STATISTICAL MODELLING - BASIC STATISTICS

1.0 INTRODUCTION

In many problems in hydrology, the data consists of measurements on a single random variable; hence we must deal with univariate analysis and estimation. The objective of univariate analysis is to analyse measurements on the random variable, which is called sample information, and identify the statistical population from which we can reasonably expect the sample measurements to have come. After the underlying population has been identified, one can make probabilistic statements about the future occurrences of the random variable, this represents univariate estimation. It is important to remember that univariate estimation is based on the assumed population and not the sample, the sample is used only to identify the population. Statistical analysis of univariate data requires several components:

i) Select a model: here this is the probability distribution function or p.d.f.

ii) Obtain a sample: this consists of a set of n independent observations on or measurements of a random variable.

iii) Fit the selected distribution model with the sample data; for most applications the graphical, method, least square method, method of moments or method of maximum likelihood are generally used.

iv) Perform goodness of fit tests for selecting the best fit distribution for the assumed population; use either the graphical method or the analytical statistical tests.

v) Use the best fit frequency distribution model to make probability statements about the likelihood of occurrence of values of the random variables.

In case the hydrologic variable is not random, the statistical analysis discussed above can not be used to make the predictions at different probability levels. Test of independence may be performed to examine whether the hydrologic variable is random or time dependent. In order to describe the time dependent characteristics of a hydrologic variable, a time series analysis is generally carried out. A time series model can be formulated and calibrated from analysing the data on hydrologic variable in which time is considered to be an independent variable. Future values of the time dependent variable may be predicted from the calibrated time series model. Methods used to analyse time series are also being used to analyse spatial data of hydrologic systems.

In statistical analysis of multivariable data, the functional forms of the relationships are studied. Linear regression analysis may be used as one of the ways to develop the suitable form of the multiple variable models wherein a dependent variable takes on values caused by variations in one or more independent or predictor variables. Such models are found to be more useful for predicting the hydrological variables using the predictor variables if they are known.

In this lecture, statistical analysis of univariate and multivariate data has been discussed. Some of the important statistical parameters are described along with the definitions of some statistical terms. Furthermore, various theoretical frequency distributions, commonly used for statistical modelling of the random hydrologic variable, are also discussed. Simple linear regression and/or multiple linear regression analysis are used to study the functional forms of the relationships consider-ing the multivariable data. This lecture note also covers the procedures involved in the simple linear

regression and multiple linear regression analysis. For further details about the various aspects of statistical analysis, one may refer the literature given in the end of the lecture note as a bibliography.

2.0 DEFINITIONS OF SOME STATISTICAL TERMS

In this section of the lecture some of the important statistical terms are defined:

Population: A population is a collection of persons or objects e.g. (i) the pupils in a school, the workers in a factory, the people in a country, (ii) motor cars produced in a factory. Each unit of the population has many different possible attributes associated with it. These attributes might be: (i) height, volume or weight which are measurable on a scale, or (ii) colour, condition which may not be numerically measurable.

Sample data: Sample data are available data from the observation of an event.

Random events: Events whose occurrence is not influenced by the occurrence of the same event earlier.

Probability density function: Probability density function (P.D.F.) is the probability of occurrence of an event.

Cumulative density function: Cumulative density function (C.D.F.) is the probability of occurrence of all the events that are equal to or less than an event.

Probability paper: A probability paper is a special graph paper on which the ordinate usually represents the magnitude of the variate and the abscissa represents the probability P, or the return period T. The ordinate and abscissa scales are so designed that the distribution plots more nearly a straight line permitting better definition of the upper and lower parts of the frequency curve. The probability paper is used to linearize the distribution so that data to be fitted appear close to the straight line.For example, the extreme value and the log normal probability papers are used for linearization of the extreme value and log normal distribution.

Plotting position: Determining the probability to assign a data point is commonly referred to as determining its plotting position.

3.0 SAMPLE STATISTICS

In any analysis of statistical data in general and of hydrolytic data in particular, certain calculations are usually made in order to determine some of the basic properties inherent in the data. For instance, the sample mean and variance are two statistics defining the most important characteristics of a given set of statistical data. In general sample statistics provide the basic information about the variability of a given data set. The most useful sample statistics measure the following characteristics:

(i) the central tendency or value around which all other values are clustered,
(ii) the spread of the sample values around mean,
(iii) the asymmetry or skewness of the frequency distribution, and
(iv) the flatness of the frequency distribution.

These statistical properties are determined by sample statistics as described below:

3.1 Measure of Central Tendency or Measures of Location

In Statistics various measures of location are described. Some of the important measures include the following:

(i) Mid range:

It is the average of the minimum and maximum values of the sample (or population) i.e.

$$\text{Mid range} = \frac{\textit{Minimum value} + \textit{Maximum value}}{2.0} \quad \ldots.(1)$$

(ii) Mode:

It is the value in the sample (or population) having most frequent occurrences i.e.

$$\text{Mode} = \text{Most frequent value} \quad \ldots.(2)$$

(iii) Median:

It is the middle value of the ranked values for a sample (or population) i.e.

$$\text{Median} = \text{Middle value of the ranked values} \quad \ldots.(3)$$

(iv) Mean:

If X_1, X_2 X_3 .... X_n represent a sequence of observations, the mean of this sequence is determined as the ratio of sum of values and number of values:

$$\bar{X} = 1/N \sum_{i=1}^{N} X_i \quad \ldots.(4)$$

Here $\bar{X}$ represent the sample mean, Population mean is generally represented by μ.

3.2 Measure of Dispersion or Variation

Some of the important measures of dispersion or variation include:

(i) Range:

It is the difference between maximum and minimum values i.e.

$$\text{Range} = \text{Max. Value} - \text{Min. Value} \quad \ldots(5)$$

(ii) Interquartile Range:

It is defined as $I_3 - I_1$, where I_1 is the value separating the lowest quarter of the ranked data from the second quarter and I_3 separates the third and fourth quarters of the ranked data. In other words, interquartile range between the 25% and 75% cumulative frequency values contains 50% of the values.

(iii) Mean Deviation:

Dispersion about the arithmetic mean is mean deviation. Thus,

$$\text{Mean Deviation} = \frac{\sum_{i=1}^{N} |X_i - \bar{x}|}{N} \quad \ldots.(6)$$

(iv) Variance:

Variance represents dispersion about the mean. Mathematically for sample it is expressed as:

$$\text{Variance} = S^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{x})^2}{N} \quad \ldots.(7)$$

(v) Standard deviation:

The unbiased estimate of population standard deviation (S) from the sample is given as the square root of the variance i.e.

$$S = [\, 1/(N-1) \sum_{i=1}^{N} (X_i - \bar{X})^2 \,]^{1/2} \quad \ldots.(8)$$

(vi) The coefficient of variation C_V is a dimensionless dispersion parameter and is equal to the ratio of the standard deviation and the mean:

$$C_V = S/\bar{X} \quad \ldots.(9)$$

This coefficient is extensively used in hydrology particularly as a regionalisation parameter.

The range and mean deviation have the same units (dimension) as the original data. The variance has the square of the units of the original data and hence can not be directly compared with the data. Therefore, the standard deviation is used because its dimensions are that of the data.

In many samples of hydrological data, especially in flood hydrology the largest value is very much larger than the second largest. Therefore the range R might not be a good indicator of the scatter inherent in the data as a whole.

The mean deviation is a good measure of spread but can not be handled easily in mathematical statistics because of the absolute value sign while the same applies to the interquartile range. The variance is more easily handled mathematically and holds a prominent place. The interquartile range is easy to evaluate but is very difficult from a mathematical point of view and hence is not much used even though it is quite good at describing spread.

3.3 Measures of Symmetry

If the data are exactly symmetrically displaced about the mean than the measure of symmetry should be zero. If the data to the right of the mean (larger) are more spread out from the mean than those on the left then, by convention, the asymmetry is positive and vice versa for negative asymmetry.

(i) Interquartile measure of asymmetry:

The interquartile measure of asymmetry (I_{as}) is defined as:

$$I_{as} = |I_3 - I_2| - |I_1 - I_2| \quad \ldots.(10)$$

where, I, I_2 and I_3 are the lower quartile, median and upper quartile respectively.

(ii) Third Central Moment:

The third moment of the sample data about the mean is given by:

$$M_3 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^3 \qquad \ldots.(11)$$

If the data are symmetrical this is zero. Otherwise it is positive or negative.

(iii) Skewness Coefficient:

The skewness coefficient or coefficient of skewness represents a non-dimensional measure the asymmetry of the frequency distribution of the data. An unbiased estimate of the coefficient is given by:

$$C_s = \frac{N\sum_{i=1}^{N}(X_i - \bar{X})^3}{(N-1)(N-2)S^3} \qquad \ldots.(12)$$

The skewness coefficient has an important meaning since it gives indication of the symmetry of the distribution of the data. Symmetrical frequency distributions have very small or negligible sample skewness coefficient C_s, while asymmetrical frequency distributions have either positive or negative coefficients. Often a small value of C_s, indicates that the frequency distri-bution of the sample may be approximated by the normal distribution function since $C_s = 0$ for this function.

Note that because of the third Central Moment has dimension equal to the cube of the data, it is not of direct use. It also depends on the units of the original data. The coefficient of skewness does not have this disadvantage and is therefore preferred. The interquartile measure of symmetry (I_{as}) is also not dimensionless.

3.4 Measures of Peakedness or Flatness

The Kurtosis coefficient measures the peakedness or the flatness of the frequency distribution near its centre. An unbiased estimate of this coefficient is given by:

$$C_k = \frac{N^2\sum_{i=1}^{N}(X_i - \bar{X})^4}{(N-1)(N-2)(N-3)S^4} \qquad \ldots.(13)$$

A related coefficient called the excess coefficient denoted by E is defined by:

$$E = C_k - 3 \qquad \ldots.(14)$$

Positive values of E indicate that a frequency distribution is more peaked around its centre than the normal distribution. Frequency distribution is known as LEPTOKURTIC. The negative

values of E indicate that a given frequency distribution is more flat around its centre than the normal. Frequency distribution is known as PLATYKURTIC.

The kurtosis for a normal distribution is 3.

Normal distribution is said to be MESOKURTIC. Both kurtosis and excess coefficient are seldom used in statistical hydrology.

3.5 Extreme Data Values

For frequency analysis, the data items should be random and independent. The observed data of various hydrological variables generally exhibit some time dependence. Therefore, as such the frequency analysis should not be carried out with such records. It has been observed that the extreme data values fulfill the requirement of independence and randomness. Therefore, these extreme values are used in frequency analysis.

The highest and lowest values may be obtained from the record of hydrological variables considered for a specific duration. For example: (i) annual maximum peak flood series may be obtained from the annual flood records of various years, (ii) annual low flow value for d-days duration may be obtained considering the lowest d-days duration flows for the year.

4.0 STANDARD ERRORS OF SAMPLE STATISTICS

Because of the short period of record the statistics calculated from the sample are only estimates of the true or population values which would be calculated if an infinitely large samples were available. The reliability of the statistics calculated from the sample can be judged from the standard errors of the estimate (SEE). Statistical Theory states there is about 68% probability that the true of population value of each statistic is within one standard error of estimate of the value calculated from the available data.

The standard errors of mean, standard deviation and coefficient of skewness are given below:

$$Se\,(\bar{X}) = S/\sqrt{N} \qquad \ldots.(15)$$

$$Se\,(S) = S/\sqrt{2N} \qquad \ldots.(16)$$

$$Se\,(C_s) = \sqrt{6N(N-1)/[(N-2)(N+1)(N+3)]} \qquad \ldots.(17)$$

The standard error of estimate for each moment becomes smaller as a longer length of record becomes available for use in the analysis.

If the Frequency Analysis is to provide useful answers, it must start with a data that is relevant, adequate and accurate.

As a preliminary step the basic data should be screened and adjusted to remove, as far as possible, any non-conformities that may exist. The following are some of the important considerations:

(i) Effect of man made changes in the regime of flow should be investigated and adjustment be made as required.

(ii) Changes in the stage discharge relation render stage records non homogeneous and unsuitable for frequency analysis studies. It is therefore preferable to work with discharges and if stage frequencies are required, refer the results to the most recent rating.

(iii) Any useful information contained in data publications and manuscripts should be made use of after proper scrutiny.

5.0 GRAPHICAL PRESENTATION OF GROUPED DATA

For the graphical presentation of grouped data in the form of histograms and cumulative histograms of frequency (or relative frequency or probability), a frequency table is prepared. In this table the range of the data variable is divided into a number of intervals of convenient size and the number of frequency f of values occurring in each interval is entered alongside. This table provides a very valuable summary. If the class intervals are made very large the table is made compact but loses detail. If the intervals are too small the table may be too bulky and not succinct enough. For the choice of class interval, the following criteria may be considered as guideline:

(a) Brooks and Carruthers rough guide:

$$\text{No of classes} \leq 5 \log (\text{no. of values}) \qquad \ldots.(18)$$

(b) Charlier's rule of thumb:

$$w = \frac{Max.\ value \quad - Min.\ value}{20} \qquad \ldots.(19)$$

where, w = size of class interval. Number of classes generally 15 to 25.

The frequency table can be prepared using the following steps:

(i) Order the variable (X_i) in increasing or decreasing order of magnitude.

(ii) Select a number of class interval (NC) and the size of the class interval ΔX. In this regard the guidelines given above may be followed.

(iii) Divide the ordered observations X_i into NC intervals (or groups).

(iv) Determine the absolute frequency n_j by counting the observations that fall within the j^{th} class interval for j=1....NC.

(v) Determine the corresponding relative frequencies as n_j/n, j=1....NC.

(vi) Compute the cumulative relative frequencies F_j, j = 1, ... NC. These cumulative frequencies approximate the probabilities as:

$F_j = F(X \leq x)$ if order is increasing, or

$F_j = F(X > x)$ if order is decreasing

(vii) Prepare the plots for the relative frequencies as well as cumulative relative frequencies on simple graph papers taking the group interval as abscissa and the relative frequencies or cumulative relative frequencies as ordinate.

6.0 STATISTICS USING GROUPED DATA

Some of the important sample statistics which can be derived using the grouped data are given below:

(i) Mean (m): It is the first moment about origin and given as:

$$m = \frac{\sum_{i=1}^{NC} f_i x_i}{\sum_{i=1}^{NC} f_i} \qquad \ldots\ldots(20)$$

Note that the first moment about the mean is zero.

(ii) Variance (S^2): It is the second moment of the grouped data about the mean. Mathematically, it is expressed as:

$$S^2 = \frac{1}{(N-1)} \sum_{i=1}^{NC} (X_i - m)^2 f_i \qquad \ldots\ldots(21)$$

where,

X_i is the mid point of ith class interval
f_i is no. of values in ith class
N is total No. of values.

The standard deviation is the square root of the variance.

(iii) Coefficient of Variation: It is a non-dimensional parameter and expressed as the ratio of standard deviation and mean, computed for grouped data.

(iv) Skewness: It is the third moment about the mean and expressed as:

$$m_3 = \frac{\sum_{i=1}^{NC} f_i (x_i - m)^3}{\sum_{i=1}^{NC} f_i} \qquad \ldots . (22)$$

(vi) Coefficient of skewness: It is the same as defined earlier in the text except the third moment and standard deviations computed for grouped data are utilised.

(vi) Kurtosis: It is fourth moment about the mean and computed as:

$$m_1 = \frac{\sum_{i=1}^{NC} f_i (x_i - m)^4}{\sum_{i=1}^{NC} f_i} \qquad \ldots . (23)$$

(vii) Coefficient of Kurtosis: It has the same meaning as defined earlier except the fourth moment about the mean and standard deviation used should be computed using the formulae for grouped data.

7.0 PROBABILITY DISTRIBUTIONS

A distribution is an attribute of a statistical population. If each element of a population has a value of X then the distri-bution describes the constitution of the population as seen through its X values. It tells whether they are in general very large or very small, that is their location on the axis. It tells whether they are bunched together or spread out and whether they are symmetrically disposed on the X axis or not. These three are described by the mean, standard deviation and skewness.

Distribution also tells the relative frequency or proportion of various X values in the population in the same way that a histogram gives that information about a sample. These relative frequencies are also probabilities and hence the distribution tells us the probability, $Pr(X \leq x)$, that the X value on an element drawn randomly from the population would be less than a particular value x. Knowing $Pr(X \leq x)$ for all X values, the laws of probability may then be used to deduce the probability of any proposition about the behaviour of a random sample of X values drawn from the population.

When the population is sufficiently large the histogram of its X values can be made with very small class intervals and the histogram can be replaced by a smooth curve, the area enclosed by any two vertical ordinates being the relative frequency or probability of X values between those ordinates.

Because of this probability interpretation, a relative frequency distribution is also called a probability distribution and the curve describing it is called a probability density function (p.d.f) whose cumulative function is called the distribution function (d.f.). Fig. 1 shows a typical shape of p.d.f. and d.f.

7.1 Continuous Probability Distributions

A large number of frequency distributions are available in literature. Here the normal, log normal (two parameters), Extreme Value type-I (Gumbel or EV1), Pearson type-III, Log Pearson type-III, General extreme Value, Gamma and Exponential distributions have been discussed. The probability density functions (P.D.F.), cumulative density functions (C.D.F.) and other properties of these distributions are given below:

Normal Distribution:

$$P.D.F.: f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \ldots\ldots(24)$$

$$C.D.F.: F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \ldots\ldots(25)$$

Parameters: μ = location parameter
σ = scale parameter

$$\text{Reduced Variate: } Z = \frac{(x-u)}{\sigma} \quad \ldots.(26)$$

$$\text{P.D.F.: } f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^{2/2}} \quad \ldots.(27)$$

$$\text{C.D.F. : } F(z) = -\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^{2/2}} dz$$

$$f(x) = \frac{1}{\sigma_y\sqrt{2\pi x}} \exp\left[-\frac{1}{2}\left(\frac{\log_e x - \mu}{\sigma_y}\right)^2\right] \quad \ldots.(28)$$

Mean of the reduced variate: $\bar{z} = 0$
Standard deviation of reduced variate $\sigma_z = 1$
Coefficient of skewness of the reduced variates = $g_z = 0$

Log Normal Distribution (Two Parameters):

$$\text{P.D.F.: } f(x) = \frac{1}{\sigma_y\sqrt{2\pi x}} \exp\left[-\frac{1}{2}\left(\frac{\log_e x - \mu_y^2}{\sigma_y}\right)^2\right] \quad \ldots.(29)$$

$$\text{C.D.F.: } F(x) = \frac{1}{\sigma_y\sqrt{2\pi}} \int_{x_o}^{x} \exp\left[-\frac{1}{2}\left(\frac{\log_e x - \mu_y}{\sigma_y}\right)^2\right] dx \quad \ldots.(30)$$

where,

$y = \log_e x$
μ = Mean of Y series
σ_y = Standard deviation of Y-series

Parameters: u_y = location parameter
σ_y = scale parameter

Reduced variate $Z = \dfrac{\log_e x - \sigma_y}{\sigma_y}$ ....(31)

P.D.F.: $f(z) = \dfrac{1}{\sqrt{2\pi}} e^{-z^{2/2}}$ ....(32)

C.D.F.: $F(z) = \int_0^z - \dfrac{1}{\sqrt{2\pi}} e^{-z^{2/2}} dz$ ....(33)

Mean of the reduced variate: $\bar{z} = 0$
Standard deviation of reduced variate: $\sigma_z = 1$
Coefficient of skewness of the reduced variates = $g_z = 0$

Gumbel Extreme Value (Type-I) Distribution(EV1):

$$P.D.F.\ f(x) = \frac{1}{\alpha} \exp\left[-\left(\frac{x-u}{\alpha}\right) - \exp\left\{-\left(\frac{x-u}{\alpha}\right)\right\}\right] \quad \ldots.(34)$$

C.D.F.: $F(x) = e^{-e^{-\left(\frac{x-u}{\alpha}\right)}}$ ....(35)

Reduced Variate : $z = \dfrac{x-u}{\alpha}$ ....(36)

P.D.F.: f(z) = exp (-z-exp(-z)) ....(37)

C.D.F. : $F(z) = e^{-e^{-z}}$ ....(38)

Mean reduced variate Z = 0.5772
Standard deviation of reduced variate $\alpha_z = \dfrac{\pi}{\sqrt{6}} = 1.2825$
Coefficient of skewness of reduced variate $g_z = 1.14$

Pearson type-III Distribution (PT3):

$$P.D.F.: f(x) = \frac{(x-x_o)^{\gamma-1} e^{-(x-x_o)/\beta}}{\beta^{\gamma}\sqrt{\gamma}} \quad \ldots.(39)$$

$$C.\,D.\,F.\ F(x) = \int_{x_o}^{x} \frac{(x - x_o)^{\gamma-1} e^{-(x-x_o)/\beta}}{\beta^{\gamma}\sqrt{\gamma}} dx \qquad \ldots\ (40)$$

Parameters: x_o = Location parameter
β = scale parameter
γ = shape parameter

$$\text{Reduced variate } Z = \frac{x - x_c}{\beta} \qquad \ldots(41)$$

$$\text{P.D.F.: } f(z) = \frac{1}{|\beta|\sqrt{\gamma}} (z)^{\gamma-1} e^{-z} \qquad \ldots(42)$$

$$\text{C.D.F.: } F(z) = \int_{o}^{z} \frac{1}{|\beta|\sqrt{\gamma}} (z)^{\gamma-1} e^{-z} \qquad \ldots(43)$$

Mean of the reduced variate: $\bar{z} = \gamma$
St. Deviation of the reduced variate: $\sigma_z = \sqrt{\gamma}$
Coefficient of skewness of the reduced variate: $g_z = 2/\sqrt{\gamma}$

Log Pearson Type-III Distribution (LP3):

$$P.D.F.: f(x) = \frac{(\log_e x - y_o)^{\gamma-1} e^{-(\log_e x - y_o)/\beta}}{|\beta|\sqrt{\gamma}\, x} \qquad \ldots\ (44)$$

$$\text{C.D.F.: } F(x) = \int_{yo}^{x} f(x)\, dx \qquad \ldots(45)$$

Parameters: y_o = Location parameter
β = scale parameter
γ = shape parameter

$$\text{Reduced variate: } Z = \frac{\log_e x - y_o}{\beta} \qquad \ldots(46)$$

$$\text{P.D.F.: } f(z) = \frac{1}{|\beta|\gamma} (z)^{\gamma-1} e^{-z} \qquad \ldots(47)$$

$$\text{C.D.F.: } F(z) = \int_{o}^{z} f(z)\, dz \qquad \ldots(48)$$

Mean of reduced variate = $\bar{z} = \gamma$
Standard dev. of reduced variate: $\sigma_z = \sqrt{\gamma}$
Coefficient of skewness of reduced variate: $g_z = 2/\sqrt{\gamma}$

General Extreme Value Distribution:

$$P.D.F.: f(x) = \frac{1}{\alpha}[1-K(\frac{x-u}{\alpha})]^{(1/K)-1} e^{-[1-K(\frac{x-u}{\alpha})]^{1/K}} \quad \ldots\ldots(49)$$

$$\text{C.D.F.: } F(x) = \text{Exp}\left[-(1-K(\frac{x-u}{\alpha}))^{1/K}\right] \quad \ldots.(50)$$

Parameters: u = Location parameter
α = Scale parameter
K = Shape Parameter

If K = 0 it leads to EV-I distribution,
K < 0 it leads to EV-II distribution
K > 0 it leads to EV-III distribution

$$\text{GEV reduced variate : } w = \frac{x-u}{\alpha} \quad \ldots.(51)$$

Here $w = (1_e - kz)/K$
Z = EV-I reduced variate

Gamma Distribution:

It is a special case of PT3 distribution.

$$\text{P.D.F.: } f(x) = \frac{(x)^{\gamma-1} e^{-x/\beta}}{\beta^{\gamma}\sqrt{\gamma}} \quad \ldots\ldots\ldots \quad \ldots.(52)$$

$$\text{C.D.F.: } F(x) = \int_o^x \frac{(x)^{\gamma-1} e^{-x/\beta}}{\beta^{\gamma}\sqrt{\gamma}} dx \quad \ldots.(53)$$

Parameters: β = Scale Parameter
γ = Shape parameter

$$\text{Gamma Reduced Variate } z = \frac{x}{\beta} \quad \ldots.(54)$$

$$\text{P.D.F.: } f(z) = \frac{1}{|\beta|\sqrt{\gamma}}(z)^{\gamma-1} e^{-z} \quad \ldots.(55)$$

$$\text{C.D.F.: } F(z) = \int_o^z \frac{1}{|\beta|\sqrt{\gamma}}(z)^{\gamma-1} e^{-z} \quad \ldots.(56)$$

Mean of the reduced variate: $\bar{z} = \gamma$
St. Deviation of the reduced variate: $\sigma_z = \sqrt{\gamma}$
Coefficient of skewness of the reduced variate : $g_z = 2/\sqrt{\gamma}$

Exponential Distribution:

$$\text{P.D.F.: } f(x) = \frac{1}{\beta} e^{\frac{-(x-x_o)}{\beta}} \quad \ldots.(57)$$

$$\text{C.D.F.: } F(x) = 1 - e^{\frac{-(x-x_o)}{\beta}} \quad \ldots.(58)$$

Parameters: x_o = Location Parameter
β = Scale Parameter

Reduced Variate : $z = \frac{x - x_o}{\beta}$ ....(59)

P.D.F.: $f(z) = e^{-z}$ ....(60)

C.D.F.: $F(Z) = 1-e^{-z}$ ....(61)

Mean of the reduced variate: $\bar{z} = 1$
St. Deviation of the reduced variate $\sigma_z = 1$
Coefficient of skewness of the reduced variate : $g_z = 2$

7.2 Discrete Probability Distributions

The use of discrete probability distributions is restricted generally to those random events in which the outcome can be described as success or failure, i.e there are only two mutually exclusive events of an experiments. Furthermore, the successive trials are independent and the probability of success remains constant from trial to trial.

The binomial or Poisson distributions can be used to find the probability of occurrence of an event r times m n successive years.

Binomial Distribution:

This apply to populations that have only two discrete but complementary events, for example rainy and non-rainy days. The probability of occurrence of the event r times in n successive years is given by:

$$P_{r,n} = n\,C_r\,P^r\,q^{n-r} = \frac{n!}{r!\,(n-r)!}\,p^r\,q^{n-r} \quad \text{....(62)}$$

where, $P_{r,n}$ = probability of a random hydrologic event of a giveı. magnitude and exceedence probability P occurring r times in n successive years. Thus, for example:

(a) The probability of an event of exceedence probability P occurring Z times in n successive years is:

$$P_{z,n} = \frac{n!}{(n-z)!\,z!}\,p^z\,q^{n-z} \quad \text{....(63)}$$

(b) The probability of the event not occurring at all in n successive years is:

$$P_{o,n} = q^n = (1-p)^n \quad \text{....(64)}$$

(c) The probability of the event occurring at lease once in n successive years:

$$P_1 = 1 - q^n = 1 - (1-p)^n \quad \text{....(65)}$$

Example:

Analysis of data on maximum one day rainfall depth at a station indicated that a depth of 280 mm had a return period of 50 years. Determine the probability of a one day rainfall depth equal to or greater than 280 mm occurring (a) once in 20 successive years, (b) two times in 15 successive years, and (c) at least once in 20 successive years (Subramanya, 1984).

Solution:

Here, P = 1/50 = 0.02

(a) n = 20, r = 1

$$P_{1,20} = \frac{20!}{19!\,1!} \times 0.02 \times (0.98)^{19}$$

$$= 20 \times 0.02 \times 0.68123$$
$$= 0.272$$

(b) n = 15, r = 2

$$P_{2,15} = \frac{15!}{13!\,2!}(0.02)^2 \times (0.98)^{13}$$

$$= \frac{15 \times 14}{2} \times 0.02 \times 0.02 \times 0.98^{13} = 0.0292$$

(c) Using Eq. (65), $P_1 = 1-(1-0.02)^{20} = 0.332$

Poisson Distribution:

The terms of a binomial expansion are a little inconvenient to compute in any large number. If n is large (>30), p is small (<0.1) and mean np is constant then binomial distribution tends to Poisson distribution:

$$P_{r,n} = \frac{\lambda^r e^{-\lambda}}{r!}$$

where, $\lambda = np$

The condition for this approximation are:

1. The number of events is discrete.
2. Two events cannot coincide.

3. The mean number of events in unit time is constant.
4. Events are independent.

Thus, it can be applied to following situations of rate events with p relatively small and n relatively large:

(i) Determining the probability of droughts in a given time period.
(ii) Determining the probability of number of rainy days at a given location.
(iii) Determining the probability of rare flood event of the 1 in 100 year type.
(iv) Determining the probability of reservoir being empty in any one year out of a long period of record.

7.3 Common Probability Functions

Probability functions which are frequently used in hydrologic analysis include the normal the student t, the chi-square and the F distributions, whereas the last three are used primarily for support in making statistical tests of hypothesis, the normal distribution is used for prediction as well as support. Since these distributions serve primarily a support function, the discussion here will centre on obtaining critical values from tables provided in the Appendices.

Normal Distribution

The probability density function and cummulative density function for Normal distribution are described in section 9.7.1 . Here μ and σ are the population parameters of the distribution. It can be shown that the best estimates of μ and σ are the sample mean ($\bar{X}$) and standard deviation (S), respectively. For computing probabilities based on sample statistics, $\bar{X}$ and S can be substituted for μ and σ respectively. Probabilities could be computed integrating the probability density function over a range of values of x. However, since there are an infinite number of values of μ and σ (or $\bar{X}$ and S), numerous such integrations would become very tedious. The problem can be circumvented by making a transformation of the random variable x.

If the random variable x has a normal distribution, a new random variable z can be obtained after transforming the x values as follows:

$$z = \frac{x - \mu}{\sigma} \qquad \ldots.(66)$$

Here random variable z will have a mean and standard deviation of 0 and 1 respectively. z is called the standardized variate or normal reduced variate. Using the above transformation the p.d.f. may be expressed as a function of z and the resulting p.d.f. is called the standard normal distribution. Since there is only one standard normal distribution, probabilities can be computed using p.d.f. of the standard normal distribution and placed in tabular form as a function of z (see Appendix-I). The table is structured with values of z at increments of 0.1 down the left margin and at increments of 0.01 across the top. The cumulative probability from $-\infty$ the desired value of z is given within the table. For example the probability that z is less than 0.23 is 0.5910. Also the probability that z is between -0.47 and 0.23 equals (0.5910-0.3192) or 0.2718. The figure shown at the top of the table provides a good understanding of the relation-ship between the probability and the value of z. The table can also be used to find the value of z that corresponds to a certain probability. For example for a probability of 0.05 in the left tail, we enter the table with 0.05 and find the corresponding value of z, which is -1.645. Since the distribution is symmetric, 5% of the area under the curve f(z) lies to the right of a value of z of 1.645.

Student t-distribution

The student t or t_1, distribution is similar to the normal distribution in that it is symmetric. However, it differs from the normal distribution in the sense that it is a function of a single parameter υ, which is often called the degrees of freedom and controls the spread of the t-distribution since it is a function of the parameter υ, which is a positive integer, there are many t distributions. Thus the table of t values has a slightly different structure than the normal table. In the table of Appendix-II, the value of υ is given in the left margin, the probability in the right tail of the distribution is given across the top, and the value of the t distribution is given within the table. For example, for $\upsilon = 7$ and a probability of 0.05, the t value is 1.895. Since the t distribution is symmetric, 5% of the area in the left tail is to the left of a t value of -1.895 for $\upsilon = 7$. For the case where one is interested in 5% of the area but with 2.5% in each tail, the critical t values would be -2.365 and 2.365 for $\upsilon = 7$. One final point, it is usually acceptable to use the normal distribution in place of t-distribution for $\upsilon > 30$.

Chi-Square Distribution

The Chi-Square (χ^2) distribution is similar to the t-distribution in that it is a function of a single parameter n, the degrees of freedom; however it differs from the t distribution in that the distribution is not symmetric. The table of chi-square values (Appendix-III) is identical in structure to the t table, with n down the left margin, the probability across the top, and the value of the random variable (χ^2) in the table. For 11 degrees of freedom, 5% of the area in the left tail is from 0 to 4.575. For 17 degrees of freedom, there is a probability of 0.025 that χ^2 will be between 30.191 and ∞.

F-distribution

The F-distribution is a function of two parameters, m and n. To obtain F values from Appendix-IV, the value of m is entered along the top of the table and the value of n along the left margin. The value of F corresponding to the appropriate probability in the right tail of the distribution is obtained from the table. For example if m = 10 and n = 20, there is a 5% chance that F will be greater than 2.35. Note that the table is not symmetric.For m = 20 and n = 10, the critical value for 5% in the right tail is 2.77. Thus the values of m and n must not be switched.

8.0 SIMPLE LINEAR REGRESSION

Form of the equation : $Y = a + bX$

Regression Coefficients:

$$b = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{y}) / \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \ldots\ldots (67)$$

$$a = \bar{y} - b\bar{x} \qquad \ldots\ldots (68)$$

Coefficient of determination (r^2):

$$r^2 = [\sum (x_i - \bar{x})(y_i - \bar{y})]^2 / (\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2) \qquad \ldots\ldots (69)$$

Coefficient of correlation (r):

$$r = \sqrt{r^2} = \sum (x_i - \bar{x})(y_i - \bar{y}) / (\sum (x_i - \bar{x})^2 \sum (x_i - \bar{y})^2)^{1/2} \quad \ldots\ldots (70)$$

Efficiency (EF):

$$EF = 1 - \frac{S}{Sy} \quad \ldots\ldots (71)$$

where,

$$S^2 = \sum (y_i - \hat{y}_i)^2 / (n-2) \quad \ldots\ldots (72)$$

$$S_y^2 = \sum (y_i - \bar{y})^2 / (n-1) \quad \ldots\ldots (73)$$

Inferences on Regression Coefficients:

(i) Standard error of $a(S_a) = S\left(\frac{1}{n} + \frac{x^{-2}}{\sum (x_i - \bar{x})^2}\right)^{1/2}$ ....(74)

(ii) Standard error of $b(S_b) = S / (\sum (x_i - \bar{x}^2)^{1/2}$ ....(75)

(iii) Confidence intervals on a:

$$1_a = a - t_{(1-\alpha/2),(n-2)} S_a \quad \ldots\ldots (76)$$

$$u_a = a + t_{(1-\alpha/2),(n-2)} S_a \quad \ldots\ldots (77)$$

(iv) Confidence intervals on b:

$$1_b = b - t_{(1-\alpha/2),(n-2)} S_b \quad \ldots\ldots (78)$$

$$u_b = b + t_{(1-\alpha/2),(n-2)} S_b \quad \ldots\ldots (79)$$

where, l_a & l_b denote lower confidence limits on a & b respectively. u_a and u_b denote upper confidence limits on a & b respectively. α is the confidence level, and $t_{(1-\alpha/2),\,(n-2)}$ represent t values corresponding to $(1-\alpha/2)$ confidence limits and (n-2) degrees of freedom (t values are given in Appendix-III in tabular form).

(v) Test of hypothesis concerning a:

Hypothesis H_o : $a = a_o$ versus H_a : $a \neq a_o$ is tested by computing:

$$t = (a - a_o)/S_a \quad \ldots\ldots (80)$$

H_o is rejected if $|t| > t_{(1-\alpha/2),\,(n-2)}$

(vi) Test of hypothesis concerning b:

Hypothesis $H_o : b = b_o$ versus $H_b : b \neq b_o$ is tested by computing:

$$t = (b - b_o)/S_b \quad \ldots.(81)$$

H_o is rejected if $| t | > t_{(1-\alpha/2),\,(n-2)}$

Significance of the overall regression:

Hypothesis $H_o : b = 0$ is tested by computing:

$$t = (b - 0)/S_b \quad \ldots.(82)$$

H_o is rejected if $| t | > t_{(1-\alpha/2),\,(n-2)\,a}$ and the regression equation explaining a significant amount of variation in Y.

Confidence intervals on Regression line:

$$L = \hat{y}_k - S_{\hat{y}_k} \cdot t_{(1-\alpha/2),\,(n-2)} \quad \ldots.(83)$$

$$U = \hat{y}_{y_k} + S_{\hat{y}_k} t_{(1-\alpha/2),\,(n-2)} \quad \ldots.(84)$$

where,

$$\hat{Y}_k = a + bx_k \quad \ldots.(85)$$

$$S_{\hat{y}_k} = S\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]^{1/2} \quad \ldots.(86)$$

= standard error of $S_{\hat{y}_k}$

L and U represent lower and upper confidence limits.

Confidence intervals on an individual predicted value of y:

$$L' = \hat{y}_k - S_{\hat{y}_k} t_{(1-\alpha/2),\,(n-2)} \quad \ldots.(87)$$

$$U' = \hat{y}_k + S_{\hat{y}} \; t_{(1-\alpha/2),\,(n-2)} \quad \ldots.(88)$$

$$S'_{\hat{y}_k} = S\left[1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]^{1/2} \quad \ldots.(89)$$

Example: The precipitation and runoff for a typical catchment for the month of July are given below in tabular form:

Year	Precipitation	Runoff
1953	42.39	13.26
1954	33.48	3.31
1955	47.67	15.17
1956	50.24	15.50
1957	43.28	14.22
1958	52.60	21.20
1959	31.06	7.70
1960	50.02	17.64
1961	47.08	22.91
1962	47.08	18.89
1963	40.89	12.82
1964	37.31	11.58
1965	37.15	15.17
1966	40.38	10.40
1967	45.39	18.02
1968	41.03	16.25

(a) Develop the rainfall runoff relationship in the form: Y = a+b X; where Y represents the runoff variable and X represents precipitation variable.

(b) What percent of the variation in runoff is accounted for by the developed regression equation.

(c) Compute the 95% confidence interval on a and b and test the hypothesis that a = 0 and the hypothesis that b = 0.500 for the above regression.

(d) Calculate the 95% confidence limits for the regression line. Calculate the 95% confidence interval for an individual predicted value of Y.

Solution:

(a) The form of the regression:

$$Y = a + b\,X$$

The regression coefficients are:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{369.432}{570.0559} = 0.648$$

$$a = \bar{y} - b\bar{x} = 14.63 - 0.648 \times 42.94 = -13.1951$$

∴ The regression equation is: Y = -13.1951 + 0.648 X

(b) The percent of variation in Y is accounted for by the regression is computed as the coefficient of determination (r^2) multiplied by 100.

$$r^2 = \frac{[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= b \times \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 0.648 \times \frac{369.432}{363.0714} = 0.66$$

Thus 66 percent of variation in Y is explained by the regression equation. The remaining 34 percent of variation is due to unexplained causes.

The coefficient of correlation (r) may be computed as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2]^{1/2}}$$

= square root of coefficient of determination

$= \sqrt{0.66} = 0.81$

Note: $0 \leq r^2 \leq 1$ and $-1 \leq r \leq 1$

(c) (i) Computation of 95% confidence intervals on a and b. The steps are:

- Compute the standard error of the regression equation using the relation:

$$S^2 = \frac{\sum_{i=1}^{n} y_i - \hat{y})^2}{(n-2)} = \frac{123.7}{16-2} = \frac{123.7}{14} = 8.83$$

$$S = 2.97$$

- Compute standard error of a (S_a):

$$S_a = S\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{1/2}$$

$$= 2.97\left(\frac{1}{16} + \frac{42.94 * 42.94}{570.0559}\right)^{1/2} = 5.39$$

- Compute standard error of b (S_b):

$$S_b = \frac{S}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}} = \frac{2.97}{(570.0559)^{1/2}} = 0.125$$

- Compute $t_{(1-\alpha/2),\,(n-2)}$ from the t-table where $\alpha = 0.05$, n-2 = 14.
  Therefore, $t_{(1-0.025),\,(16-2)} = t_{0.975,\,14} \approx 2.14$.

- Compute 95% confidence intervals on a:

$l_a = a - t_{(1-\alpha/2),\,(n-2)} \cdot S_a$

$= -13.1951 - 2.14 \times 5.39$
$= -24.73$

$u_a = a + t_{(1-\alpha/2),\,(n-2)} \cdot S_a$

$= -13.1951 + 2.14 \times 5.39$
$= -1.66$

- Compute 95% confidence intervals on b:

$l_b = b - t_{(1-\alpha/2),\,(n-2)} \cdot S_b$

$= 0.648 - 2.14 \times 0.125 = 0.38$

$$u_b = b + t_{(1-\alpha/2),\,(n-2)} \cdot S_b$$

$$= 0.648 + 2.14 \times 0.125 = 0.92$$

(ii) Testing the hypothesis $H_o : a = 0$ versus $a \neq 0$

- Compute $t = \frac{a - 0.00}{S_a} = \frac{-13.1951}{5.39} = -2.44$

- Since $t_{(1-\alpha/2),\,(n-2)} = t_{0.975,14} = 2.14$

and $|t| > t_{0.995,14}$, we reject $H_o : a = 0$

Testing the hypothesis $H_o : b = 0.5$ versus $H_a : b \neq 0.5$

- Compute $t = \frac{b - 0.5}{S_b} = \frac{0.648 - 0.50}{0.125} = 1.184$

- Since $|t| < 1.184$, we cannot reject H_o.

From the above tests, it is observed that the intercept is significantly different from zero. However, the slope is not significantly different from 0.5.

Comment: The significance of the overall regression can be evaluated by testing H_o: $b = 0$. Under this hypothesis,

$$t = \frac{b - 0.00}{S_b} = \frac{0.648}{0.125} = 5.184$$

Since $|t| > t_{0.975,14}$, we reject Ho. The regression equation is explaining a significant amount of the variation in Y.

(d) (i) Computation of 95% limits for the regression line:

- Compute the standard error of $\hat{Y}_k$ as:

$$S_{\hat{Y}_k} = S\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^{1/2}$$

$$= 2.97\left[\frac{1}{16} + \frac{(x_k - 42.94)^2}{570.0559}\right]^{1/2}$$

$$S_{\hat{Y}_k} = 2.97\left[0.0625 + \frac{(x_k - 42.94)^2}{570.0559}\right]^{1/2}$$

$$L = \hat{y}_k - S_{\hat{y}_k} . t_{(1-\frac{\alpha}{2}),(n-2)}$$

$$U = \hat{y}_k + S_{\hat{y}_k} \; t_{(1-\frac{\alpha}{2}),(n-2)}$$

No	X	Y	$\hat{Y}$	L	U
1	42.39	13.26	14.27	12.67	15.87
2	33.48	3.31	8.50	5.52	11.48
3	47.67	15.17	17.69	15.66	19.72
4	50.24	15.50	19.36	16.85	21.87
5	43.28	14.22	14.85	13.25	16.44
6	52.60	21.20	20.89	17.86	23.91
7	31.06	7.70	6.93	3.39	10.47
8	50.02	17.64	19.22	16.75	21.68
9	47.08	22.91	17.31	15.38	19.25
10	47.08	18.89	17.31	15.38	19.25
11	40.89	12.82	13.30	11.62	14.98
12	37.31	11.58	10.98	8.79	13.16
13	37.15	15.17	10.87	8.66	13.09
14	40.38	10.40	12.97	11.24	14.70
15	45.39	18.02	16.21	14.50	17.93
16	41.03	16.20	13.39	11.72	15.06

(ii) Computation of 95% confidence limits for individual predicted values of Y.

- Compute the standard error of individual predicted values as:

$$S_{\hat{y}_k} = S\left[1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^{1/2}$$

$$= 2.97\left[1 + \frac{1}{16} + \frac{(x_k - 42.94)^2}{570.0559}\right]^{1/2}$$

$$= 2.97\left[1.0625 + \frac{(x_k - 42.94)^2}{570.0559}\right]^{1/2}$$

- Compute 95% confidence limits for individual predicted values of Y.

$$L' = \hat{y}_k - S_{\hat{y}_k} \cdot t_{(1-\frac{\alpha}{2}),(n-2)}$$

$$U' = \hat{y}_k + S_{\hat{y}_k} \cdot t_{(1-\frac{\alpha}{2}),(n-2)}$$

$$L' = -13.1951 + 0.648\, x_k - 2.97\left[1.0625 + \frac{(x_k - 42.94)^2}{570.0559}\right]^{1/2} \cdot 2.14$$

No	X	Y	$\hat{Y}$	L'	U'
1	42.39	13.26	14.27	7.71	20.83
2	33.48	3.31	8.50	1.47	15.52
3	47.67	15.17	17.69	11.02	24.37
4	50.24	15.50	19.36	12.52	26.20
5	43.28	14.22	14.85	8.29	21.40
6	52.60	21.20	20.89	13.84	27.93
7	31.06	7.70	6.93	-0.35	14.21
8	50.02	17.64	19.22	12.39	26.04
9	47.08	22.91	17.31	10.66	23.96
10	47.08	18.89	17.31	10.66	23.96
11	40.89	12.82	13.30	6.72	19.88
12	37.31	11.58	10.98	4.25	17.70
13	37.15	15.17	10.87	4.14	17.61
14	40.38	10.40	12.97	6.38	19.56
15	45.39	18.02	16.21	9.63	22.80
16	41.03	16.20	13.39	6.81	19.97

These confidence intervals are plotted in Fig. 9.1.

Extrapolation:

The extrapolation of a regression equation beyond the range of x used in estimating a and b is discouraged for two reasons. First, as can be seen from Fig. 9.1, the confidence intervals on the regression line become very wide as the distance from $\bar{x}$ is increased. Second, the relation between Y and X may be non-linear over the entire range of X and only approximately linear for the range of X investigated. A typical example of this is shown in Fig. 9.2.

9.0 MULTIPLE LINEAR REGRESSION

Form of the equation:

$$y_1 = \beta_1 x_{1,1} + \beta_2 x_{1,2} + ..\beta_p x_{1,p}$$

$$y_2 = \beta_1 x_{2,1} + \beta_2 X_{2,2} + ...\beta_p x_{2,p}$$

... .

$$y_n = \beta_1 x_{n,1} + \beta_2 x_{n,2} + ...\beta_p x_{n,p}$$

where, y_i is the ith observation on the dependent variable.

$$y_i = \sum_{j=1}^{p} \beta_j \, x_{i,j} \qquad \ldots.(90)$$

p = number of independent variables.

In matrix notation

$$\underset{(nx1)}{Y} = \underset{(nxp)}{X} \; \underset{(px1)}{\beta} \qquad \ldots.(91)$$

Correlation matrix R, of the independent variables:

$$\text{Let } Z_{i,j} = (X_{i,j} - \bar{x_j})/S_j \qquad \ldots.(92)$$

where, $\bar{x_j}$ and S_j are the mean and standard deviation of the j^{th} independent variable:

$$Z = [Z_{i,j}] \qquad \ldots.(93)$$

So the correlation matrix is:

$$R = Z' \, Z/(n-1) = [R_{i,j}] \qquad \ldots.(94)$$

where, $R_{i,j}$ is the correlation between the i^{th} and j^{th} independent variables. R is a symmetric matrix since $R_{i,j} = R_{j,i}$

Regression Co-efficients:

$$\underset{(px1)}{\beta} = \underset{(pxn)(nxp)}{(X'X)^{-1}} \; \underset{(pxn)}{X'} \; \underset{(nx1)}{Y} \qquad \ldots.(95)$$

X' is transpose of matrix X of size (pxn).

Coefficient of Determination (R^2):

$$R^2 = (\beta' X' Y - n\bar{Y}^2) / (Y' Y - n\bar{Y}^2) \qquad \ldots.(96)$$

Here, β' is transpose of vector β of size (1xp), and
Y' is transpose of vector Y of size (1 x n)

Coefficient of Correlation (R):

R = square root of coefficient of Determination.

Efficiency (EF):

$$EF = 1 - \frac{S}{S_y} \qquad \ldots.(97)$$

where,

$$S^2 = \sum (y_i - \hat{y}_i)^2 / (n-p) \quad \text{....(98)}$$

$$S_y^2 = \sum (y_i - \bar{y})^2 / (n-1) \quad \text{....(99)}$$

Inferences on Regression Coefficients:

(1) Standard errors of β_i

Let C = X' X, then $C^{-1} = (X' X)^{-1}$, and

$$Var(\beta_i) = S_{\beta_i}^2 = C_{ii}^{-1} S^2 \quad \text{....(100)}$$

where, C_{ii}^{-1} is the ith diagonal element of $(X'X)^{-1}$

$$S_{\beta_i} = C_{ii}^{-1} S^2 \quad \text{....(101)}$$

(ii) Confidence intervals on β_i

$$L_{\beta_i} = \beta_i - t_{(1-\alpha/2),(n-p)} S_{\beta_i}$$

$$U_{\beta_i} = \beta_i - t_{(1-\alpha/2),(n-p)} S_{\beta_i} \quad \text{....(102)}$$

(iii) Test of hypothesis concerning β_i:

Hypothesis $H_o : \beta_i = \beta_o$ versus $H_a : \beta_i \neq \beta_o$ is tested by computing:

$$t = \frac{(\beta_i - \beta_o)}{S_{\beta_i}} \quad \text{....(103)}$$

H_o is rejected if $| t | > t_{(1-\alpha/2),\,(n-p)}$

(iv) Test of hypothesis that the ith independent variable is not contributing significantly in explaining the variation in the dependent variable hypothesis $H_o : \beta_i = 0$ versus $H_a : \beta_i \neq 0$ is tested by computing:

$$t = \beta_i / S_{\beta_i} \quad \text{....(104)}$$

H_o is accepted if $| t | < t_{(1-\alpha/2),\,(n-p)}$.
In such a situation it is advisable to delete the ith independent variable from the model.

Significance of the overall regression:

Hypothesis $H_o : \beta_2 = \beta_3 = .. \beta_p = 0$ versus H_a : at least one of these β's is not zero is tested by computing the test statistic:

$$F = \frac{((\beta' X' Y) - n\,Y^2)/(p-1)}{(Y'Y - \beta' X' Y)/(n-p)} \quad \text{....(105)}$$

H_o is rejected if $| F |$ exceeds $F_{(1-\alpha),\,(p-1),\,(n-p)}$ which values are given in Appendix-IV in tabular form.

Confidence Intervals on Regression Line:

$$L = \hat{Y}_k - t_{(1-\alpha/2),(n-p)} S_{\hat{Y}_k} \quad \ldots.(106)$$

$$U = Y_k + t_{(1-\alpha/2),(n-p)} S_{\hat{Y}_k} \quad \ldots.(107)$$

where,

$$\hat{Y}_K = X_K \hat{\beta} \quad \ldots.(108)$$

$$S_{\hat{Y}_k}^2 = S^2 X_K (X'X)^{-1} X_K \quad \ldots.(109)$$

Confidence Intervals on individual Predicted Value of Y:

$$L = \hat{Y}_k - t_{(1-\alpha/2),(n-p)} S_{\hat{Y}_k} \quad \ldots(110)$$

$$U = \hat{Y}_k + t_{(1-\alpha/2),(n-p)} S_{\hat{Y}_k} \quad \ldots.(111)$$

$$S^2_{\hat{Y}_k} = S^2 (1 + X_k (XX)^{-1} X_k) \quad \ldots.(112)$$

10.0 ABSOLUTE AND RELATIVE SENSITIVITIES

$$y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p x_{i,p} \quad \ldots.(113)$$

Absolute Sensitivities are given by:

$$AS_j = \left(\frac{\delta y}{\delta x_j}\right), j = 1, 2 \ldots.p \quad \ldots.(114)$$

Linear Sensitivity equation:

$$\Delta y'_j = \left(\frac{\delta y}{\delta x_j}\right) \Delta x_j, j = 1, 2\ldots.p \quad \ldots.(115)$$

Relative Sensitivities:

$$RS_j = \left(\frac{\delta y}{\delta x_j}\right) \frac{x_j}{y}, j = 1, 2\ldots., p\ldots \quad \ldots.(116)$$

Example:

The mean annual peak flood in (Q) in thousand of cumec, catchment area (A) in thousands of square kilometere and average annual maximum 24-hour rainfall (I) depth in cm for some of the gauged catchments of a typical region are given below:

No.	Q	A	I
1	15.50	1.250	1.7
2	8.50	0.871	2.1
3	85.00	5.690	1.9
4	105.00	8.270	1.9
5	24.80	1.620	2.1
6	3.80	0.175	2.4
7	1.76	0.148	3.2
8	18.00	1.400	2.7
9	8.75	0.297	2.9
10	8.25	0.322	2.9
11	3.56	0.178	2.8
12	1.90	0.148	2.7
13	16.50	0.872	2.1
14	2.80	0.091	2.9

(a) Estimate the regression coefficients for the model: $Q = \beta_1 + \beta_2 A + \beta_3 I$ and also estimate R^2.

(b) Test the hypothesis that the regression equation is not explaining a significant amount of the variation of Y.

(c) Test the hypothesis Ho : $\beta_2 = 0$

(d) Test the hypothesis Ho : $\beta_3 = 0$

(e) Calculate the 95% confidence limits on β_2.

(f) Calculate the 95% confidence limits on regression line at the A = 4000 square kilometre and I = 2.0 cm.

Solution:

(a) Since $Q = \beta_1 + \beta_2 A + \beta_3 I$

or

$$y_1 = \beta_1 . X_{1,1} + \beta_2 x_{1,2} + \beta_3 X_{1,3}$$
$$y_2 = \beta_1 . X_{2,1} + \beta_2 x_{2,2} + \beta_3 X_{2,3}$$
$$y_3 = \beta_1 . X_{3,1} + \beta_2 x_{3,2} + \beta_3 X_{3,3}$$
$$\vdots \quad \vdots \quad \vdots \quad \vdots$$
$$y_{14} = \beta_1 . X_{14,1} + \beta_2 x_{14,2} + \beta_3 X_{14,3}$$

In matrix notation:

$$\underset{(14x1)}{\underline{Y}} = \underset{(14x3)}{\underline{X}} \; . \underset{(3x1)}{\underline{\beta}}$$

here, $\underline{Q} = \underline{Y}$

$X_{1,1} = X_{2,1} = x_{3,1} = \ldots . \; X_{14,1} = 1.0$
$X_{1,2} = A_1 \qquad X_{1,3} = I_1$
$X_{2,2} = A_2 \qquad X_{2,3} = I_2$
$\ldots \qquad \ldots$
$\ldots \qquad \ldots$
$X_{14,2} = A_{14} \qquad X_{14,3} = I_{14}$

$$\therefore \underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ .. \\ .. \\ y_{14} \end{bmatrix} ; \underline{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & X_{1,3} \\ X_{2,1} & X_{2,2} & X_{2,3} \\ X_{3,1} & X_{3,2} & X_{3,3} \\ .. & .. & .. \\ .. & .. & .. \\ X_{14,1} & X_{14,2} & X_{14,3} \end{bmatrix}$$

$$\underset{(3x14)}{X'} = \begin{bmatrix} x_{1,1} & X_{2,1} & .. & .. & X_{14,1} \\ x_{1,2} & X_{2,2} & .. & .. & X_{14,2} \\ X_{1,3} & X_{2,3} & .. & .. & X_{14,3} \end{bmatrix}$$

$$X'X = \begin{bmatrix} \Sigma x_{i,1}^2 & \Sigma X_{i,2} X_{i,1} & \Sigma x_{i,3} X_{i,1} \\ \Sigma x_{i,2} X_{i,1} & \Sigma x_{i,2}^2 & \Sigma x_{i,3} x_{i,2} \\ \Sigma x_{i,3} x_{i,1} & \Sigma x_{i,3} x_{i,2} & \Sigma x_{i,3}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 14.0 & 21.33 & 34.30 \\ 21.33 & 108.741 & 43.34 \\ 34.30 & 43.34 & 86.99 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 3.71678 & -0.18094 & -1.37537 \\ -0.18094 & 0.02028 & 0.06124 \\ -1.37537 & 0.06124 & 0.52332 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \Sigma X_{i,1} \cdot y_i \\ \Sigma x_{i,2} \cdot y_i \\ \Sigma x_{i,3} \cdot y_i \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_{i,2} \cdot y_i \\ \Sigma x_{i,3} y_i \end{bmatrix} = \begin{bmatrix} 304.14 \\ 1465.8927 \\ 627.800 \end{bmatrix}$$

$$Since \quad \beta = (X'X)^{-1} X'Y = \begin{bmatrix} 1.6570 \\ 13.1510 \\ 0.0112 \end{bmatrix}$$

Therefore, the parameter estimates are:

$$\beta_1 = 1.657, \beta_2 = 13.1510, \ and \ \beta_3 = 0.0112$$

Thus the regression equation is:

$$Q = 1.657 + 13.151 A + 0.0112 I$$

$$Now \quad R^2 = \frac{(\beta' X'Y - n\bar{y}^2)}{(Y'Y - n\bar{y}^2)}$$

$$= \frac{[19,788.911 - 14(21.7229)^2]}{[19,960.066 - 14(21.7229)^2]} = 0.9872$$

This means that 99% of the variation in Y is explained by the regression equation.

The ANOV table for this example would be:

Source	d.f.	s.s.	m.s.
Mean	1	6,606.381	-
Regression	2	13,182.600	6591.30
Residual	11	171.09	15.554
Total	14	19,960.066	

From the ANOV table $R^2 = \frac{13,182.60}{(19,960.066 - 6,606.381)} = 0.99$

and variance of the regression equation = $S^2 = 15.554$

Therefore, the standard error of the regression equation is: S = 3.94

Comments:

If a large number of significant figures are not carried in computing $(X'X)^{-1}$ matrix, significant errors can result. To demonstrate this, the elements of the X'X and X'Y matrices were rounded to two decimal places resulting in estimates for $\underline{\beta}$ of $\beta_1 = 1.10$; $\beta_2 = 12.24$, *and* $\beta_3 = 5.28$.

Thus number of significant figures that are carried in the calculations should be as large as practical. In reporting the results, the number of significant figures should be reduced. The reported results on the regression of the above example might be:

$$Q = 1.66 + 13.15\ A + 0.01\ I$$

(b) Testing the hypothesis that the regression equation is not explaining a significant amount of the variation of Y. This H_o is equivalent to $H_o : \beta_2 = \beta_3 = 0$ versus H_a - at least one of β_2 *or* $\beta_3 \neq 0$.

The test is conducted by calculating the F-stastic as:

$$= \frac{\textit{Mean square due to regression}}{\textit{Residual mean square}} = \frac{6591.30}{15.554} \quad (\textit{from ANOV Table}) = 4$$

From the table given in Appendix-IV, critical value of:

$$F_{(1-\alpha),\ (p-1),\ (n-p)} = F_{0.95,\ 2,\ 11} = 3.98$$

Since the computed F-statistic exceeds the critical F, Ho is therefore rejected. It indicates that the regression equation is explaining a singificant amount of variation in Y. Rejection of Ho does not imply that all the independent variables considered are important - it only implies that at least one of these variables is explaining a significant amount of the variation in Y.

(c) Testing the hypothesis $H_o : \beta_2 = 0$ *vs.* $H_a : \beta_2 \neq 0$.

For β_2, t- statistic may be computed as:

$$t = \frac{\beta_2 - \beta_2}{S_{\beta_2}}$$

here $\beta_2 = 13.151$, $\beta_2 = 0$, *and* S_{β_2} is computed using the relation:

$$S^2_{\beta_2} = C_{22}^{-1} . S^2$$

here C = X'X and $C^{-1} = (X'X)^{-1}$.

Here, $S_{\beta_2} = 0.0208 \times 15.554$

$$S_{\beta_2} = (0.0208 \times 15.554)^{1/2} = 0.562.$$

$$\therefore t = \frac{13.151 - 0}{0.562} = 23.4$$

Since critical value of $t_{(1-\alpha/2),(n-p)} = t_{0.975,11}$ from the table is 2.201, therefore $|t| > t_{0.975,11}$. Hence H_0 is rejected indicating that catchment area is explaining a significant amount of the variation in Y.

(d) Testing the hypothesis $H_o : \beta_3 = 0 \quad vs. \quad H_a : \beta_3 \neq 0$

For β_3, t-statistic may be computed as:

$$t = \frac{\hat{\beta}_3 - \beta_3}{S_{\beta_3}}$$

here $\beta_3 = 0.0112$ *and* S_{β_3} computed using the relation:

$$S_{\beta_3}^2 = C_{33}^{-1} S^2 = 0.5233 \times 15.554$$

$$S_{\beta_3} = (0.5233 \times 15.554)^{1/2} = 2.85$$

$$t = \frac{0.0112 - 0}{2.85} = 0.004$$

Since $|t| < t_{0.975,11}$ (=2.201), we can not reject Ho. The mean annual maximum 24-hour rainfall depth is not explaining a significant amount of the variation in the mean annual peak flow.

(e) 95% confidence limits on β_2 are calculated using the equation:

$$l = \beta_2 - S_{\beta_2} \cdot t_{(1-\alpha/2),(n-p)}$$

$$U = \beta_2 + S_{\beta_2} \cdot t_{(1-\alpha/2),(n-p)}$$

$$l = 13.15 - 0.562 \times 2.201 = 11.91$$

$$u = 13.15 + 0.562 \times 2.201 = 14.39$$

(f) 95% confidence limits on regression line at A = 4000 square kilometer and I = 20 cm is computed using:

$$lr = \hat{y}_k - t_{(1-\alpha/2),(n-p)} \cdot S_{\hat{y}_k}$$

$$Ur = \hat{y}_k + t_{(1-\alpha/2),(n-p)} \cdot S_{\hat{y}_k}$$

here $\hat{y}_k = 1.6570 + 13.151\ A + 0.0112\ I$
$= 1.6570 + 13.151 \times 4 + 0.0112 \times 2$
$= 54.28$

$$S_{Y_k}^2 = S^2 X_K (X'X)^{-1} X_K$$

$X_K = (1.0, 4.0, 2.0)$

$$X_K (X'X)^{-1} X'_K = 0.16529$$

Here $(X'X)^{-1}$ is computed as given in part (a) of this example.

$$\therefore S_{\hat{y}_k}^2 = 15.554 \times 0.16529$$

$$S_{\hat{y}_k} = 1.60$$

$$\therefore \text{lr} = 54.28 - 2.201 \times 1.60 = 50.76$$

$$\text{Ur} = 54.28 + 2.201 \times 1.60 = 57.80$$

Comments:

The hypothesis $H_o : \beta_2 = 0$ and $\beta_3 = 0$ were both tested in this example as though the tests were independent. In fact β_2 and β_3 are not independent. The Cov (β_2, β_3) can be determined from $C_{23}^{-1} S^2$ as 0.0612 x 15.554 = 0.9519. The correlation between β_2 and β_3 can be estimated from Cov $(\beta_2, \beta_3) / \sigma_{\beta_2} . \sigma_{\beta_3}$ as 0.9519/0.562 (2.85) = 0.59. The test of $H_o : \beta_3 = 0$ is made relative to the full model which includes all of the β's. The acceptance of H_o implies that $\beta_3 = 0$ given that β_1 and β_2 are in the model. In general, if there are p β's and $H_o : \beta_i = 0$ is tested for each of them with the result that K of the hypotheses can be accepted, one can not eliminate these K variables from the model on the basis of this test alone since each of the individual $H_o : \beta_i = 0$ assumes all of the other p-1 β's are still in the model. To eliminate K variable at once, the test must be based on the F-statistic computed using the equation:

$$F = \frac{(Q_2 - Q_2^*) / K}{Q_1 / (n-p)} \qquad \ldots(117)$$

where, Q_2 = sum of squares due to regression on the full model with (p-1) degrees of freedom;
Q_2^* = Sum of squares due to regression on the reduced model with (p-K-1) degrees of freedom;
Q_1 = Residual sum of squares on the full model with n-p degrees of freedom.

The statistic F will have an F distribution with K and n-p degrees of freedom. H_o is rejected if F exceeds $F_{(1-\alpha), K, n-p}$.

As an example of the application of the above equation, the $H_o : \beta_3 = 0$ will be tested. The ANOV for the full model is given during the solution of this example. The reduced model is simply $Y = \beta_1 + \beta_2 X$, where X is the watershed area in thousand of square km. Since this is a simple regression situation, we can compute the sum of squares due to regression from $b \sum (X_i - \bar{X}) (y_i - \bar{y})$ where $b = \sum (x_i - \bar{x}) (y_i - \bar{y}) / \sum (x_i - \bar{x})^2$. The result of this calculation is the sum of squares due to regression for the reduced model is 13,182.60.

$$\therefore F = \frac{(Q_2 - Q_2^*) / K}{Q_1 / (n-p)} = \frac{(13,182.60 - 13,182.60) / 1}{171.090 / 11} = 0.0$$

Since $F_{0.95, 1, 11} = 4.84$ so we accept $H_o : \beta_3 = 0$.

Since $H_0 : \beta_3 = 0$ was accepted, the next logical step is to eliminate I from the model and consider only A. In so doing the resulting regression equation is:

$$Q = 1.69 + 13.15 A$$

STEPWISE REGRESSION:

Most commonly used procedure for selecting the best regression equation is stepwise regression. This procedure consists of building the regression equation one variable at a time by adding at each step the variable that explains the largest amount of the remaining unexplained variation. After each step all the variables in the equation are examined for significance and discarded if they are no longer explaining a significant amount of the variation. Different steps involved in the stepwise regression are as follows:

(i) Add the first variable as the one which has highest simple correlation with dependent variable.

(ii) Add the second variable as the one which explains the largest variation in the dependent variable that remains unexplained by the first variable added.

(iii) Test the significance of first variable and retain or discard depending on the results of this test.

(iv) Add the third variable as the one that explains the largest portion of the variation that is not explained by the variables already in the equation.

(v) Test the variables in the equation for significance.

(vi) Repeat the steps (iv) & (v) until all of the variables not in the equation are found to be insignificant and all of the variables in the equation are significant.

This is a very good procedure to use but care must be exercised to see that the resulting equation is rational. Of course, the real test of how good the resulting regression model is depends on the ability of the model to predict the dependable variable for observations on the independent variables that were not used in estimating the regression coefficients. To make a comparison of this nature, it is necessary to randomly divide the data into two parts. One part of the data is then used to develop the model and the other part to test the model. Unfortunately, many times in hydrologic applications, there are not enough observations to carry out this procedure.

With regard to extrapolation, the comments given relative to simple regression are equally applicable to multiple regression. In multiple regression, an additional problem arises. It is some times difficult to tell the range of the data.

11.0 MULTIVARIATE ANLAYSIS

11.1 Principal Components Analysis

In multiple linear regression analysis a dependent variable is dependent on several other variables which are considered to be independent. When data are collected on these independent variables,

these variables are many times correlated. This correlation indicates that some of the information contained in one variable is also contained in some of the other remaining independent variables. The objective of the principal components analysis is transform the original correlated variables into uncorrelated or orthogonal components. These components are linear functions of the original variables. Such a transformation can be written as:

$$\underline{Z} = \underline{X}\,\underline{A} \qquad \text{....(118)}$$

where, $\underline{Z}$ is an nxp matrix of n values for each of p components,

$\underline{X}$ is an nxp matrix of n observations on p variables. Since we are dealing with variances and covariances, all x's will be assumed to be deviations from their respective means to that $\underline{X}$ is a matrix of deviations from means.

$\underline{A}$ is a pxp matrix of coefficients defining the linear transformation i.e. matrix of characteristic coefficients.

Other notation:

$$Z = \{Z_i, Z_2, \ldots Z_p\} = [\, Z_{ij} \,], i = 1, 2, 3, \ldots . n;\ j = 1, 2, 3, \ldots p$$

$$Z_j = [\, Z_{ij} \,], i = 1, 2, 3, \ldots n$$

Since the original p variate set of observations contained in $\underline{X}$ contains correlation, it might be possible to characterise the variance of $\underline{X}$ with q < p orthogonal components. Thus it is desired to construct $\underline{Z}$ so that each component, Z_j (an nx1 column vector) explains the maximum amount of the variance of $\underline{X}$ left unexplained by the first j-1 components. In this way it may be found that the first q components explain most of the system variance and that the last p-q components explain little of the system variance.

The principal components alongwith their correlation with the variables are computed in the following steps:

(i) Compute the variance-covariance matrix of $\underline{X}$ i.e. $\underline{S}$ using the relationships:

$$\underline{S} = [S_{ij}]\ i = 1, 2, \ldots p;\ j = 1, 2, \ldots p$$

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n} X_{K_j} X_{K_i} \qquad \text{....(119a)}$$

$$\text{or, } \underline{S} = X'X/(n-1) \qquad \text{....(119b)}$$

(ii) Compute the total system variance v defined as the sum of the variances of the original variables using the relationships:

$$V = Trace\ \ \underline{S} = \sum_{i=1}^{p} S_{ii} \qquad \text{....(120)}$$

(iii) Compute the characteristics roots, λ (also known as lagrangian Multiplier or Eigen value) solving the following equation:

$$|\, \underline{S} - \lambda I \,| = 0 \qquad \text{....(121)}$$

where, I is identity matrix.

(iv) Compute the coefficients of the characteristic vector for the first principal component solving:

$$(\underline{S} - \lambda_1 I)\, \underline{a}_1 = \underline{0} \qquad \ldots.(122)$$

Subject to the constraint: $\sum_{i=1}^{p} a_{i,1}^2 = 1$

(v) Repeat step (iv) to compute the coefficients of the characteristics vector corresponding to the other principal components.

(vi) Compute the values for principal components using:

$$\underline{Z} = \underline{X}\, \underline{A} \qquad \ldots.(123)$$

where, $\underline{A}$ is matrix of characteristics coefficients obtained from steps (iv) and (v).

(vii) Compute the elements of correlation matrix between the variables and the principal components using the relationship:

$$\text{Cor}\,(\underline{x}_i, \underline{z}_j) = \lambda_j^{1/2}\, a_{i,j}/S_i \qquad \ldots.(124)$$

For example the correlation between $\underline{x}_2$ and $\underline{z}_1$ is:

$$\text{Cor}\,(\underline{x}_2, \underline{z}_1) = \lambda_1^{1/2}\, a_{2,1}/S_2 \qquad \ldots.(125)$$

Some important properties of principal components:

(i) $\underline{z}_i$ and $\underline{z}_j$ are uncorrelated for i = j

(ii) $\text{Var}\,(\underline{z}_i) = a'_i\, \underline{S}\, \underline{a}_i = \lambda_i$ ....(126)

(iii) $\lambda_1 \geq \lambda_2 \geq \lambda_p$

(iv) $V = \text{Trace}\ \underline{S} = \sum_{K=1}^{p} \lambda_i = \sum_{K=1}^{p} Var\,(\underline{z}_i)$ ....(127)

(v) $\underline{z} = X\, A$, where, $\underline{z} = (\underline{z}_1, \underline{z}_2, \ldots \underline{z}_p)$ and $A = (\underline{a}_1, \underline{a}_2, \ldots \underline{a}_p)$

(vi) Fraction of the total variance accounted for by the j^{th} principal component is $\lambda_j/\text{trace}\ \underline{S}$.

Example:

Consider the data in the following table. Let $\underline{X}$ be a 13x3 matrix made up of 13 observations on A, S and L. Compute the principal components of $\underline{X}$ based on the covariance matrix. Compute the correlation between the variables and the components.

RO	PREC	A	S	L	P	D	RS	F	RR
17.38	44.37	2.21	50.00	2.38	7.93	0.98	0.38	1.36	332.00
14.62	44.09	2.53	7.00	2.55	7.65	1.23	0.48	2.37	55.00
15.48	41.25	5.63	19.00	3.11	11.61	2.11	0.57	2.31	77.00
14.72	45.50	1.55	6.00	1.84	5.31	0.94	0.49	3.87	68.00
18.37	46.09	5.15	16.00	4.14	11.35	1.63	0.39	3.30	68.00
17.01	49.12	2.14	26.00	1.92	5.89	1.41	0.71	1.87	230.00
18.20	44.03	5.34	7.0	4.73	12.59	1.30	0.27	0.94	44.00
18.95	48.71	7.47	11.00	4.24	12.33	2.35	0.52	1.20	72.00
13.94	44.43	2.10	5.00	2.00	6.81	1.19	0.53	4.76	40.00
18.64	47.72	3.89	18.00	2.10	9.87	1.65	0.60	3.08	115.00
17.25	48.38	0.67	21.00	1.15	3.93	0.62	0.48	2.99	352.00
17.48	49.00	0.85	23.00	1.27	3.79	0.83	0.61	3.53	300.00
13.16	47.03	1.72	05.00	1.93	5.19	0.99	0.52	2.33	39.00

Solution:

$\underline{S}$ is computed from equation (119).

$$\underline{S} = \begin{bmatrix} 4.465 & -4.519 & 2.177 \\ -4.519 & 155.769 & -2.955 \\ 2.177 & -2.955 & 1.322 \end{bmatrix}$$

$|\underline{S}-\lambda I|$ is computed as:

$$(\underline{S}-\lambda I) = \begin{bmatrix} 4.465-\lambda & -4.519 & 2.177 \\ -4.519 & 155.769-\lambda & -2.955 \\ 2.177 & -2.955 & 1.322-\lambda \end{bmatrix}$$

$$|\underline{S}-\lambda \angle| = (4.465-\lambda)(155.769-\lambda) + (-4.519)(-2.995)(2.177) + (2.177)(-4.519)(-2.955) - (2.177)^2 (155.769-\lambda) - (-4.519)^2 (1.322-\lambda) - (4.465-\lambda)(-2.955)^2 = 0$$

$$\lambda^3 - 161.548\lambda^3 + 872.130\lambda - 171.154 = 0$$

The solution to this cubic equation are:

$\lambda_1 = 155.963$
$\lambda_2 = 5.387$
$\lambda_3 = 0.207$

Note that $\Sigma\lambda_i$ = Trace $\underline{S}$ = 161.557

The first principal component accounts for $100\lambda_1$/Trace $\underline{S}$ = 100 (155.963)/ 161.557 = 96.54 percent of the total system variance. The coefficients of the characteristic vectors can be computed from equation (122). For example for the first principle component we have:

$$(\underline{S} - \lambda_1 I)\ \underline{a}_1 = \underline{0}$$

$$\begin{bmatrix} 4.465 - 155.963 & -4.519 & 2.177 \\ -4.519 & 155.769 - 155.963 & -2.955 \\ 2.177 & -2.955 & 1.322 - 155.933 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} = Q$$

or

$$\begin{aligned} -151.498\, a_{11} - 4.519\, a_{21} + 2.177\, a_{31} &= 0 \\ -4.519\, a_{11} - .194\, a_{21} - 2.955\, a_{31} &= 0 \\ 2.177\, a_{11} - 2.955\, a_{21} - 154.641\, a_{31} &= 0 \end{aligned}$$

Solving these three equations simultaneously for a_{11}, a_{21}, and a_{31}, results in:

$a_{21} = -51.43 a_{31}$ and $a_{11} = 1.5503 a_{31}$.

Using the constraint that $a^2_{11} + a^2_{21} + a^3_{31} = 1$, the solution is $a_{11} = 0.30$, $a_{21} = -.999$ and $a_{31} = .020$. Similarly for λ_2 and λ_3 we get:

$a_{12} = .892$ $a_{22} = .036$ $a_{32} = .451$
$a_{13} = -.452$ $a_{23} = .004$ $a_{33} = .892$

Thus,

$$A = \begin{bmatrix} .030 & .892 & -.452 \\ -.999 & .036 & .004 \\ .020 & .451 & .892 \end{bmatrix}$$

The values for the principal components can now be calculated from:

$$\underline{Z} = \underline{X}\,\underline{A}$$

The correlation matrix between the variables and the components can be computed from equation (124). For example the correlation between $\underline{x}_2$ and z_1 is:

$$Cor\ (\underline{x}_2, Z_1) = \lambda_1^{1/2}\ a_{21} / S_2 = 155.963^{1/2}\ (-0.999) / 155.769^{1/2} = -0.9995$$

The resulting correlation matrix is:

$$\begin{bmatrix} 0.178 & 0.979 & -0.097 \\ -1.000 & 0.007 & 0.000 \\ 0.212 & 0.911 & 0.353 \end{bmatrix}$$

The above example illustrates that using the $\underline{S}$ matrix in a principal component analysis presents some problems if the units of the X variables differ greatly. In the above example, the magnitude of the observations associated with the second variable were much greater than those associated with the other two variables. Consequently the variance of $\underline{x}_2$ was much greater than either Var ($\underline{x}_2$) or Var ($\underline{x}_3$). $\underline{x}_2$ accounted for 100 Var ($\underline{x}_2$)/Trace S or 96.4% of the system variance. This means that the first principal component is merely a restatement of $\underline{x}_2$. This can also be seen from the fact that the correlation between $\underline{x}_2$ and $\underline{z}_1$, is 1.000.

Principal Component Analysis for Standardized Variables:

In most hydrologic studies the problem of non-commensurate units on X's has been handled by standardizing the X's through the transformation $(x_{ij} - \bar{x}_j) / S_j$. The covariance matrix of the standardized variables becomes the correlation matrix i.e. $\underline{S}=\underline{R}$. The principal component analysis is then done on $\underline{R}$. The total system 'variance' now becomes Trace $\underline{R}$ = p since $\underline{R}$ has ones on the diagonal.

The principal component analysis for standarized variables involves the following steps:

(i) Compute the standardized variables y_{ij} using the relation:

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j} \qquad \ldots.(128)$$

(ii) Compute the elements r_{ij} of correlation matrix $\underline{R}$ using the relationship:

$$R = [r_{i,j}] = r'r / (n-1) \qquad \ldots.(129)$$

. . . . (129)

where, $\underline{R}$ is a symmetrical matrix since $r_{i,j} = r_{j,i}$

(iii) Compute the total system variance, V as:

$$V = \text{Trace } \underline{R} = P \qquad \ldots.(130)$$

(iv) Compute the characteristic roots, λ solving the following equation:

$$|\underline{R} - \lambda I| = 0 \qquad \ldots.(131)$$

(v) Compute the coefficients of characteristic vectors corresponding to each principal components solving:

$$(\underline{R} - \lambda_j I)\,\underline{a}_j = \underline{0} \qquad \ldots.(132)$$

Subject to the constraint: $\sum_{i=1}^{p} a_{ij} = 1$

(vi) Compute the numerical values of principal components as:

$$\underline{Z} = \underline{YA} \qquad \ldots.(133)$$

(vii) Compute the correlation matrix between the standardized variables and components, cor (y_i, z_j), using the relationship:

$$cor\ (\underline{y}_i, Z_j) = \lambda_j^{1/2} a_{ij} \qquad \ldots.(134)$$

These correlations are sometimes called factor loadings. The factor loadgings can be used to attach physical significance to the components. If a particular component is highly correlated with 1, 2, or 3 variables, then the component is a reflection of these variables. For example, in a study of watershed geomorphic factors, it might be found that a component is highly correlated with the average stream slope and the basin relief ratio. This being the case, that particular component might be termed a measure of watershed steepness.

Example:

Repeat the above example using $\underline{R}$ instead of $\underline{S}$.

Solution:

$$R = \begin{bmatrix} 1.000 & -.1713 & .8958 \\ -.1713 & 1.0000 & -.2059 \\ .8958 & -.2059 & 1.0000 \end{bmatrix}$$

$$|\underline{R} - \lambda I| = (1-\lambda)^3 - (1-\lambda)\,(.8768435) + .06343748 = 0$$

which has solutions:

$\lambda_1 = 1.9692$
$\lambda_2 = 0.9273$
$\lambda_3 = 0.1035$

In this formulation $\underline{z}_1$ accounts for 100 (1.9692)/3 or 65.64% of the system 'variance' while $\underline{z}_2$ and $\underline{z}_3$ account for 30.91% and 3.45% respectively.

The corresponding characteristic vectors are:

$$A = (a_1 a_2 a_3) = \begin{bmatrix} .679 & .208 & -.704 \\ -.265 & .964 & .029 \\ .684 & .167 & .710 \end{bmatrix}$$

The factor loadings computed from $\lambda_i^{1/2}$ a $_{ij}$ are:

$$\begin{bmatrix} .953 & .200 & -.226 \\ -.372 & .928 & .009 \\ .960 & .161 & .228 \end{bmatrix}$$

Since component 1 is highly correlated with both area and length, this component might be called a 'size' component. Likewise component 2 might be called a slope component. In terms of explaining the 'variance' of R, component 3 could be eliminated since it explains only 3.40% of the variance and is not correlated with any of the variables. We cannot eliminate any variables, however, since component 1 is strongly dependent on X_1, and X_3 while component 2 depends on X_2.

In terms of explaining the variance of R, we have reduced our problem from one of considering a 13x3 $\underline{X}$ matrix with correlations to a 13x2 $\underline{Z}$ matrix without correlations (assuming Z_3 is discarded).

The values for the components are computed from:

$$\underline{Z} = \underline{X}\underline{A}$$

where

$$\underline{X} = [(x_{ij} - \bar{x_j}) / S_j] = \begin{bmatrix} -.046 & -2.69 & -0.16 \\ -.030 & -.076 & -0.01 \\ 1.16 & 0.20 & 0.47 \\ -0.77 & -0.84 & -0.63 \\ 0.94 & -0.04 & 1.37 \\ -0.49 & 0.76 & -0.56 \\ 1.03 & -0.76 & 1.88 \\ 2.03 & -0.44 & 1.46 \\ -0.51 & -0.92 & -0.49 \\ 0.34 & 0.12 & -0.41 \\ -1.18 & 0.36 & -1.23 \\ -1.10 & 0.52 & -1.13 \\ -0.69 & -0.92 & -0.55 \end{bmatrix}$$

$$\underline{Z} = \underline{XA} = \begin{bmatrix} -1.13 & 2.47 & 0.28 \\ -.02 & -0.80 & 0.18 \\ 1.06 & 0.52 & -0.48 \\ -0.73 & -1.07 & -0.07 \\ 1.58 & 0.39 & 0.31 \\ -0.92 & 0.54 & -0.03 \\ 2.19 & -0.20 & 0.59 \\ 2.49 & 0.25 & -0.41 \\ -0.44 & -1.07 & -0.02 \\ -0.08 & 0.12 & -0.52 \\ -1.74 & -0.10 & -0.03 \\ -1.66 & -0.09 & -0.01 \\ -0.60 & -1.12 & -0.07 \end{bmatrix}$$

11.2 Multiple regression on principal components

Many times a principal components analysis is the first step in the development of a prediction model for some dependent variable, Y. Once the principal components are derived, they are used as the independent variables in a multiple regression anlaysis with the dependent variable, Y. Because of the differing units usually present in the original independent variables, the principal components are generally abstracted from the correlation matrix.

The steps involved in performing a multiple regression on principal components are outline here:

(i) Standardize the independent variables and centre the dependent variable so that $\underline{X} = [x_{ij}]$ and $\underline{Y} = [y_i]$.

where,

$$x_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j} \text{ and } y_i = Y_i - \bar{Y}$$

where, Y_i is the ith observation on Y. $\bar{Y}$ is the mean of Y, X_{ij} is the ith observation on jth variable and $\bar{X}_j$ and S_j are mean and standard deviation of jth variable. Centring Y is not necessary. It eliminates the need for an intercept and simplifies notation.

(ii) Compute the characteric roots λ_j and corresponding vectors a_j using the procedure described in the previous section of this lecture. Then determine the matrix of principal components $\underline{Z}$ from $\underline{Z} = \underline{X}\,\underline{A}$ with $\underline{A}$ being a p x p matrix whose jth column is a_j, the characteristic vector.

(iii) Develop the regression model:

$$\underline{Y} = \underline{Z}\,\beta \text{ or } Y_i = \sum_{j=1}^{p} \beta_j Z_{ij} \qquad \ldots\ldots(135)$$

where, $\underline{Y}$ is an n x 1 vector whose elements are the n observations of the centred dependent variable, $\underline{Z}$ is an n x p matrix whose elements, Z_{ij} represent the ith value of the jth principal component.

(iv) Estimate $\underline{\beta}$ using the following relationship:

$$\beta_j = \frac{\bar{a}_j \underline{X}\, \underline{Y}}{(n-1)\lambda_j} \qquad \ldots\ldots(136)$$

(v) Estimate the variances and covariances of β, i.e.

$$\text{Cov}\,(\beta_i, \beta_j) = 0 \text{ for } i \neq j \qquad \ldots\ldots(137)$$

$$Var\,(\beta_j) = \frac{\sigma^2}{(n-1)\,\lambda_j} \text{ for } i = j \qquad \ldots\ldots(138)$$

where, σ is the standard error of the regression equation. Cov $(\beta_i, \beta_j) = 0$ for $i \neq j$ indicates that β_i is independent of β_j for $i = j$.

(vi) Perform t-test to judge statistical significance of β_j with null hypothesis: $H_o : B_j = 0$ (note that $\beta_o = 0$). The t-test statistics is computed using relationship:

$$t = \frac{\bar{a}_j \underline{X}' \underline{Y}}{\sqrt{(n-1)\lambda_j}\sigma} \qquad \ldots\ldots(139)$$

There is no reason to believe before the regression is performed that this test statistic will be non-significant for small values of λ_j. Therefore, the regression should be performed on all of the components and then the components that prove to be non-significant can be eliminated.

(vii) Transform the resulting regression equation into an equation in terms of the original X variables. This can be done since:

$$\underline{Y_i} = Y_i - \bar{Y}, \quad x_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j}$$

$Z_{ij} = \sum_{K=1}^{p} a_{kj} X_{ik}$ *and* $\beta_j's$ are constant. Thus the regression equation becomes:

$$Y_i = \bar{Y} + \sum_{j=1}^{p} \hat{\beta}_j \sum_{k=1}^{p} a_{kj} \left[\frac{(X_{ik} - \bar{X}_k)}{S_k}\right] \qquad \ldots\ldots(140)$$

Which may be further simplified by collecting the different terms as:

$$Y_i = \beta_o^* + \sum_{j=1}^{p} \beta_j^* X_{ij} \qquad \ldots . (141)$$

where, the β^*'s are constants. If only q ($q<p$) components are retained in the final regression equation and the components are rearranged so that the first q-components are retained, the first summation in Eq. (141) would run from 1 to q; however, the second summation would still run from 1 to p. This means the summation in Eq. (141) would run from 1 to p. It also means that even though the equation contains only q components, all p of the original variables must be measured to predict Y.

Advantages & Limitations of Regression Principal Components

In the linear regression on principle components β_j is independent of β_j for $i \neq j$. Therefore the numerical value for the $\beta's$ retained in the regression will not be altered by eliminating any number of other $\beta's$. This is the distinct advantage of having an orthogonal matrix of independent varaibles.

A second advantage of having independent $\beta's$ is that the interpretation of β's in terms of the independent variables is greatly simplified. Thus if some hydrologic meaning can be attached to a component through an examination of the factor loadings, hydrologic significance can also be attached to the Unfortunately in most hydrologic applications of principal components analysis, a clear and distinct interpretation of the principal components has not been possible. This in turn means the hydrologic significance of the is unclear as well.

Another advantage for using regression on principal components as compared to normal multiple regression is that the resulting regression coefficients are more stable when applied to a new set of data because of coefficients are fitted on the basis of only statistically significant orthogonal components. This could imply that using an equation based on regression on principal components for prediction on a sample not included in the equation development would have a smaller standard error on this sample than would a normal multiple regression equation. If this is the case, it would be an important advantage for the regression on principal components technique. Adequate demonstration of this hypothesis needs to be developed, however.

A disadvantage of using principal components in a regression is that even if all but one of the components is eliminated, all of the original variables (the X's) must still be measured since each component is a function of all of the X's.

Some of the original X variables can be eliminated from the analysis before any regressions are performed by examining the factor loadings and eliminating variable that are not highly correlated with any of the components. The remaining X variables are then resubmitted to a principal components analysis with the multiple regression being performed on the new components. This procedure has advantage of reducing the number of variables that must be measured to use the resulting regression equation. It has the disadvantage of eliminating X variables rather arbitrarily (there is no statistical test for the significance of the factor loadings) without ever having them in a position to determine their usefulness in explaining the variatiion in the dependent variable Y.

11.3 Factor Anlaysis

The purpose of factor anlaysis is to partition a p-variate observed vector into some factors common to all of the p variables and some factors unique to each of the p variables. A factor model might be written:

$$\underset{pxl}{\underline{X}} = \underset{pxm}{\underline{G}} \; \underset{mxl}{\underline{F}} + \underset{pxp}{\underline{H}} \; \underset{pxl}{\underline{U}} \qquad \ldots . (142)$$

where, $\underline{X}$ is a p-variate vector of observed variables, $\underline{G}$ is a matrix of coefficients, $\underline{F}$ is a vector of common factors, $\underline{H}$ is a diagonal matrix of coefficients, and $\underline{U}$ is a vector of unique factors. A common application of factor analysis in hydrology has been conducted by ignoring the unique factors and considering the common factors. Further discussion about the application of factor analysis in hydrology may be found elsewhere (Matala and Reichor, 1967; and Wallis, 1967).

11.4 Cluster Analysis

The main objective of a regional analysis is to develop regional regression models which can be used to estimate the hydrologic variables at ungauged sites. Hydrologic data from several gauging stations in hydrologically homogeneous regions are collected and analysed to obtain estimates of the regression parameters. Identification of these hydrological homogeneous regions is a vital component in any regional analysis. One method used to identify these regions is a multivariate statistical procedure known as cluster.

Cluster anlaysis is a method used to group objects with similar characteristics. Two clustering methods are used for this purpose. The first group of procedures are known as hierarchical method, and they attempt to group objects by a series of successive mergers. The most similar objects are first grouped and as the similarity decreases, all subgroups are progressively merged into a single cluster. The second group is collectively referred to as nonhierarchical clustering techniques and, if required, can be used to group objects into a specified number of clusters. The clustering process starts from an initial set of seed points, which will form the nuclei of the final clusters.

The most commonly used similarity measure in cluster anlaysis is the Euclidean distance defined by:

$$D_{ij} = \left[\sum_{k=1}^{p} (z_{ik} - z_{jk})^2 \right]^{1/2} \qquad \ldots . (143)$$

where D_{ij} is the Euclidean distance from site i to site j, p is the number of variables inlcuded in the computation of the distance (i.e. the basin and climatic variables, and z_{ik} is a standardized value for variable k at site i.

In many applications the variables describing the objects to be clustered (discharges, watershed areas, stream lengths, etc.) will not be measured in the same units. It is reasonable to assume that it would not be sensible to treat say, discharge measured in cubic meter per second, area in square kilometer and stream length in kilometer as equivalent in determining a measure of similarity. The solution suggested most often is to standardize each variable to unit variance prior to anlaysis. This is done

by dividing the variables by the standard deviations calculated from the complete set of objects to be clustered. The standardization process eliminates the units from each variable and reduced any differences in the range of values among the variables.

To get a feel for how cluster analysis works, consider six precipitation stations and their associated annual precipitation in mm:

Station	1	2	3	4	5	6
Precipitation	1000	1200	600	700	500	1100

It is desired to see if these stations can be grouped into homogeneous groups based on the average annual precipitation

The first thing that is done is to standardize the precipitation values. For this set of data, the mean is 850 and the standard deviation is 288. Exhibit contains the data and results. Equation (143) is used to calculate D_{ij}. For example $D_{1,2}$ is $\sqrt{(0.52 - 1.21)^2}$ which equals (0.52 - 1.21) or 0.69. The results for all of the $D_{i,j}$ are shown in Table A of Exhibit 1.

The next step is to find the minimum value of the similarity measure, $D_{i,j}$. This value is seen to be 0.35. The value 0.35 appears several times. The pair (3,4) was arbitrarily chosen as the first similar pair. Table B of exhibit 1 contains the $D_{i,j}$ values from Table 1 except for the (3,4) row. This row contains the minimum of $D_{3,j}$ and $D_{4,j}$ for $j = 1, 2, 5$, and 6. For example $D_{3,1}$ is 1.39 and $D_{4,1}$ is 1.04. Therefore, the (3, 4), 1 entry in Table B is 1.04. Other values in the (3, 4) row are similarly determined.

Again the minimum entry in Table B is found to be 0.35 corresponding to the (1, 6) pair. Thus (1, 6) is clustered as in Table C and entries for Table C are determined from Table B in the same manner as entries in Table B were determined from Table A. The next step results in (1, 6) and 2 being clustered to form (1, 2, 6). This is followed by (3, 4) being clustered with 5 to form (3, 4, 5).

Exhibit 2 is similar to Exhibit 1 except the value of precipitation for the third station is changed from 600 to 1050 mm. Carrying through the anlaysis as was done for Exhibit 1 results in forming the clusters (4, 5) and (1, 2, 3, 6).

In Exhibit 3, the third station value is changed to 1800 mm. The cluster results are (1, 2, 4, 5, 6) and 3. In all of these analyses, the $D_{i,j}$ entry is a measure of the similarity that exists. For example in Exhibit 3, the $D_{i,j}$ values of 0.22 indicate strong similarity. The values of 0.44 shows that stations 4 and 5 are not as similar as are stations 1, 2, and 6. The value 0.67 shows that the cluster (4, 5) and (1, 2, 6) are less similar than either 4 and 5 or 1, 2, and 6. Finally the value 1.33 shows that 3 is not very similar to the cluster (1, 2, 4, 5, 6).

Clustering may stop when there is a significant jump in the similarity measure. In Exhibit C one might conclude with three clusters (1, 2, 6), (4, 5), and (3), or with two clusters (1, 2, 4, 5, 6) and 3.

Exhibit 4 extends the anlaysis to consideration of two measures of the stations being considred, precipitation and potential evapotranspiration. Again Table A was constructed from equation (143). For example the $D_{1,2}$ entry is calculated from standardized values of:

$D_{1,2} = \sqrt{(-0.11 - 0.33)^2 + (-1.21 - 1.21)^2}$ *or* 2.47.

The analysis is completed based on Table A in the same manner as for Exhibits 1-3. Here a satisfying clustering doesn't appear to exist. It looks as though 2 and 6 might be clustered but possibly the other stations can not be clustered.

Exhibit 5 is based on the rati of precipitation over potential evapotranspiration. Using this system measure, 2, 4, and 6 certainly form a cluster. Depending on the purpose of the analysis, one might conclude that: (1, 3) and (2, 4, 5, 6) represent the final clustering.

12.0 REMARKS

Statistical analysis of hydrological variables provide useful information about the nature of distribution, the data on the hydrological variabless follow. It may be used to predict the magnitude and associated frequency. Regression techniques are being widely used for developing the prediction equations for different hydrological variables. These prediction equations are useful for estimating the dependent hydrological variables, which are difficult to be monitored, based on the independent variables which can be easily monitored. Multivariate analysis techniques such as principal component analysis, cluster anlaysis techniques, etc. are also applied for solving many of the hydrological problems.

BIBLIOGRAPHY

1. Chow, V.T., 1964. *Hand book of Applied Hydrology*. Mc-Graw Hill New York.

2. Haan, C.T., 1977. *Statistical Methods in Hydrology*. The Iowa State University Press, Amer., U.S.A.

3. Kite, G.W., 1977. *Frequency and Risk Analysis in Hydrology*. Water Resources Publications, Colorado.

4. Linseley, R.K., Kohler, H.A. and Paulhus, J.L.H., 1975. *Hydrology for Engineers*. Mc-Graw Hill, International Book Company.

5. Matalar, N.C. and B. J. Reiher, 1967. *Some comments on the use of factor anlaysis*. Water Resources Research, Vol. 3(1):213-224.

6. McGuess, R.H. and Snyder, W.M., 1985. *Hydrologic Modelling, Statistical Methods and Applications*. Prentice-Hall, Englewood Cliffs, New Jersey.

7. Wallis, J.R., 1967. *When is it safe to extend a prediction equation? - An answer based on factor and discriminant function analysis*. Water Resources Research, Vol. 3 (2) : 375-384.

8. Yevjevich, V., 1972. *Probability and Statistics in Hydrology*. Water Resources Publications, Fort Collins, Colorado.

** *** **

Exhibit 1

sta	1	2	3	4	5	6	mean	stdev
precip	1000	1200	600	700	500	1100	850	288
z	0.52	1.21	-0.87	-0.52	-1.21	0.87	0	1

Table A.

	1	2	3	4	5	6
1	0.00	0.69	1.39	1.04	1.74	0.35
2		0.00	2.08	1.74	2.43	0.35
3			0.00	[0.35]	0.35	1.74
4				0.00	0.69	1.39
5					0.00	2.08
6						0.00

Table B

	3,4	1	2	5	6
3,4	0	1.04	1.74	0.35	1.39
1		0	0.69	1.74	[0.35]
2			0	2.43	0.35
5				0	2.08
6					0

Table C

	1,6	3,4	2	5
1,6	0	1.04	[0.35]	1.74
3,4		0	1.74	0.35
2			0	2.43
5				0

Table D

	1,2,6	3,4	5
1,2,6	0	1.04	1.74
3,4		0	[0.35]
5			0

Table E

	3,4,5	1,2,6
3,4,5	0	[1.04]
1,2,6		0

Exhibit 2

sta	1	2	3	4	5	6	mean	stdev
precip	1000	1200	1050	700	500	1100	925	268
z	0.28	1.03	0.47	-0.84	-1.59	0.65	0	1

Table A

	1	2	3	4	5	6
1	0.00	0.75	0.19	1.12	1.87	0.37
2		0.00	0.56	1.87	2.61	0.37
3			0.00	1.31	2.05	[0.19]
4				0.00	0.75	1.49
5					0.00	2.24
6						0.00

Table B

	3,6	1	2	4	5
3,6	0	[0.19]	0.37	1.31	2.05
1		0	0.75	1.12	1.87
2			0	1.87	2.61
4				0	0.75
5					0

Table C

	1,3,6	2	4	5
1,3,6	0	[0.37]	1.12	1.87
2		0	1.87	2.61
4			0	0.75
5				0

Table D

	1,2,3,6	4	5
1,2,3,6	0	1.12	1.87
4		0	[0.75]
5			0

Table E

	4,5	1,2,3,6
4,5	0	[1.12]
1,2,3,6		0

Exhibit 3

sta	1	2	3	4	5	6	mean	stdev
precip	1000	1200	1800	700	500	1100	1050	451
z	-0.11	0.33	1.66	-0.78	-1.22	0.11	-7E-18	1

Table A

	1	2	3	4	5	6
1	0.00	0.44	1.78	0.67	1.11	**0.22**
2		0.00	1.33	1.11	1.55	0.22
3			0.00	2.44	2.89	1.55
4				0.00	0.44	0.89
5					0.00	1.33
6						0.00

Table B

	1,6	2	3	4	5
1,6	0	**0.22**	1.55	0.67	1.11
2		0	1.33	1.11	1.55
3			0.00	2.44	2.89
4				0.00	0.44
5					0.00

Table C

	1,2,6	3	4	5
1,2,6	0	1.33	0.67	1.11
3		0	2.44	2.89
4			0.00	**0.44**
5				0

Table D

	4,5	1,2,6	3
4,5	0	**0.67**	2.44
1,2,6		0	1.33
3			0

Table E

	1,2,4,5,6	3
1,2,4,5,6	0	**1.33**
3		0

Exhibit 4

sta	1	2	3	4	5	6	mean	stdev
precip	1000	1200	1800	700	500	1100	1050	451
z1	-0.11	0.33	1.66	-0.78	-1.22	0.11	-7E-18	1
PET	500	1200	600	700	1000	1100	850	288
z2	-1.21	1.21	-0.87	-0.52	0.52	0.87	2E-17	1

Table A

	1	2	3	4	5	6
1	0.00	2.47	1.81	0.96	2.06	2.09
2		0.00	2.47	2.06	1.70	**0.41**
3			0.00	2.47	3.20	2.33
4				0.00	1.13	1.65
5					0.00	1.38
6						0.00

Table B

	2,6	1	3	4	5
2,6	0	2.09	2.33	1.65	1.38
1		0	1.81	**0.96**	2.06
3			0.00	2.47	3.20
4				0.00	1.13
5					0.00

Table C

	1,4	2,6	3	5
1,4	0	1.65	1.81	**1.13**
2,6		0	2.33	1.38
3			0.00	3.20
5				0

Table D

	1,4,5	2,6	3
1,4,5	0	**1.38**	1.81
2,6		0	2.33
3			0

Table E

	1,2,4,5,6	3
1,2,4,5,6	0	**1.81**
3		0

Exhibit 5

sta	1	2	3	4	5	6	mean	stdev
ratio	2	1	1.75	1	0.5	1	1.2083	1
z	1.42	-0.37	0.97	-0.37	-1.27	-0.37	2E-16	1

Table A

	1	2	3	4	5	6
1	0.00	1.79	0.45	1.79	2.69	1.79
2		0.00	1.35	0.00	0.90	0.00
3			0.00	1.35	2.24	1.35
4				0.00	0.90	0.00
5					0.00	0.90
6						0.00

Table B

	2,4	1	3	5	6
2,4	0	1.79	1.35	0.9	0
1		0	0.45	2.69	1.79
3			0.00	2.24	1.35
5				0.00	0.90
6					0.00

Table C

	2,4,6	1	3	5
2,4,6	0	1.79	1.35	0.9
1		0	0.45	2.69
3			0.00	2.24
5				0

Table D

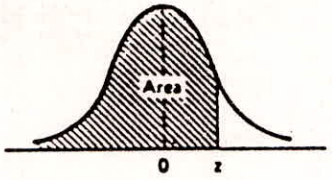
	1,3	2,4,6	5
1,3	0	1.35	2.24
2,4,6		0	0.9
5			0

Table E

	2,4,5,6	1,3
2,4,5,6	0	1.35
1,3		0

APPENDIX - I

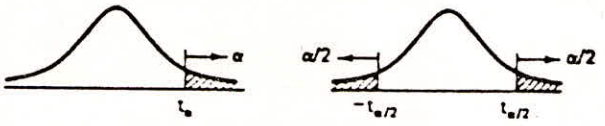
TABLE - I Normal Distribution Probabilities



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

APPENDIX-II

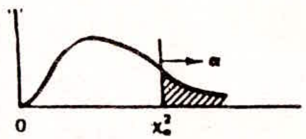
TABLE-II t-Distribution Probabilities



	Level of Significance for One-Tailed Test							
	.250	.100	.050	.025	.010	.005	.0025	.0005
	Level of Significance for a Two-Tailed Test							
Degrees of Freedom	.500	.200	.100	.050	.020	.010	.005	.001
1.	1.000	3.078	6.314	12.706	31.821	63.657	27.321	536.627
2.	.816	1.886	2.920	4.303	6.965	9.925	14.089	31.599
3.	.765	1.638	2.353	3.182	4.541	5.841	7.453	12.924
4.	.741	1.533	2.132	2.776	3.747	4.604	5.598	8.610
5.	.727	1.476	2.015	2.571	3.365	4.032	4.773	6.869
6.	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.959
7.	.711	1.415	1.895	2.365	2.998	3.499	4.029	5.408
8.	.706	1.397	1.860	2.306	2.896	3.355	3.833	5.041
9.	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.781
10.	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.587
11.	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.437
12.	.695	1.356	1.782	2.179	2.681	3.055	3.428	4.318
13.	.694	1.350	1.771	2.160	2.650	3.012	3.372	4.221
14.	.692	1.345	1.761	2.145	2.624	2.977	3.326	4.140
15.	.691	1.341	1.753	2.131	2.602	2.947	3.286	4.073
16.	.690	1.337	1.746	2.120	2.583	2.921	3.252	4.015
17.	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.965
18.	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.922
19.	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.883
20.	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.850
21.	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.819
22.	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.792
23.	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.768
24.	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.745
25.	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.725
26.	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.707
27.	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.690
28.	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.674
29.	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.659
30.	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.646
35.	.682	1.306	1.690	2.030	2.438	2.724	2.996	3.591
40.	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.551
45.	.680	1.301	1.679	2.014	2.412	2.690	2.952	3.520
50.	.679	1.299	1.676	2.009	2.403	2.678	2.937	3.496
55.	.679	1.297	1.673	2.004	2.396	2.668	2.925	3.476
60.	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.460
65.	.678	1.295	1.669	1.997	2.385	2.654	2.906	3.447
70.	.678	1.294	1.667	1.994	2.381	2.648	2.899	3.435
80.	.678	1.292	1.664	1.990	2.374	2.639	2.887	3.416
90.	.677	1.291	1.662	1.987	2.368	2.632	2.878	3.402
100.	.677	1.290	1.660	1.984	2.364	2.626	2.871	3.390
125.	.676	1.288	1.657	1.979	2.357	2.616	2.858	3.370
150.	.676	1.287	1.655	1.976	2.351	2.609	2.849	3.357
200.	.675	1.286	1.653	1.972	2.345	2.601	2.839	3.340
∞	.6745	1.2816	1.6448	1.9600	2.3267	2.5758	2.8070	3.2905

APPENDIX – III

TABLE – III χ^2 Distribution



Degrees of Freedom \ α	.999	.995	.990	.975	.950	.900	.700	.500	.200	.100	.050	.020	.010	.005	.001
1.	.000	.000	.000	.001	.004	.016	.148	.455	1.642	2.706	3.842	5.405	6.637	7.905	10.809
2.	.002	.010	.020	.051	.103	.211	.713	1.386	3.219	4.604	5.995	7.822	9.221	10.589	13.691
3.	.024	.072	.115	.216	.352	.584	1.424	2.366	4.642	6.252	7.817	9.841	11.325	12.838	16.292
4.	.090	.205	.297	.484	.711	1.064	2.195	3.357	5.989	7.765	9.492	11.660	13.280	14.862	18.459
5.	.209	.411	.553	.831	1.145	1.610	3.000	4.352	7.291	9.237	11.073	13.385	15.088	16.762	20.751
6.	.377	.673	.871	1.236	1.635	2.204	3.828	5.349	8.559	10.658	12.596	15.033	16.810	18.550	22.477
7.	.597	.988	1.237	1.688	2.167	2.833	4.671	6.346	9.804	12.020	14.070	16.624	18.471	20.270	24.327
8.	.850	1.341	1.642	2.179	2.732	3.489	5.527	7.344	11.031	13.363	15.512	18.171	20.082	21.938	26.119
9.	1.135	1.728	2.086	2.699	3.324	4.168	6.393	8.343	12.243	14.686	16.925	19.683	21.654	23.583	28.061
10.	1.446	2.152	2.555	3.244	3.938	4.864	7.266	9.342	13.443	15.990	18.311	21.165	23.194	25.154	29.765
11.	1.819	2.597	3.047	3.815	4.576	5.576	8.148	10.341	14.633	17.278	19.683	22.633	24.733	26.716	31.633
12.	2.188	3.064	3.568	4.402	5.225	6.303	9.034	11.340	15.813	18.551	21.030	24.056	26.246	28.249	33.070
13.	2.577	3.560	4.102	5.006	5.890	7.041	9.926	12.340	16.986	19.814	22.367	25.477	27.717	29.870	34.693
14.	3.018	4.066	4.653	5.644	6.568	7.789	10.821	13.339	18.152	21.067	23.691	26.879	29.169	31.370	36.373
15.	3.449	4.588	5.226	6.260	7.260	8.546	11.721	14.339	19.317	22.310	25.000	28.266	30.605	32.857	37.843
16.	3.936	5.133	5.807	6.905	7.960	9.311	12.626	15.339	20.466	23.546	26.301	29.640	32.027	34.331	39.292
17.	4.350	5.687	6.400	7.560	8.670	10.083	13.550	16.338	21.616	24.771	27.593	31.002	33.455	35.771	40.926
18.	4.864	6.231	7.004	8.225	9.388	10.864	14.440	17.337	22.761	25.992	28.877	32.553	34.811	37.248	42.444
19.	5.351	6.825	7.627	8.904	10.114	11.650	15.351	18.338	23.902	27.208	30.148	33.696	36.216	38.633	43.949
20.	5.848	7.422	8.252	9.587	10.849	12.442	16.265	19.337	25.039	28.415	31.418	35.028	37.591	40.046	45.440
21.	6.398	8.018	8.886	10.278	11.590	13.238	17.182	20.337	26.171	29.619	32.678	36.350	38.957	41.469	46.919
22.	6.919	8.622	9.528	10.976	12.336	14.040	18.100	21.337	27.304	30.817	33.933	37.668	40.314	42.843	48.387
23.	7.447	9.247	10.187	11.685	13.088	14.846	19.020	22.337	28.431	32.012	35.178	38.975	41.662	44.220	49.853
24.	8.027	9.869	10.846	12.397	13.845	15.657	19.945	23.337	29.558	33.199	36.421	40.277	43.004	45.595	51.293
25.	8.576	10.498	11.510	13.115	14.607	16.471	20.866	24.337	30.678	34.384	37.660	41.573	44.338	46.973	52.733
26.	9.130	11.132	12.190	13.837	15.377	17.291	21.792	25.337	31.796	35.566	38.896	42.863	45.663	48.336	54.187
27.	9.735	11.789	12.868	14.565	16.149	18.113	22.718	26.336	32.913	36.745	40.119	44.147	46.986	49.666	55.584
28.	10.306	12.438	13.551	15.304	16.925	18.938	23.646	27.336	34.028	37.920	41.344	45.426	48.301	51.036	56.998
29.	10.882	13.093	14.220	16.042	17.705	19.766	24.576	28.336	35.140	39.092	42.563	46.699	49.610	52.378	58.401
30.	11.509	13.767	14.943	16.784	18.488	20.598	25.507	29.336	36.251	40.261	43.782	47.968	50.914	53.713	59.805
35.	14.554	17.160	18.496	20.563	22.461	24.793	30.178	34.339	41.787	46.063	49.811	54.250	57.398	60.313	66.734
40.	17.846	20.669	22.139	24.425	26.508	29.053	34.878	39.339	47.261	51.796	55.753	60.445	63.710	66.802	73.490
45.	21.183	24.273	25.878	28.357	30.611	33.355	39.591	44.337	52.721	57.497	61.652	66.562	69.978	73.201	80.165
50.	24.609	27.957	29.685	32.349	34.763	37.693	44.319	49.336	58.157	63.159	67.501	72.619	76.178	79.523	86.745
55.	28.111	31.762	33.549	36.390	38.957	42.064	49.061	54.336	63.571	68.789	73.308	78.625	82.309	85.780	93.245
60.	31.678	35.503	37.465	40.476	43.187	46.463	53.814	59.336	68.966	74.390	79.078	84.586	88.396	91.982	99.679
65.	35.303	39.353	41.424	44.596	47.449	50.882	58.578	64.336	74.344	79.966	84.817	90.507	94.438	98.156	106.057
70.	38.980	43.249	45.423	48.750	51.739	55.155	63.351	69.335	79.709	85.521	90.528	96.393	100.441	104.243	112.383
80.	46.466	51.145	53.551	57.147	60.391	64.282	72.920	79.335	90.400	96.572	101.876	108.075	112.366	116.348	124.901
90.	54.104	59.171	61.734	65.641	69.126	73.285	83.515	89.335	101.049	107.559	113.143	119.656	124.130	128.324	137.267
100.	61.869	67.303	70.049	74.216	77.929	82.362	92.133	99.335	111.667	118.493	124.340	131.147	135.820	140.193	149.505
125.	81.726	88.007	91.166	95.941	100.178	105.217	116.256	124.335	138.071	145.638	152.092	159.580	164.706	169.495	179.651
150.	102.073	109.132	112.655	117.980	122.692	128.278	140.460	149.336	164.365	172.577	179.579	187.663	193.219	198.380	209.310
200.	145.807	152.224	156.421	162.724	168.279	174.858	189.051	199.334	216.605	226.017	233.993	243.191	249.455	255.281	267.379

Table iv Cumulative F Distribution (m Numerator and n Denominator Degrees of Freedom).

α	n \ m	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
.90		39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.3	62.8	63.1	63.3
.95		161	200	216	225	230	234	237	239	241	242	244	246	248	250	252	253	254
.975	1	648	800	864	900	922	937	948	957	963	969	977	985	993	1000	1010	1010	1020
.99		4,050	5,000	5,400	5,620	5,760	5,860	5,930	5,980	6,020	6,060	6,110	6,160	6,210	6,260	6,310	6,340	6,370
.995		16,200	20,000	21,600	22,500	23,100	23,400	23,700	23,900	24,100	24,200	24,400	24,600	24,800	25,000	25,200	25,400	25,500
.90		8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.46	9.47	9.48	9.49
.95		18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
.975	2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5
.99		98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5
.995		199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
.90		5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17	5.15	5.14	5.13
.95		10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.57	8.55	8.53
.975	3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	14.1	14.0	13.9	13.9
.99		34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.5	26.3	26.2	26.1
.995		55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7	43.4	43.1	42.8	42.5	42.1	42.0	41.8
.90		4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.93	3.92	3.90	3.87	3.84	3.82	3.79	3.78	3.76
.95		7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.69	5.66	5.63
.975	4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.46	8.36	8.31	8.26
.99		21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.8	13.7	13.6	13.5
.995		31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.7	20.4	20.2	19.9	19.6	19.5	19.3
.90		4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.17	3.14	3.12	3.11
.95		6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.43	4.40	4.37
.975	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.23	6.12	6.07	6.02
.99		16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.38	9.20	9.11	9.02
.995		22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.4	13.1	12.9	12.7	12.4	12.3	12.1
.90		3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.80	2.76	2.74	2.72
.95		5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.74	3.70	3.67
.975	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.07	4.96	4.90	4.85
.99		13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.06	6.97	6.88
.995		18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.2	10.0	9.81	9.59	9.36	9.12	9.00	8.88
.90		3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.56	2.51	2.49	2.47
.95		5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.30	3.27	3.23
.975	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.36	4.25	4.20	4.14
.99		12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.82	5.74	5.65
.995		16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.53	7.31	7.19	7.08
.90		3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.38	2.34	2.31	2.29
.95		5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.01	2.97	2.93
.975	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.89	3.78	3.73	3.67
.99		11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.03	4.95	4.86
.995		14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.40	6.18	6.06	5.95

Table IV Contd.....

.90		3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.25	2.21	2.18	2.16
.95		5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.79	2.75	2.71
.975	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.56	3.45	3.39	3.33
.99		10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.48	4.40	4.31
.995		13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.62	5.41	5.30	5.19
.90		3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.15	2.11	2.08	2.06
.95		4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.70	2.62	2.58	2.54
.975	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.31	3.20	3.14	3.08
.99		10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.08	4.00	3.91
.995		12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.07	4.86	4.75	4.64
.90		3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.01	1.96	1.93	1.90
.95		4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.38	2.34	2.30
.975	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.96	2.85	2.79	2.72
.99		9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.54	3.45	3.36
.995		11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.33	4.12	4.01	3.90
.90		3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.87	1.82	1.79	1.76
.95		4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.16	2.11	2.07
.975	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.64	2.52	2.46	2.40
.99		8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.05	2.96	2.87
.995		10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.69	3.48	3.37	3.26
.90		2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.74	1.68	1.64	1.61
.95		4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.95	1.90	1.84
.975	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.35	2.22	2.16	2.09
.99		8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.61	2.52	2.42
.995		9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.12	2.92	2.81	2.69
.90		2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.61	1.54	1.50	1.46
.95		4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.74	1.68	1.62
.975	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2 31	2.20	2.07	1.94	1.87	1.79
.99		7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.21	2.11	2.01
.995		9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.63	2.42	2.30	2.18
.90		2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.48	1.40	1.35	1.29
.95		4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.53	1.47	1.39
.975	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.82	1.67	1.58	1.48
.99		7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.84	1.73	1.60
.995		8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.19	1.96	1.83	1.69
.90		2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.54	1.48	1.41	1.32	1.26	1.19
.95		3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.43	1.35	1.25
.975	120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.69	1.53	1.43	1.31
.99		6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.66	1.53	1.38
.995		8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	1.98	1.75	1.61	1.43
.90		2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.34	1.24	1.17	1.00
.95		3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.32	1.22	1.00
.975	∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.57	1.39	1.27	1.00
.99		6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.47	1.32	1.00
.995		7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.79	1.53	1.36	1.00