested for significance. In addition, the significance of the difference between the two regression coefficients for between and within years can be tested to determine whether the classification by years has an effect upon the degree of association of the variables.

The coefficients of regression are given in Table 8-III-5 with the t statistic

$$t_{N-k} = A\sqrt{\frac{(N - k)\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}} \quad \text{(8-III-18)}$$

used to test the significance of the within-years coefficient.

2. Study of Regression Effect. In order to correct the yearly means for the regression effect, the variance must be partitioned. First, it is partitioned into that due to the regression of streamflow with precipitation and that due to deviations from the regression line. The latter part, then, is further partitioned into that within years and that between years.

The variation due to regression is

$$A = \frac{\left[\sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y})\right]^2}{\sum_i \sum_j (x_{ij} - \bar{x})^2} \quad \text{(8-III-19)}$$

Table 8-III-5. Test of Coefficients of Regression

Source of covariance	Degrees of freedom	Coefficient of regression
Between classes (years).....	$k - 1$	$A = \dfrac{\sum_i n_i(x_i - \bar{x})(y_i - \bar{y})}{\sum_i n_i(\bar{x}_i - \bar{x})^2}$
Within classes (years).....	$N - k$	$B = \dfrac{\sum_i \sum_j (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)}{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}$
Total.....	$N - 1$	$C = \dfrac{\sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y})}{\sum_i \sum_j (x_{ij} - \bar{x})^2}$

The variation between years can be determined in two ways, however, depending upon whether the between- or within-years regression coefficient is used as an adjusting factor. The total variation due to deviations from the regression line is

$$B = \sum_i \sum_j (y_{ij} - \bar{y})^2 - \frac{\left[\sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y})\right]^2}{\sum_i \sum_j (x_{ij} - \bar{x})^2} \tag{8-III-20}$$

That owing to within-years variation corrected for its regression coefficient is

$$C = \sum_i \sum_j (\bar{y}_{ij} - \bar{y}_i)^2 - \frac{\left[\sum_i \sum_j (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)\right]^2}{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2} \tag{8-III-21}$$

and that due to between-years variation corrected for its regression coefficient is

$$D = \sum_i n_i(\bar{y}_i - \bar{y})^2 - \frac{\left[\sum_i n_i(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})\right]^2}{\sum_i n_i(\bar{x}_i - \bar{x})^2} \tag{8-III-22}$$

whereas the difference between Eqs. (8-III-20) and (8-III-21) gives the variation due to between-years variation corrected for the within-classes regression coefficient. The analysis of covariance is given in Table 8-III-6.

The within-classes sum of squares divided by its degrees of freedom is the estimate of the variance used for testing. An F test may be applied first to the regression effect

$$F_{1,N-k-1} = \frac{(N - k - 1)A}{C} \tag{8-III-23}$$

Table 8-III-6. Analysis of Covariance

Source of variation	Degrees of freedom	Sum of squares
Regression	1	A
Deviations about regression line	$N - 2$	B
Within classes	$N - k - 1$	C
Between classes based on between-classes regression	$k - 2$	D
Between classes based on within-classes regression	$k - 1$	$B - C$

then to either or both measures of the between-years variation,

$$F_{k-2,N-k-1} = \frac{(N - k - 1)D}{(k - 2)C} \tag{8-III-24}$$

and

$$F_{k-1,N-k-1} = \frac{(N - k - 1)(B - C)}{(k - 1)C} \tag{8-III-25}$$

although the second test is more meaningful, if a class effect exists. To test whether the regression coefficients based on within- and between-years variation are significantly different, the difference of the two estimates of between-years variation may be used. This variation, which is the result of the difference in regression coefficients, may be used with one degree of freedom to compute an estimate of the variance, and this tested against the within-groups variance. The resulting F test

$$F_{1,N-k-1} = \frac{(N - k - 1)(B - C - D)}{C} \tag{8-III-26}$$

may be used to test whether the two regression coefficients are significantly different. The null hypothesis tested in this example is that there is no time variation of the relation of streamflow to precipitation, the assumption being that any time variation which does exist is the effect of reforestation.

II. ANALYSIS OF TIME SERIES

A. Introduction

1. Definition of Time Series. A *time series* is a sequence of values arrayed in order of their occurrence which can be characterized by statistical properties. The sequence of values is represented by $x(t_1)$, $x(t_2)$, $x(t_3)$, . . . , where $t_1 < t_2 < t_3 \cdots$. The daily hydrograph is a graphical representation of a time series of daily discharges. Other examples of hydrologic time series are the annual sequences of floods, low flows, and mean discharges. A time series may be a function of time explicitly or a function of any single variable which takes the place of time. Examples of sequences ordered by distance rather than time are the width and roughness of a stream channel as a function of distance.

Generally, it is possible to classify time series as being either of two types: *stationary* or *nonstationary*. Assume that a time series is divided into several segments and that a statistical parameter such as the mean is used to characterize the data within each section. If the expected value of the statistical parameter is the same for each section, the time series is said to be stationary. If the expected values are not the same, the time series is nonstationary. In stationary time series, absolute time is not important, and the series may be assumed to have started somewhere in the infinite past. However, in nonstationary time series, it is necessary to consider absolute time since the series cannot be assumed to have begun prior to the time of the initial observation.

2. Characteristics of Time Series. Most of the statistical methods used in hydrologic studies are based on the assumption that the observations are independently distributed in time. The occurrence of an event is assumed to be independent of all previous events. This assumption is not always valid for hydrologic time series. Observations of daily discharges do not change appreciably from one day to the next. There is a tendency for the values to cluster, in the sense that high values tend to follow high values and low values tend to follow low values. Thus the daily discharges are not independently distributed in time. The dependence between monthly discharges is less than that between daily discharges, and the dependence between annual discharges is less than that between monthly discharges. Thus the dependence between hydrologic observations decreases with an increase in the time base.

Hydrologic time series may be considered as composed of the sum of two components: a *random element* and a *nonrandom element.* A nonrandom element is said to exist when observations separated by k time units are dependent. If the values of x_i are linearly dependent upon the values of x_{i+k}, then the correlation between x_i and x_{i+k} may be taken as the measure of dependence. This correlation is referred to as the *kth-order serial correlation.*

The *serial correlation coefficient* is analogous to the product-moment correlation coefficient for two sets of data. If x_i and x_{i+k} are considered as two sets of data then the kth-order serial correlation coefficient is defined as

$$r_k = \frac{\frac{1}{N-k}\sum_{i=1}^{N-k} x_i x_{i+k} - \frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k} x_i\right)\left(\sum_{i=1}^{N-k} x_{i+k}\right)}{\left[\frac{1}{N-k}\sum_{i=1}^{N-k} x_i^2 - \frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k} x_i\right)^2\right]^{1/2}\left[\frac{1}{N-k}\sum_{i=1}^{N-k} x^2{}_{i+k} - \frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k} x_{i+k}\right)^2\right]^{1/2}} \qquad \text{(8-III-27)}$$

where N is the length of the time series. For $k = 0$, it follows that $r_0 = 1$, and for $k \geq 1$, $-1 \leq r_k \leq 1$.

If a time series is random, $r_k = 0$ for all values of $k \geq 1$. However, for a sample of finite size, computed values of r_k may differ from zero because of sampling errors. Since N is small for most hydrologic sequences, the sampling errors are very large, so that it is necessary to test the values of r_k to determine if they are significantly different from zero. A test of significance for r_k is given below.

An example of the computation of the first-order serial correlation coefficient r_1 for the low flows (annual minimum daily discharges) for Middle Branch Westfield River near Goss Heights, Mass., is shown in Table 8-III-7. The period of record is from 1913 to 1950 ($N = 38$). In the table, the columns headed x_i and x_{i+1} give the low-flow values from 1913 to 1949 and from 1914 to 1950, respectively.

In order to determine r_2, it is necessary that x_i and x_{i+2} denote the low-flow values from 1910 to 1953 and from 1912 to 1955, respectively. Similarly, by forming two sets of data, the values of r_3, r_4, etc., can be determined.

3. Properties of the Nonrandom Element. The nonrandom element may be composed of both a trend, or a long-term movement, and an oscillation about the trend. Both of these parts need not be present in a particular time series. The first step in analyzing a time series is to separate the nonrandom element from the random element.

Trend is usually thought of as a smooth motion of the series over a long period of time. For any given time series, the sequence of values will follow an oscillatory pattern. If this pattern indicates a more or less steady rise or fall, it is defined as a *trend.* However, no matter what the length of a time series is, it can never be stated

Table 8-III-7. Computation of the First-order Serial Correlation Coefficient

x_i	x_{i+1}	x_i^2	x^2_{i+1}	$x_i x_{i+1}$
1.6	0.4	2.56	0.16	0.64
0.4	0.4	0.16	0.16	0.16
0.4	2.9	0.16	8.41	1.16
2.9	5.4	8.41	29.16	15.66
5.4	5.0	29.16	25.00	27.00
5.0	7.5	25.00	56.25	37.50
7.5	5.0	56.25	25.00	37.50
5.0	14	25.00	196	70.00
14	15	196	225	210.00
15	2.5	225	6.25	37.50
2.5	3.0	6.25	9.00	7.50
3.0	9.1	9.00	82.81	27.30
9.1	4.0	82.81	16.00	36.40
4.0	6.8	16.00	46.24	27.20
6.8	14	46.24	196	95.20
14	4.0	196	16.00	56.00
4.0	4.7	16.00	22.09	18.80
4.7	4.8	22.09	23.04	22.56
4.8	2.1	23.04	4.41	10.08
2.1	4.6	4.41	21.16	9.66
4.6	6.0	21.16	36.00	27.60
6.0	5.5	36.00	30.25	33.00
5.5	2.5	30.25	6.25	13.75
2.5	6.9	6.25	47.61	17.25
6.9	10	47.61	100	69.00
10	2.6	100	6.76	26.00
2.6	4.6	6.76	21.16	11.96
4.6	2.5	21.16	6.25	11.50
2.5	4.4	6.25	19.36	11.00
4.4	4.5	19.36	20.25	19.80
4.5	4.8	20.25	23.04	21.60
4.8	11	23.04	121	52.80
11	3.5	121	12.25	38.50
3.5	3.6	12.25	12.96	12.60
3.6	2.6	12.96	6.76	9.36
2.6	1.8	6.76	3.24	4.68
1.8	3.6	3.24	12.96	6.48
Σ 193.6	195.6	1,483.84	1,494.24	1,134.70
$\Sigma/(N-1)$ 5.23	5.29	40.10	40.38	30.67

$$r = \frac{33.37 - (5.23)(5.29)}{[40.10 - (5.23)^2]^{\frac{1}{2}}[40.38 - (5.29)^2]^{\frac{1}{2}}} = 0.24$$

with certainty that an apparent trend is not part of a slow oscillation, unless the series ends.

An oscillatory pattern is often confused with a cyclical pattern. For a *cyclical time series,* the maximum and minimum values occur at equal intervals of time with constant amplitude. The random element, if present, tends to distort this pattern. In an oscillatory time series, the amplitude and the interval of time between maximum

and minimum values are distributed about mean values. A cyclical time series is oscillatory, but an oscillatory time series is not necessarily cyclical.

B. Trend Analysis

1. Use of Moving Averages. Various methods of removing trend are available. All the methods, however, are not fully understood as to how they affect the time series. The most general method involves the *fitting of a polynomial* to the data. This method has two principal objections: (1) the coefficients of the polynomial must be defined by high-order moments which are unreliable because of their large sampling errors since N is small, and (2) the coefficients of the polynomial must be recomputed each time a new value is added to the time series because they are based on the available data of the time series.

An alternative method of trend elimination is that of *moving averages*, which consists of finding a polynomial which will fit part of the record and using different polynomials for different parts of the record. This method permits the addition of new values without altering the previously fitted polynomials.

In order to remove the trend, it is necessary to smooth out irregularities in the time series. Assume that the observations $x_1, x_2, \ldots, x_N$ are taken at equal intervals of time. The method of moving averages consists of determining overlapping means of m successive weighted values. An example of moving averages of $m = 3$ is

$$\begin{aligned} y_2 &= \frac{b_1x_1 + b_2x_2 + b_3x_3}{3} \\ y_3 &= \frac{b_1x_2 + b_2x_3 + b_3x_4}{3} \\ &\cdots\cdots\cdots\cdots \\ y_{N-1} &= \frac{b_1x_{N-2} + b_2x_{N-1} + b_3x_N}{3} \end{aligned} \qquad \text{(8-III-28)}$$

The weights of the moving average, b_1, b_2, and b_3, are such that their sum equals 3. In general, for moving averages of m,

$$\sum_{i=1}^{m} b_i = m \qquad \text{(8-III-29)}$$

The weights may be either positive or positive and negative. A simple moving average refers to the case where each of the weights equals 1. Although a simple moving average tends to smooth out the data, it does not preserve the main features of the time series as well as a weighted moving average.

It is convenient to use odd values of m so that the computed values of y correspond in time to the middle value of the x's being averaged. A moving average of m applied to a sequence of N terms yields a sequence of $N - 2n$ terms, where $n = (m - 1)/2$. Thus, if $m = 3$, $n = 1$, so that one term is lost at the beginning and end of the time series. Although it is possible to use moving averages of $m = 2, 3, \ldots, N - 1$, it is necessary that m be small relative to N.

Generally, even a smooth trend obtained by the method of moving averages cannot be represented conveniently by a mathematical equation. If a mathematical trend is fitted to the data, a simple relation should be used unless logic indicates otherwise. The simplest mathematical expression is a straight line. However, a time series is apt to be such that a single linear trend cannot be used throughout the time of observation. In such cases, it is possible to use linear trends for portions of the time series.

After a trend has been established, it is possible to remove the trend from the data in one of several ways. One way is to take as a new variable the deviations about the trend line. It is necessary that these deviations constitute a stationary time series. This procedure of trend removal is widely used in hydrologic studies. With some

time series, such as tree-ring series, the deviations of the data about the trend line decrease with time [6]. Hence, in this case, dividing the deviations by their corresponding trend values yields a series which may be considered as stationary.

As an example of trend analysis, simple moving averages of $m = 3$ and $m = 5$ are applied to the low-flow data for the Middle Branch Westfield River near Goss Heights, Mass. These results are shown graphically in Fig. 8-III-4. Both moving averages indicate an apparent trend. This apparent trend may, however, be part of an oscillatory movement. With such a short series, it is difficult to prove that the apparent trend is significant, and not part of the oscillatory movement of the series.

2. Slutzky-Yule Effect. Assume that a time series has an oscillatory component about a trend. If a moving average is used to determine the trend, a long-period oscillation tends to be included as part of the trend. Oscillations which are comparable in period with the length of the moving average m, or even shorter, are damped out. The moving average also introduces an oscillatory movement into the random element. These consequences of the moving-average method are referred to as the *Slutzky-Yule effect* [7, 8].

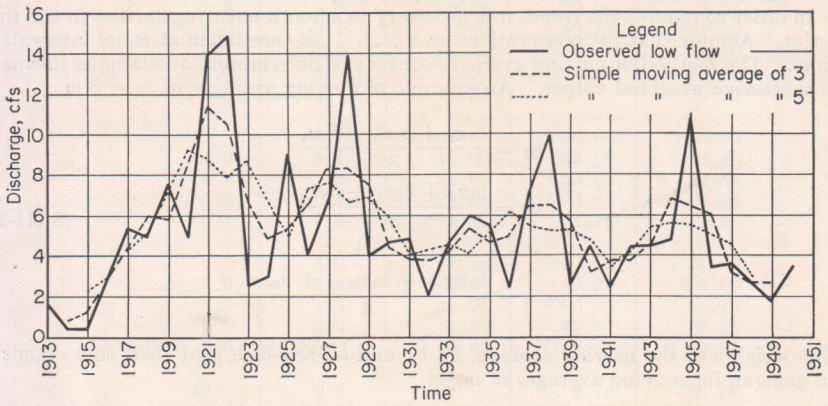

FIG. 8-III-4. Annual daily low flow for Middle Branch Westfield River near Goss Heights, Mass.

If a simple moving average is used, the variance of the induced oscillation is $1/m$ times the variance of the random element, and the average length of this induced oscillatory movement is $360°/\cos[(m-1)/(m+1)]$. If a weighted, instead of a simple, moving average is used, the Slutzky-Yule effect is magnified. Because of the Slutzky-Yule effect, care must be exercised in discussing the oscillatory character of a time series if its trend has been removed by means of the moving-average method.

C. Tests for Serial Dependence

1. Parametric Test of Significance. The variables x_i and x_{i+k} used to determine the serial correlation coefficients are actually parts of the same time series. The serial correlation coefficients cannot be tested for significance by means of the test for the ordinary product-moment correlation between two random series unless N is very large. A reliable test of significance must be based on small-sample theory.

Anderson [9] developed a test of significance based on a normal random time series which is circular. A *circular time series* is defined as a time series where the last value is followed by the first value so that the series repeats itself. As N tends to infinity, the confidence limits based on a circular time series converge to those based on an open time series. If N is small, only the low-order (k small) serial correlation coefficients may be tested for significance. Blackman and Tukey [10] recommend that k/N should not exceed 0.10. This rule appears to be satisfactory for deciding upon the

highest-order serial correlation which may be tested for significance by Anderson's method.

With respect to the first-order serial correlation coefficient r_1, Anderson showed that, for a normal random time series of N values, the expected value and the variance of r_1 are $-1/(N-1)$ and $(N-2)/(N-1)^2$, respectively. Since r_1 is nearly normally distributed, the confidence limits (CL) for a computed value of r_1 are given by

$$\mathrm{CL}(r_1) = -\frac{1}{N-1} \pm t_\alpha \frac{\sqrt{N-2}}{N-1} \tag{8-III-30}$$

where t_α is the standardized normal variate corresponding to the probability level $1-\alpha$.

If the computed value of r_1 lies within the confidence limits, then r_1 is considered to be insignificantly different from zero at the probability level $1-\alpha$. An insignificant r_1 is a necessary, but not a sufficient, condition for deciding that a time series is random. In order that a time series be regarded as random, it is necessary that r_k for $k \geq 1$ be insignificant. Because of sampling errors, the serial correlation coefficients for some values of k will be found to be significant even if the observed time series is a sample from a random time series. However, the number of significant serial correlation coefficients should not be greater than that expected by chance from the total number of serial correlation coefficients tested. The reader is referred to Anderson's paper [9] for tests of significance of r_k's where $k \geq 2$.

For the low-flow data for Middle Branch Westfield River near Goss Heights, Mass., r_1 is 0.24. At the 95 per cent level, $t_\alpha = 1.96$. Thus the 95 per cent confidence limits are 0.30 and -0.34. Since r_1 is below the upper confidence limit, it may be regarded as insignificant at the 95 per cent level.

2. Nonparametric Tests of Significance. A nonrandom time series has an oscillatory component. The observed values in a purely random time series fluctuate erratically about some mean value. The fact that a time series exhibits more or less erratic fluctuations suggests that the number of times that the values are above or below a given value is indicative of the randomness or nonrandomness of the time series.

A nonparametric method of determining if a time series is random is that of the *median cross*. For a time series of N values, the median is determined. From the sequence of N values, the number of times that the series crosses the median is determined. Let this number be denoted by n. The expected value and variance of n are $(N-1)/2$ and $(N-1)/4$, respectively. Since n is nearly normally distributed, it is possible to test if n is significantly different from $(N-1)/2$ by

$$t = \frac{n - (N-1)/2}{\sqrt{(N-1)/4}} \tag{8-III-31}$$

If the absolute value of t is greater than the absolute value of t_α, the normal deviate at the probability level $1-\alpha$, the time series is regarded as nonrandom. If the contrary is true, the time series is considered to be random.

Another nonparametric method for determining if a time series is random is the *turning-point test*. A turning point is associated with a value x_i, where either $x_{i+1} > x_i < x_{i-1}$ or $x_{i+1} < x_i > x_{i+1}$. Let m denote the number of turning points. The expected value and the variance of m are $2(N-2)/3$ and $(16N-29)/90$, respectively. Since m is nearly normally distributed,

$$t = \frac{m - 2(N-2)/3}{\sqrt{(16N-29)/90}} \tag{8-III-32}$$

The significance or nonsignificance of m is determined in the same manner described for n in the median-cross test. By using the median-cross and the turning-point tests to determine if the Middle Branch of Westfield River data are random, the t values are

0.82 and 0, respectively. Since both t values are less than 1.64, the data are assumed to be random. It is interesting to note that the nonparametric tests indicate randomness and the parametric test indicates nonrandomness. The parametric test is stronger and more reliable than the nonparametric tests.

D. Generating Processes

1. Definition. A *generating process* is the manner by which the causal forces act to produce a time series. Some processes can be expressed mathematically, in which case it is possible to determine directly the various statistical characteristics of the time series. Often a time series is approximated by a certain process. The choice of the process is based upon how well the mathematical structure of the process conforms to the physical characteristics of the time series. The processes which have been used in hydrologic studies are (1) the moving average, (2) the sum of harmonics, and (3) the autoregression.

2. Moving-average Process. The moving-average process may be expressed as

$$x_i = b_0 + b_1 y_i + b_2 y_{i-1} + \cdots + b_m y_{i-(m-1)} \tag{8-III-33}$$

where y is a random variable and m is the extent of the moving average. Equation (8-III-33) may be taken as the model representing the relation between annual runoff x and annual effective precipitation y, where m is the extent of the carryover due to the water-retardation characteristics of the river basin. For such a model, the weights $b_0, b_1, \ldots, b_m$ must all be positive and sum to unity. By virtue of the moving average on the y's, the generated series x is not random. The serial correlation coefficients for the x's are given by Wold [11]:

$$r_k = \frac{\sum_{i=0}^{m} b_i b_{i+k}}{\sum_{i=0}^{m} b_i^2} \qquad 0 \le k \le m - 1 \tag{8-III-34}$$

where

$$r_k = 0 \qquad k \ge m \tag{8-III-35}$$

It should be noted in Eqs. (8-III-34) and (8-III-35) that dependence between values does not extend throughout the time series. Values separated by m or more time units are independent.

3. Sum-of-harmonics Process. A simple model of the generating process of the sum of harmonics is

$$x_i = A \sin \theta i + y_i \tag{8-III-36}$$

where A and θ are the amplitude and period of cyclicity, respectively, and y is a random component. Equation (8-III-36) may be taken as a model representing seasonal discharges. For example, if the x's denote monthly discharges and if there is a distinct period of high flow and of low flow, then $\theta = \pi/6$, and i would represent the months from 1 to 12. The generated x's are nonrandomly distributed in time. The serial correlation coefficients are defined by

$$r_k = \frac{A^2}{2 \text{ Var } (x)} \cos \theta k \tag{8-III-37}$$

where the variance of x, Var (x), is defined by

$$\text{Var } (x) = \frac{A^2}{2} + \text{Var } (y) \tag{8-III-38}$$

Equation (8-III-36) is a special case, where only one harmonic is involved. It is often argued that there are hidden periodicities in hydrologic data of annual sequences.

The periodicities are called hidden since the superposing of many series of different harmonics yields a series which is seemingly random. A recent study involving the search for hidden periodicities in rainfall has been made by Abbott [12].

4. Autoregression Process. An *autoregression process* is used in hydrologic studies for representing sequences whose nonrandomness is due to storage in the basin (groundwater, lake, or channel storage). There are many autoregressive models; however, the first-order process is defined as

$$x_{i+1} = r_1 x_i + \epsilon_{i+1} \tag{8-III-39}$$

where r_1 is the first-order serial correlation coefficient for the x's and ϵ is a random component. This process is often referred to as the *first-order Markov process*. For this process, the serial correlation coefficients are given by

$$r_k = r_1{}^k \tag{8-III-40}$$

If r_1 is positive, then all values of r_k are positive and $r_1 > r_2 > \cdots$. If r_1 is negative, then r_k is positive for even values of k and negative for odd values of k. The absolute value of r_k decreases as k increases.

5. Correlograms. A *correlogram* is a graphical representation of the r_k's as a function of k where the values of r_k are plotted as ordinates against their respective values of k as abscissas. In order to reveal the features of the correlogram better, the plotted points are joined each to the next by a straight line.

From Eqs. (8-III-34) and (8-III-35), it is seen that the correlogram for a moving average may oscillate, depending upon the b's, but it will vanish for all values of $k > m$. It is seen from Eq. (8-III-37) that the correlogram for a harmonic process will oscillate with period θ and amplitude $A^2/2$ Var (x). The period of oscillation of the correlogram is the same as that for the time series itself. For the autoregression process, it is seen by Eq. (8-III-40) that, if r_1 is positive, the correlogram will decrease monotonically from $r_0 = 1$ to $r_\infty = 0$. If r_1 is negative, the correlogram will oscillate with period unity above the abscissa with a decreasing but nonvanishing amplitude.

The correlogram provides a theoretical basis for distinguishing among the three types of oscillatory time series. From a set of data, the serial correlation coefficients can be determined and the correlogram can be constructed. The shape of the correlogram is indicative of the generating process in the manner described above. In practice, the number of observations forming a sequence is small, so that observed correlograms always show less damping than theoretical correlograms because the observed serial correlation coefficients are inflated by sampling errors. Thus one cannot easily discern what the generating process is simply by observing the correlogram.

At present there is no adequate small-sample test for distinguishing among the generating processes. Quenouille [13] has developed tests of significance of the correlograms for various autoregressive models. However, these tests are based on the length of sequence being very large.

E. Effect of Serial Correlation

1. Estimation of the Variance. Serial correlation represents a tendency for fluctuations about the mean to perpetuate themselves. In nonrandom hydrologic time series, r usually is positive, so that high values tend to follow high values and low values tend to follow low values. Thus values near x_i yield little new information concerning the true fluctuation of the events about the mean. The amount of information which is furnished varies inversely with r_1. If $r_1 = 0$, then each successive event furnishes new information. If $r_1 = 1$, then each event contains all the available information, so that each successive event furnishes no new information. Thus, for a given nonrandom time series of length N, the information given by the N values is equal to that given by a random time series of length N', where $N' < N$. N' is often referred to as the *effective length of record*. With respect to a given value of N, the larger r_1 is, the smaller N' is.

For a sequence of N' events, taken from a random time series, the unbiased estimator of the variance is given by

$$\hat{\sigma}^2 = \frac{N'}{N'-1}\left[\sum_{i=1}^{N'}\frac{(x_i - \bar{x})^2}{N'}\right] = \frac{N'}{N'-1}S^2 \qquad \text{(8-III-41)}$$

so that the expected value of $\hat{\sigma}^2$ is σ^2. For a sequence of N events, from a time series generated by a first-order Markov process, the unbiased estimator of the variance is given by

$$\hat{\sigma}^2 = \left[1 - \frac{1 - r_1^2}{N(1-r_1)^2} + \frac{2r_1(1 - r_1^N)}{N^2(1-r_1)^2}\right]^{-1} S^2 \qquad \text{(8-III-42)}$$

If $r_1 = 0$ and $N = N'$, Eq. (8-III-42) reduces to Eq. (8-III-41). By equating Eqs. (8-III-41) and (8-III-42), it is possible to determine N' for given values of r_1 and N. A graphical procedure facilitates the determination of N'. In Fig. 8-III-5 a family of curves is shown for $S^2/\hat{\sigma}^2$ versus N as a function of r_1. As N tends to infinity,

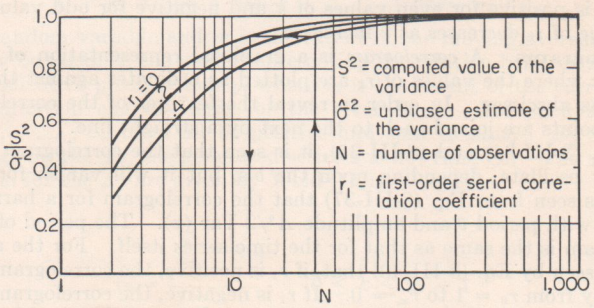

FIG. 8-III-5. Relation between $s^2/\hat{\sigma}^2$ and N as a function of r_1.

$S^2/\hat{\sigma}^2$ tends to unity for all values of r_1. The larger r_1 is, the slower is the rate of convergence. For a given sequence, N is known and r_1 can be determined by Eq. (8-III-27). Thus, starting with the value of N on the abscissa, a vertical line is drawn upward until it intersects the curve corresponding to r_1. From this point of intersection, a horizontal line is drawn to the left until it intersects the curve for $r_1 = 0$. From this point of intersection, a vertical line is drawn downward to the abscissa scale to determine N'. An example is shown in Fig. 8-III-5. It is assumed that $N = 30$ and $r_1 = 0.4$, so that $N' = 12$.

2. Correlation and Regression Analyses. In hydrologic studies, one is often interested in whether or not two or more variables are associated (see also Sec. 8-II). Extensive theory has been developed for determining the degree of association by means of the correlation coefficient when each variable is randomly distributed. The correlation between two nonrandom time series can be determined, but cannot be tested for significance in the same manner as the correlation between two random variables. If N pairs of observations are available, each observation cannot be considered as contributing new information about the correlation if the two time series are nonrandom.

The test of significance for the correlation between two random variables is based on the t test [2], where

$$t = r\sqrt{\frac{n}{1 - r^2}} \qquad \text{(8-III-43)}$$

Where r denotes the correlation between the two variables, and n, which is equal to $N - 2$, where N is the number of pairs of observations, denotes the number of degrees of freedom.

In order to test the correlation between two nonrandom time series for significance, it is necessary to replace n by the effective number n' of degrees of freedom. From Bartlett's work [14] it can be shown that, for very large sample sizes,

$$n' = \frac{n}{1 + 2r_1r_1' + 2r_2r_2' + \cdots} \qquad \text{(8-III-44)}^1$$

where r_1, r_2, . . . are the serial correlation coefficients for one of the time series and r_1', r_2', . . . are the serial correlation coefficients for the other time series. To determine n', it is necessary to compute the serial correlation coefficients for many orders. This is very laborious, and because of the sampling errors associated with the serial correlation coefficients, it is not possible to determine n' accurately.

A useful formula for n' is obtained by considering each of the time series to be generated by a first-order Markov process. For this process, $r_k = r_1^k$ and $r_k' = (r_1')^k$. By using these relations in Eq. (8-III-44), n' becomes

$$n' = n\left(\frac{1 - r_1r_1'}{1 + r_1r_1'}\right) \qquad \text{(8-III-45)}$$

By Eqs. (8-III-44) and (8-III-45), it can be seen that if either time series is random, $n' = n$. This is consistent with the fact that each observation of the random time series contributes completely new information on the value of the correlation between the random and nonrandom time series.

Regression analysis is used in hydrologic studies to establish the relation between a given variable (referred to as the dependent variable) and one or more variables (referred to as the independent variables). The classical theory of regression analysis is based on the assumption that each variable is randomly distributed. If both the dependent and independent variables are time series, it is necessary to determine if the variables are random or not. An ordinary regression analysis involving time series is valid under two conditions: (1) if either the dependent or the independent variables are random, and (2) if the deviations from the line of regression are serially independent.

The serial dependence of the deviations from the line of regression may be determined by means of serial correlation coefficients. However, in testing the serial correlations of the deviations for significance, it is necessary to allow for the fitting of the regression line. No exact test of significance is available. An approximate test of significance is given by Durbin and Watson [15]. This test, summarized in Table 8-III-8, is based on giving correction terms for determining the effective number of deviations.

Table 8-III-8. Corrections to Number of Observations

Number of independent variables	Level of significance	
	$P = 0.05$	$P = 0.02$
1	(−1) (20)	(−1) (16)
2	(−5) (35)	(−5) (30)
3	(−10) (60)	(−10) (50)
4	(−15) (100)	(−15) (75)

Table 8-III-8 is used in the following manner. Assume that a regression analysis involving two time series is based on $N = 20$ pairs of observations. The serial correlation of the deviations from the line of regression may be tested for significance by

[1] See also Eq. (8-II-27).

Eq. (8-III-43), where $N = 20 - 1 = 19$ and $N = 20 + 20 = 40$. At the 95 per cent confidence level t is approximately 2. Thus, for $N = 19$, $r = 0.46$, and for $N = 40$, $r = 0.31$. If the computed serial correlation is greater than $r = 0.46$, then the serial correlation is significantly greater than 0 at the 95 per cent level. A computed serial

Table 8-III-9. Data for Studying the Effect of Serial Correlation on Correlation and Regression Analyses

Year	x	y	ϵ_x	ϵ_y	y'
1913	21	1.6	.....	.....	−3.03
1914	22	0.4	13.69	−0.32	−4.43
1915	20	0.4	12.34	0.22	−4.03
1916	45	2.9	38.04	2.72	−6.53
1917	30	5.4	29.34	4.09	−1.03
1918	32	5.0	21.56	2.55	−1.83
1919	24	7.5	12.86	5.23	2.27
1920	25	5.0	16.65	1.60	−0.43
1921	31	14	22.30	11.73	7.37
1922	33	15	22.21	8.66	7.97
1923	20	2.5	8.52	−4.30	−1.93
1924	18	3.0	11.04	1.87	−1.03
1925	16	9.1	9.74	7.74	5.47
1926	14	4.0	8.43	−0.12	0.77
1927	20	6.8	15.13	4.99	2.37
1928	43	14	36.04	10.92	4.97
1929	20	4.0	5.04	−2.34	−0.43
1930	15	4.7	8.04	2.89	1.27
1931	14	4.8	8.78	2.67	1.57
1932	15	2.1	10.13	−0.07	−1.33
1933	21	4.6	15.78	3.65	−0.03
1934	23	6.0	15.69	3.92	0.97
1935	29	5.5	21.00	2.78	−0.73
1936	20	2.5	9.91	0.01	−1.93
1937	26	6.9	19.04	5.77	1.27
1938	24	10	14.95	6.87	4.77
1939	22	2.6	13.65	−1.93	−2.23
1940	25	4.6	17.34	3.42	−0.83
1941	14	2.5	5.30	0.42	−0.73
1942	14	4.4	9.13	3.27	1.17
1943	19	4.5	14.13	2.51	0.27
1944	21	4.8	14.39	2.76	0.17
1945	41	11	33.69	8.83	2.37
1946	39	3.5	24.73	−1.48	−4.73
1947	30	3.6	16.43	2.01	−2.83
1948	19	2.6	8.56	0.97	−1.63
1949	18	1.8	11.39	0.62	−2.23
1950	21	3.6	14.74	2.78	−1.03

correlation is nonsignificant if it is less than $r = 0.31$. If a computed serial correlation lies between these two values, then there is doubt about the significance at the 95 per cent level, since it is not certain if the 95 per cent level is reached [13].

If the residuals are serially uncorrelated, an ordinary regression analysis is valid. However, if the residuals are serially correlated, it is necessary to take this fact into

account in the regression analysis. Quenouille [13] suggests that this may be done in one of two ways. The first way consists in calculating the deviations from a serial regression of the dependent variable upon previous values of itself and using these deviations in a regression analysis on the independent variables and their previous values. The second way is to make a regression analysis of the dependent variable upon the independent variables and upon previous values of the dependent variables and itself. The first method may be used to predict the random variation in the dependent variable from the independent variables. By the second method, values of the dependent variable may be estimates from the independent variables and previous observations.

In order to clarify the above discussions an example is given. In Table 8-III-9, under the columns headed by x and y, respectively, the correlation between x and y, $r(xy)$, is 0.47, and the first-order serial correlations for the x and y series are $r_1(x) = 0.35$ and $r_1(y) = 0.24$, respectively. By assuming that both series are generated by a first-order Markov process, then the effective number of degrees of freedom is, according to Eq. (8-III-40), 32. By using Eq. (8-III-43), $t = 2.91$. Since the value of t at the 95 per cent level, 2.056, is less than 2.91, the correlation between x and y is significant.

The equation for the regression of y on x is

$$y = 0.31 + 0.20x \tag{8-III-46}$$

The deviations from this regression are given in Table 8-III-9 under the column headed by y'. The first-order serial correlation of the deviations, $r_1(y')$, is 0.333. By using the corrections, given in Table 8-III-8, to the number of observations and applying Eq. (8-III-43), it is seen that $r_1(y')$ is significant at the 95 per cent level. Since the x, y, and y' series are nonrandom, an ordinary regression analysis cannot be made.

The deviations from the serial regression of the independent variable upon previous values of itself are given by

$$x_{i+1} - 0.453x_i = (\epsilon_x)_{i+1} \tag{8-III-47}$$

These deviations are given in Table 8-III-9 under the column headed by ϵ_x. Similarly, the deviations from the serial regression of the dependent variable upon previous values of itself can be determined. These deviations are given in Table 8-III-9 under the column headed by ϵ_y. The first-order serial correlation coefficients for these two sets of deviations are $r_1(\epsilon_x) = 0.226$ and $r_1(\epsilon_y) = -0.176$. Both of these coefficients are insignificant at the 95 per cent level. Thus the ϵ_x and ϵ_y series may be considered as random.

If the first method of accounting for the serial correlation is used, it is necessary to determine the regression of ϵ_y on ϵ_x. This regression gives

$$(\epsilon_y)_{i+1} = -0.782 + 0.232(\epsilon_x)_{i+1} \tag{8-III-48}$$

so that y_{i+1} might be predicted using

$$y_{i+1} = -0.782 + 0.348y_i + 0.232x_{i+1} - 0.105x_i \tag{8-III-49}$$

If the second method is used, a multiple regression of y_{i+1} on y_i, x_{i+1}, and x_i must be carried out. This regression gives

$$y_{i+1} = 1.210 + 0.602y_i + 0.244x_{i+1} - 0.205x_i \tag{8-III-50}$$

In order to use either Eq. (8-III-49) or Eq. (8-III-50), it is necessary that the deviations from the line of regression be serially independent.

III. REFERENCES

1. Hoel, P. G.: "Introduction to Mathematical Statistics," John Wiley & Sons, New York, 1954.

2. Weatherburn, C. E.: "A First Course in Mathematical Statistics," Cambridge University Press, London, 1952.
3. Bartlett, M. S.: Properties of sufficiency and statistical tests, *Proc. Roy. Soc., London*, ser. A, vol. 160, pp. 268–282, 1937.
4. Characteristics of low flow volume-duration-frequency statistics, *U.S. Army Corps Engrs. Tech. Bull.* 1, 1960.
5. Double-mass curves, *U.S. Geol. Surv. Water-Supply Paper* 1541-B, 1960.
6. Schulman, Edmund: "Dendroclimatic Changes in Semiarid America," University of Arizona Press, Tucson, Ariz., 1954, pp. 29–30.
7. Slutzky, Eugen: The summation of random causes as the source of cyclic processes, *Econometrika*, vol. 5, pp. 105–146, 1937.
8. Yule, G. U.: On the time series problem, *J. Roy. Statist. Soc.*, vol. 84, pp. 497–526, 1921.
9. Anderson, R. L.: Distribution of the serial correlation coefficient, *Ann. Math. Statist.*, vol. 8, no. 1, pp. 1–13, 1941.
10. Blackman, R. B., and J. W. Tukey: "The Measurement of Power Spectra," Dover Publications, Inc., New York, 1959.
11. Wold, Herman: "A Study in the Analysis of Stationary Time Series," Almqvist and Wiksell, Stockholm, 1954.
12. Abbott, C. G.: A long-range forecast of United States precipitation, *Smithsonian Inst. Misc. Collections*, vol. 139, no. 9, 1960.
13. Quenouille, M. H.: "Associated Measurements," Academic Press Inc., New York, 1952, pp. 165–187.
14. Bartlett, M. S.: Some aspects of the time-correlation problem in regard to tests of significance, *J. Roy. Statist. Soc.*, vol. 98, pp. 536–543, 1935.
15. Durbin, James, and G. S. Watson: Testing for serial correlation in least squares regression, *Biometrika*, vol. 37, pp. 409–428, 1950.