SOME ELEMENTS OF STATISTICS for Agriculturists and Foresters

(Training Manual 2)

by

Jacques Antoine
FAO Computer Expert

FAO/UNDP ASSISTANCE TO THE SECOND AGRICULTURE RESEARCH PROJECT
BGD/83/010

BARC COMPUTER CENTRE
DHAKA, AUGUST 1986

SOME ELEMENTS

OF

STATISTICS

for

Agriculturists and Foresters

(Training Manual 2)

by

Jacques Antoine

FAO Computer Expert

FAO/UNDP ASSISTANCE TO THE SECOND
AGRICULTURE RESEARCH PROJECT

BGD/83/010

BARC COMPUTER CENTRE

Dhaka, August 1986

P R E F A C E

This manual is a compilation of the Course Materials used in Part 2 of a Computer Programming Course for scientific programming trainees from BARC (Bangladesh Agriculture Research Council), BARI (Bangladesh Agriculture Research Institute), BRRI, (Bangladesh Rice Research Institute), BJRI, (Bangladesh Jute Research Institute), SRDI (Soil Research Development Institute) which is being held at BARC Computer Centre.

It is a follow on to Training Manual 1 on Basic Computer Mathematics which were taught in Part 1 of the course. It contains some basic concepts of statistics and statistical inferences and a set of selected standard statistical procedures and tests in parametric statistics. The emphasis is laid on the computational steps required to perform the statistical tests presented in the Manual, because of their usefullness in the preparation of related computer program algorithms used in statistical programming exercises of Part 3 of the Course on BASIC and FORTRAN programming.

The contents of this Manual can be taught in 15 to 30 hours depending on the level of statistical knowledge of the trainees.

BARC, August 1986

Contents

Contents page

Contents | page

Statistics deal with methods used in the collection, presentation, analysis and interpretation of data.

In a narrower sense the term statistics is used to denote the data themselves or numbers derived from the data as for example, averages. Thus we speak of crop statistics, rainfall statistics etc.

1. BASIC IDEAS ON DATA COLLECTION

The data is usually numerical data collected from a sample, that is a part of the population under study. Here the word population is not used in the common sense of "population of human beings". It is rather used to represent the totality of observations with which the user of statistical methods is concerned. A population must always be defined to make sure that it contains only elements of the same nature.

Example 1

a) *The set of meteorological stations in Bangladesh constitutes a population of weather stations.*

b) *The area of Soil Association shown on a map is a population of points.*

c) *The total number of fishes in a lake is also a population.*

Populations may be finite or infinite. The first and third populations of example 1 are finite populations, because the number of meteorological stations in a country or the number of fishes in a lake is a finite number. The second population of example 1 is an infinite population since there is an infinite number of points in a mapped Soil Association.

In many cases the agriculturist is dealing with either infinite populations or finite populations with many members. In such cases the population under study would be too large to be enumerated and/or measured, because the cost of these operations would be prohibitive.

In fact, it is not necessary to operate on the whole population. Statistics give us the means of selecting only a set of elements from the population, called a sample, study it and draw conclusions on the whole population. However it is essential that given procedures of selection of samples are followed, if the statistical conclusions are to be meaningful.

One important thing is that the sample be representative of the population. One way in which a representative sample may be obtained is by a process called random sampling. Random sampling means that each element or member of the population under study has an equal chance of being included in the sample. There are various techniques for obtaining a random sample. A simple one that has proved very useful in many situations involving small populations is to assign numbers to each member of the population, write these numbers on small pieces of paper, place them in a box, and then draw numbers from the box by mixing them thoroughly before each drawing. If the population is large a table of random numbers available in many statistical books can be used to assign numbers to elements of the population; it is also possible to do this on a computer or programmable pocket calculator by using a subroutine for producing random numbers. Such subroutines are readily available on most computers.

Another important point is that the enumeration and/or measurement of the elements of the selected random sample should be as accurate as possible. There are two types of errors that can be done in performing any enumeration or measurement operation

(a) *Random errors*

(b) *Systematic errors*

Random errors have a more or less irregular pattern of occurrence. For instance repeated pH measurements carried out on the same soil lab sample with the same instrument will always differ to some extent; this requires of course that the measuring instrument is precise enough. Then, some of the values will lie above the true pH value of the sample the other ones will lie under the true pH: we would expect about a half of the values to lie above and the other half to lie under the true value. Random errors occur in any measurements, they cannot be eliminated but careful measurement may help reduce their magnitude.

Systematic errors are <u>one-sided errors</u>. For example, when measuring length with a tape, if the tape is not kept tight, every measurement is too long resulting in a one-sided positive error.

If the pH measurements are performed by a soil lab technician with a tendency of reading values that are less than the ones actually shown on the instrument (scale), most of the measurements will be lower than the true values.

Errors of this kind are dangerous since they cause all observed values to be either above or under the expected results. Such measurements are said to be <u>biased</u>. Care should be taken to avoid this kind of errors as far as possible.

Now, provided that the prescribed procedure to select a random sample is followed and an accurate assessment of the elements of the sample is made, the sample data collected is used,

a) *to get estimates of the true values of certain population parameters.*

b) *to get estimates of the errors made in estimating those parameters through the sample.*

Let us consider the population of soil samples in a soil series at a particular location in Bangladesh. We might want to investigate one specific soil characteristic at that location, for example the pH of the soil; here the parameter we are interested in might be the <u>mean pH value of the soil series at that location</u>.

To estimate this parameter, we use the <u>mean pH of the sample of soil samples</u>. Such a pH average calculated from a sample is called a <u>statistic</u>. Any value calculated from a sample is a statistic, whereas a <u>parameter</u> is a population value. Statistics are used to estimate parameters that are mostly unknown. In order to avoid confusion latin letter are usually used to symbolize statistics and greek letters to symbolize parameters. Since many random samples are possible from the same population, we would expect a statistic to vary somewhat from sample to sample. A statistic is therefore a <u>random variable</u> like each observation in the random sample from which the statistics has been evaluated, but a parameter of a specific population is a fixed value.

2. <u>DATA PRESENTATION</u>

It is often helpful to summarize the raw sample data collected before processing it in order to detect certain characteristics of the sample. This can be done by setting up a frequency table of the data and/or present it in a graphical form, as an histogram.

Let us consider an example. Suppose that from a forest plantation under study a sample of 100 trees are selected and their height measurements recorded. By building height classes where every tree within a class is considered having the mean height of that class, we might have got the following <u>frequency table</u>:

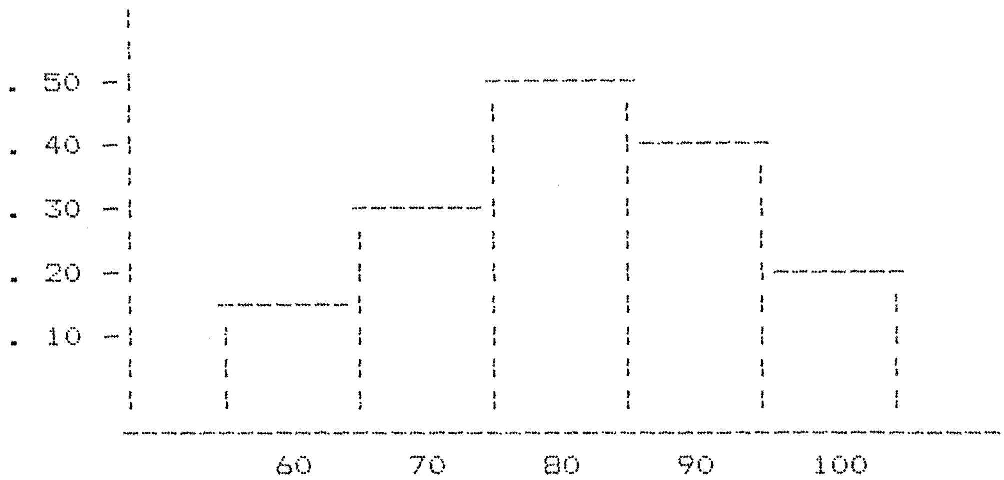
<u>Table 1</u>

height classes (dm)	height mean (dm)	absolute frequencies	relative frequencies
56 - 65	60	5	0.05
66 - 75	70	18	0.18
76 - 85	80	42	0.42
86 - 95	90	27	0.27
96 - 100	100	8	0.08
		100	1.00

The relative frequencies are obtained by dividing the absolute frequencies by 100 (= the number of sample trees).

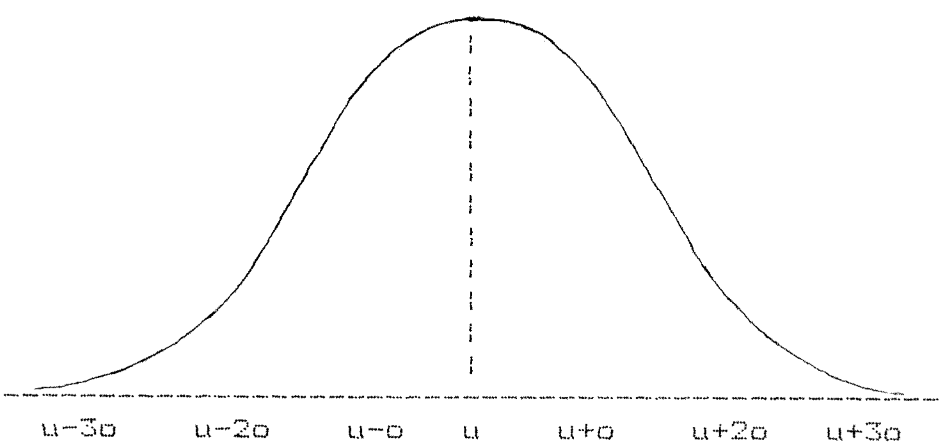
We can use the class middles (abcisses) and the relative frequencies (ordinates) to draw the following histogram (Fig.1).

Fig. 1



The above histogram is typical of many frequency distributions obtained from various types of measurements, like height and diameter measurements, crop yields etc., They can be considered as rough bell-shaped distributions. The most useful of these bell-shaped distributions is the so called <u>normal distribution</u>, the graph of which is given in Fig.2.

<u>Fig. 2</u>



The normal distribution belongs to a class of distributions called continuous probability distribution.

One other important distribution which belong to the class of so called discrete probability distribution is the binomial distribution. It is used to analyse phenomena that can be described in terms of numbers of times an event of interest will happen in a certain number of trials. For instance, if p is the probability that an event will happen in any single trial (called the probability of a success) and q, with q = 1 - p, is the probability that the event will fail to happen in any single trial (called the probability of failure) then the probability pattern of the event can be said to correspond to the binomial distribution. Assuming N trials some properties of the binomial distribution are as follows.

Mean u = Np

Variance o^2 = Npq

Standard deviation o = $\sqrt{Npq}$

Example 2

If a fair coin is tossed 196 times the mean (or expected) numbers of heads is u = Np = 196 * 1/2 = 98, and the standard deviation is $o = \sqrt{Npq} = \sqrt{196*1/2*1/2} = \sqrt{\frac{196}{4}} = \sqrt{49} = 7$

NOTE: *A FAIR COIN IS A COIN FOR WHICH THE PROBABILITY OF GETTING A HEAD OR A TAIL ON TOSSING IS EQUAL TO 1/2*

Two important parameters of the normal distribution are its mean denoted by the greek letter u and its standard deviation denoted by the greek letter o. Sample estimates of these parameters are respectively x and s. The following geometrical properties of the normal curve (see figure 2) is of fundamental importance for statistical inferences:

a) The area under the normal curve between $\mu - \sigma$ and $\mu + \sigma$ is about 68% of the total area.

b) The area under the normal curve between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 95% of the total area.

c) The area under the normal curve between $\mu - 3\sigma$ and $\mu + 3\sigma$ is about 99% of the total area.

3. DATA ANALYSIS AND INTERPRETATION

3.1 Some useful statistics

The normal distribution given in fig. 2 is completely determined by its mean μ, its standard deviation σ or variance σ^2. The mean is used for measuring the centre of the distribution and the standard deviation or variance is used for measuring the dispersion of the distribution. The population mean μ is estimated through the sample mean $\bar{x}$; the population variance σ^2 is estimated through the sample variance s^2 and the population standard deviation σ is estimated through the sample standard deviation s. Now we are going to see how to compute these three statistics from the sample values.

3.1.1 Sample Mean

The mean or arithmetic mean of a random sample of n observations $x_1, x_2, \ldots\ldots\ldots, x_n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 3.1

Find the mean of the random sample whose observations are

10, 15, 13, 8, 9

Solution

$$\bar{x} = \frac{10 + 15 + 13 + 8 + 9}{5} = 11$$

If the observations x_1, x_2, ..., x_k occur f_1, f_2, ..., f_k times respectively, which means that the observations can be arranged into k classes, the arithmetic mean is given by:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_kx_k}{f_1 + f_2 + \dots + f_k}$$

$$= \frac{\sum_{j=1}^{k} f_jx_j}{n}$$

where

$n = f_1 + f_2 + \dots + f_k$ is the total frequency or number of observations in the sample

Example 3.2

Let us consider the data of Table 1. To calculate the mean height of the sampled trees we can use directly the data in the second and third columns of the table as follows:

$$\bar{x} = \frac{(5*60) + (18*70) + (42*80) + (27*90) + (8*100)}{100}$$

$$= 81.5 \text{ dm} = 8.15 \text{ m}$$

3.1.2 Sample variance

The sample variance of a random sample of n observations x_1, ..., x_n is given by

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

Note that the sample variance is like the mean an average: It is the average of the squares of the deviations of the observations from their mean, but with a relatively small correction; the divisor is not n but n-1. The reason for using n-1 as a divisor rather than n is that the mean x is used in the formula as an estimator of the parameter u, and therefore one so called degree of freedom is lost in estimating the true population mean by x, as it is the case for any other parameter, so that after remain n-1 degrees of freedom associated with the variance s^2.

Example 3.3

Find the variance of the random sample whose observations are

5, 4, 9, 6, 5, 4, 7, 8

Solution

$$\bar{x} = \frac{5 + 4 + 9 + 6 + 5 + 4 + 7 + 8}{8} = 6$$

Thus

$$s^2 = \frac{\sum_{i=1}^{8} (x_i - 6)^2}{7}$$

$$= \frac{(5-6)^2 + (4-6)^2 + (9-6)^2 + (6-6^2 + \ldots + (8-6)^2}{7}$$

$$= \frac{24}{7}$$

If the mean $\bar{x}$ is a number that has been rounded off, using the variance formula in the above form may result in a large error in the value of the variance. To avoid this the following form of the variance formula can be used:

$$s^2 = \frac{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n\ (n-1)}$$

or

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

This formula is particularly suitable for calculating of the variance on computers, because it is more efficient than the first one in the sense that it is not necessary to evaluate the mean first and then the variance; the mean is rather a by-product of the variance calculation. To illustrate this let us treat example 3.3 again but using the second formula. The sums and sums of squares of the observations needed for the calculation are contained in the following table

Table 2

x_i	x^2
5	25
4	16
9	81
6	36
5	25
4	16
7	49
8	64

$$\sum_{i=1}^{n} x_i = 48 \qquad \sum_{i=1}^{n} x^2 = 312$$

Hence

$$s^2 = \frac{312 - \frac{(48)^2}{8}}{7} = \frac{312 - 288}{7} = \frac{24}{7}$$

Thus to implement this formula on the computer the observation values need to be read only once and the same storage cell is reused for each of the successive data items, whereas the implementation of the other formula would require either reading the observation values twice; once for computing the mean and the second time to compute the variance, or the observation values are read once but then kept all in storage for variance computation, which would mean that a substantive amount of memory may be needed to store the data in case of large samples.

3.1.3 Sample standard deviation

The sample standard deviation is the positive square root of the sample variance

$$s = \sqrt{s^2}$$

Thus the standard deviation of the sample of example 3.3 is

$$s = \sqrt{\frac{24}{7}}$$

3.2 Statistical Inferences

3.2.1 Standard Error, Confidence Intervals, Statistical Hypothesis and Tests of Significance

We have seen how to compute the location statistic mean and the dispersion statistics variance and standard deviation from a sample. Using these statistics we might wish to make various statements concerning the values of the corresponding population parameters. But generalization from a statistic to a parameter can be made with confidence, only if we understand the fluctuating behaviour of our variable statistic when computed from different random samples from the same population. One parameter we are very often called upon to make decision about is the population mean. The distribution of the sample mean that describes its fluctuating behaviour is used to make descisions about the population mean.

Here an important statistic is the standard deviation of the samples distribution of the mean that is called the standard error of the mean. The standard error of the mean depends on the size of the population, the size of the samples and the method of choosing the random samples. The standard error of the mean is given by

$$o_{\bar{x}} = \frac{o}{\sqrt{N}}$$

where o is the population standard deviation and N the population size. It is estimated by

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation and n the sample size.

The standard error is a measure of the error made in estimating the true population mean by the sample mean or a measure of the confidence limits of the sample mean. Since the true mean is usually unknown, direct calculation of this error is not possible, but confidence intervals can be built giving the range about the sample mean within which the population mean can be expected to lie with a certain probability. Such confidence intervals are commonly used in statistical tests.

To be able to make statistical decisions, it is useful to make assumptions or guesses about the populations under study, the parameters of which are unknown. Such assumptions are used to define statistical hypothesis. For instance, if we want to decide whether two populations have different means we formulate the hypothesis that there is no difference between the means. Such an hypothesis is often called null hypothesis and any hypothesis which differs from the null hypothesis is called an alternative hypothesis.

Now assuming that the null hypothesis is true, if we find that results observed in a random sample differ significantly from those expected under the null hypothesis on the basis of pure chance using sampling theory, we would conclude that the observed differences are significant and we would reject the null hypothesis. Procedures which enable us to decide whether to reject or accept a null hypothesis or to determine whether observed samples differ significantly from expected results (see Chi-square test below) are called tests of significance.

A statistical decision based on a test of significance that is itself based on probability theory can never have a 100% reliability. We might have made an error in rejecting the null hypothesis when in fact we should not have rejected it. However, we can specify the maximum probability with which we would be willing to risk such an error this probability or error is called the level of significance of the test and is denoted by α. In the following tests we will use a level of significance of 0.05 which is the most common one in agriculture statistics.

The meaning of 0.05 or 5% level of significance is that, in testing a null hypothesis against an alternative hypothesis, there are about 5% chances in 100 that we would reject the null hypothesis when it should not be rejected, or we are about 95% confident that we have made the right decision.

3.2.2 Standardized Variable and use in building Confidence intervals

Statistical methods are somewhat general methods that apply to a wide range of problems in various fields. For example, although medical data is in most cases different from agricultural data the same statistical procedures may be used for processing both medical and statistical data. This is because in statistics the original observations (raw data) are often transformed to new variables called standardized variables that are dimensionless quantities, independent of the units used for the original variables.

Such a standardized form of an observation from a sample is obtained by subtracting the mean of all sample observations from the observations and dividing the result by the standard deviation of the sample; it is denoted by z. Therefore

$$z = \frac{x - \bar{x}}{s}$$

If a random variable x is normally distributed, the corresponding standardized variable z will also have a normal distribution that is called the standard normal distribution because its mean is 0 and its standard deviation is 1.

The standard normal distribution is shown in figure 3 that is similar to figure 2. Correspondingly the standard normal curve has the following geometrical properties that are also of basic importance for statistical inferences:

a) *The area under the standard normal curve between -1 and +1 is about 68% of the total area.*

b) *The area under the standard normal curve between -2 and +2 is about 95% of the total area.*

c) *The area under the standard normal curve between -3 and +3 is about 99% of the total area.*

Fig. 3

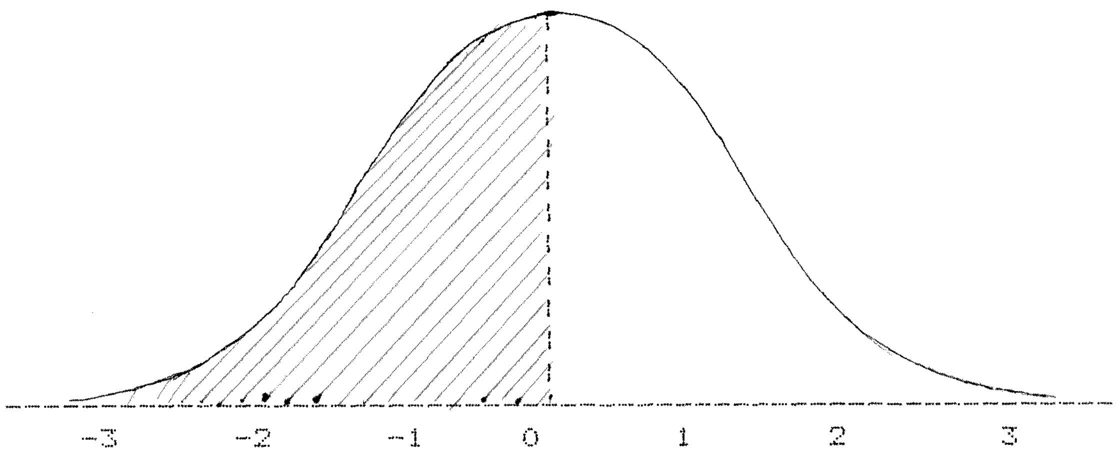


Table A in the Appendix contains values that represent areas under the curve of figure 3. The values of z in this table are given to two decimal places, with the second decimal place determining the column of the table to use. As an illustration, suppose we wish to find the area under the curve left to the mean 0 (hatched area).

In table A we read down the first column until the z value -0.0 or 0.0 is reached, then across to the entry in the column headed 0.00 to find 0.5000; this is the desired area.

The relationship between the standardized variable and the original variable x can also be expressed by

$$x = u + z$$

This relationship enables us to find the point z on the standard normal distribution that corresponds to any point x on the original normal distribution. The standardized variable is therefore of great importance in statistical tests.

A standardized variable called the <u>t statistic</u> is often used in establishing confidence intervals for the true means of populations on the basis of small ($n < 30$) as well as large samples ($n >= 30$). The t statistic is given by

$$t = \frac{\bar{x} - u}{s / \sqrt{n}}$$

or

$$t = \frac{\bar{x} - u}{s\bar{x}}$$

The variable t has a distribution called the <u>t distribution</u>, some commonly used values of which are contained in Table C in the Appendix.

For a particular random sample of size n the mean $\bar{x}$ and estimate of the standard error of the mean, $s_{\bar{x}}$ are computed and the (1-a) 100% confidence interval is given by

$$\boxed{\bar{x} - t\,\alpha/2 \,.\, \bar{s}_x < u < \bar{x} + t\,\alpha/2 \,.\, \bar{s}_x}$$

Where $t\,\alpha/2$ is the critical t value with n-1 degrees of freedom at a level of significance α. The critical t value is found in the table of the t distribution.

Example 3.4

Suppose that from a teak plantation located at an homogeneous site the following statistics are calculated from a random sample of 150 trees.

$\bar{x}$ = mean under bark volume per ha = 200 cubic meter

$s_{\bar{x}}$ = estimate of standard error of the mean = 9.5 cu.m/ha

Construct a confidence interval in which the true mean (mean volume under bark per ha) can be expected to lie with 95% probability or at level of significance = 0.05.

Solution

To construct the confidence interval the critical value $t_{\alpha/2}$ or $t_{0.025}$ is needed. It is found in the table of the t distribution as follows:

We read down the first column until the number of degrees of freedom is reached. Since n = 150 we reach inf. in the last row of column 1 that corresponds to any number of degrees of entry in the column headed 0.025 to find $t_{0.025}$ = 1.96. Thus the confidence interval is:

$$200 - (1.96*9.5) < u < 200 + (1.96*9.5)$$

or

$$200 - 18.62 < u < 200 + 18.62$$

or

$$181.38 < u < 218.62$$

Interpretation

We can be 95% confident that the mean volume under bark per ha in the teak plantation is in the range of 181.38 to 218.62 cu.m/ha.

Exercise 3.1

The following data represent rainfall measurements (in cm) at a certain location over a period of 20 successive years.

202, 205, 190, 198, 220, 192, 175, 208, 230, 218
172, 196, 215, 204, 183, 195, 217, 200, 219, 187

Assuming that annual rainfall is a normally distributed random variable with unknown mean and variance and successive years are independent from each other, construct a confidence interval in which the true mean (mean annual rainfall at that location) can be expected to lie with 95% probability.

3.3 NORMALITY TEST

One important requirement for the use of some common statistical tests, like t-test, is that the population under study is normally distributed. In many cases the user of statistical methods will not be wrong in assuming that the population concerned is normally distributed. However, it is advisable to check whether the assumption of normality holds to be sure that the results of statistical tests that require normality of the distribution of the population observations are reliable.

A simple test involving the standardized variable z can be used to get a good idea about the normality of a population. It may be carried out as follows:

Given a random sample $x_1, x_2, \ldots\ldots\ldots, x_n$ from the population, the test procedure is as follows:

1) Calculate the mean of the sample

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2) Calculate the sample standard deviation

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

3) Compute the values z_i for the standardized variable z corresponding to each observation

$x_i, \quad i = 1,2, \ldots\ldots, n$

$$z_i = \frac{x_i - \bar{x}}{s}$$

4) Evaluate the percentage of the z_i values that lie between the limits (-2, +2).

5) <u>The decision</u>

If at last 95% of the values for the standardized variable are between the limits (-2, +2) the population from which the sample has been selected can be considered as normally distributed, otherwise the population is considered to be not normally distributed.

Example 3.5

Given the random sample 2,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,9, 10. Test the hypothesis of normality of the population from which the sample has been selected.

1) The mean of the sample is

$$\bar{x} = 5.7$$

2) The standard deviation of the sample is

$$s = 2.105$$

3) The values of the standardized variable are:

$$z_1 = \frac{2-5.7}{2.105} = -1.76$$

$$z_2 = z_3 = \frac{3-5.7}{2.105} = -1.28$$

$$z_4 = z_5 = z_6 = \frac{4-5.7}{2.105} = -0.81$$

$$z_7 = z_8 = z_9 = z_{10} = \frac{5-5.7}{2.105} = -0.33$$

$$z_{11} = z_{12} = z_{13} = \frac{6-5.7}{2.105} = 0.14$$

$$z_{14} = z_{15} = z_{16} = \frac{7-5.7}{2.105} = 0.62$$

$$z_{17} = z_{18} = \frac{8-5.7}{2.105} = 1.09$$

$$z_{19} = \frac{9-5.7}{2.105} = 1.57$$

$$z_{20} = \frac{10-5.7}{2.105} = 2.04$$

4) 19 out of the 20 values for the standardized variable lie between the limits (-2, +2). Only z_{20} = 2.04 lies outside these limits. Therefore the percentage of the values z_i for the standardized variables z that lie between the limits (-2, +2) is

$$\left(\frac{19}{20} * 100\right)_{\%} = 95\%$$

5) <u>The decision</u>

Since 95% of the values for the standardized variables are between the limits (-2, +2), the population from which the sample has been collected can be considered as a normally distributed population.

In practice the user of statistical procedures needs not be so strict in applying the normality test to decide whether to apply confidently standard statistical tests to collected sample data to be analysed. This is due to the important fact that for samples of size n > 30, called large samples, the sampling distributions of many statistics, such as the mean, are approximately normal, the approximation becoming better with increasing sample size; this holds even though the original population under study may not be normally distributed. In such cases the user is better put on the safer side, by using samples as large as possible. But there are many instances where the user may not be able to come by with large samples (n > 30) to be analysed. Though the approximation for normal distribution of sample statistics is not good for small samples n < 30, there are modifications that can be made to make standard statistical tests valid on such data; they are used on a study of sampling distributions of statistics for small samples called small or exact sampling theory. Two important distributions derived from this theory are the chi-square distribution and the "Student's" t distribution (already mentioned).

Exercise 3.2

Given the random sample

20,11,16,8,9,33,14,17,12,16,23,19,12,18,
21,19,11,9,15,17,13,22,17,38,20,14,21,15,
16,12,25,17,20,15,23,24,14,19,13,16.

Test the hypothesis of normal distribution of the population from which the sample has been selected.

3.4 THE CHI-SQUARE TEST

The Chi-square test is often used to test whether observed frequencies of events in a sample are compatible with the expected frequencies.

The Chi-square test supplies a measure of the discrepancy existing between observed and expected frequencies. The statistics X^2 used in the chi-square test is given by

$$X^2 = \frac{(n_1 - k_1)^2}{k_1} + \frac{(n_2 - k_2)^2}{k_2} + \ldots\ldots + \frac{(n_m - k_m)^2}{k_m}$$

$$= \sum_{i=1}^{m} \frac{(n_i - k_i)^2}{k_i}$$

where

m = the number of classes or events

n_i = observed frequency in the i_{th} class

k_i = expected frequency in the $_{ith}$ class

The total frequency N is

$$N = \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} k_i$$

The number of the degrees of freedom is

$$n.d.f = m-l-1$$

where l is the number of parameters that might have to be estimated in computing the expected frequencies.

If $X^2 = 0$, observed and expected frequencies agree exactly, while if $X^2 > 0$ they do not agree exactly. (Note that X^2 cannot be < 0). The larger the value of X^2 the greater the discrepancy between observed and expected frequencies.

3.4.1 The Chi-square test procedure

1) Compute the observed frequencies n_i of the events e_i in the sample, for i = 1,2, ... , m

2) Calculate, if necessary, the expected frequencies by multiplying the probabilities p_i of the events by the total frequency N

$$k_i = N*p_i \quad \text{for} \quad i = 1,2, \ldots\ldots, m$$

3) Evaluate

$$X^2 = \sum_{i=1}^{m} \frac{(n_i - k_i)^2}{k_i}$$

4) The significance test

Compare the computed value for X^2 with the critical value at the selected significant level α. The critical value is contained in the table of the distribution and is located at the junction of the number equal to the number of degrees of freedom of the Chi-square statistics and the column corresponding to α, (see Table B in the Appendix).

5) The decision

If the computed value of X^2 is greater than the critical value from the table conclude that observed frequencies differ significantly from expected frequencies and reject the hypothesis that observed frequencies agree with expected frequencies; otherwise accept the hypothesis.

Example 3.6

We know that in tossing a fair dice the probability of getting any one of the possible outcomes or events 1,2,3,4,5, 6 is p = 1/6. Now suppose that we have got a dice and want to test whether the dice is fair. We can do this by using the Chi-square test in the following manner:

(1) We toss the coin a certain number of times, let us say 180 times, and count the number of times each of the events 1,2,3,4,5,6 occurs: our counts are the observed frequencies n_i, the sum of which must be 180.

(2) the expected frequencies are easily obtained: we expect each of the events 30 times from tossing a fair dice 180 times

Let us summarize the result of 1) and 2) in the following table

Table 3

Face	1	2	3	4	5	6
Observed frequencies	31	29	25	35	32	28
Expected frequencies	30	30	30	30	30	30

(3) The value of the Chi-square statistic is:

$$X^2 = \frac{(n_1 - k_1)^2}{k_1} + \frac{(n_2 - k_2)^2}{k_2} + \frac{(n_3 - k_3)^2}{k_3}$$

$$+ \frac{(n_4 - k_4)^2}{k_4} + \frac{(n_5 - k_5)^2}{k_5} + \frac{(n_6 - k_6)^2}{k_6}$$

$$X^2 = \frac{(31-30)^2}{30} + \frac{(29-30)^2}{30} + \frac{(25-30)^2}{30}$$

$$+ \frac{(35-30)^2}{30} + \frac{(32-30)^2}{30} + \frac{(28-30)^2}{30}$$

$$= \frac{1}{30} + \frac{1}{30} + \frac{25}{30} + \frac{25}{30} + \frac{4}{30} + \frac{4}{30}$$

$$= 2$$

(4) No parameter has been estimated, thus l=0; since the number of classes or events is k = 6, the number of degrees of freedom is

$$\begin{aligned} n.d.f &= m-1-1 \\ &= 6-0-1 \\ &= 5 \end{aligned}$$

The critical value $X^2_{0.95}$ (significance level $\alpha = 0.05$) for 5 degrees of freedom is 11.07 (See Table B, in the Appendix). The computed value for Chi-square (2) is smaller than the corresponding table (critical) value for Chi-square that is 11.07.

(5) Since the computed value for Chi-square is less than the table value for Chi-square, we accept the hypothesis that the dice is fair.

NOTE

MAKE SURE THAT THE EXPECTED FREQUENCIES ARE AT LEAST EQUAL TO 5 ($k_i = 5$) IN ORDER TO GET RELIABLE RESULT FROM THE CHI-SQUARE TEST.

We should look with suspicion upon circumstances where X^2 is too close to zero since it is rare that observed frequencies agree to well with expected frequencies. To examine such situations we can determine whether the computed value of X^2 is less than $X_{.95}$ or $X_{.99}$, in which cases we would decide that the agreement is too good at the 0.05 or 0.01 levels of significance respectively.

Exercise 3.3

50 agricultural workers have undergone a training in using a new method to perform a particular rice harvesting operation. This new method is supposed to bring an improvement in terms of the time needed to execute the operation. A current method is considered as the standard method and time data recorded before training for the standard method has been used to compute the expected frequencies. The observed frequencies are obtained from data recorded by a time keeper at the end of the training. Table 4 below shows the observed and expected frequencies. Test the hypothesis that the observed frequencies agree with the expected frequencies, in other words that there is no significant pattern for the 2 methods. A significance level $\alpha = 0.05$ is to be used for the test.

Table 4

time in min.	up to 10	10 to 12	12 to 14	14 to 16	16 to 18	18 +
observed frequencies	8	10	19	9	3	1
expected frequencies	5	5	15	10	9	6

3.5 THE t-TEST

The t-test is appropriate for testing statistics of small as well as large samples from normally distributed populations.

The t-test is often used to test the hypothesis that 2 normally distributed populations whose standard deviations are equal ($\sigma_1 = \sigma_2$) also have equal means ($u_1 = u_2$). There are two tests depending on the null hypothesis:

a) *the one-sided t-test*

b) *the two-sided t-test*

<u>In case a)</u> the null hypothesis that the means of the two populations are equal ($u_1 = u_2$) is tested against the alternative hypothesis that one of the two means is greater than the other, for instance ($u_1 > u_2$).

<u>In case b)</u> the null hypothesis of equality of the two populations mean ($u_1 = u_2$) is tested against the alternative that the means are not equal, i.e. $u_1 \# u_2$.

3.5.1 The t-Test Procedure

Given two independent random samples from the two populations. The observations in the first sample are denoted by

$$x_1, x_2, \ldots\ldots\ldots, x_{n1}$$

The observations in the second sample are denoted by

$$y_1, y_2, \ldots\ldots\ldots, y_{n2}$$

The t-test is performed as follows:

(1) Compute the means

$$\bar{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$$

$$\bar{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$$

and the variances

$$s^2_1 = \frac{1}{n-1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

$$s^2_2 = \frac{1}{n^2-1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$$

(2) Compute

$$t = \sqrt{\frac{n_1 n_2 \quad (n_1+n_2-2)}{n_1 + n_2}} * \frac{\bar{x} - \bar{y}}{\sqrt{(n_1-1)s^2_1 + (n_2-1)s^2_2}}$$

if the sizes of the samples are not equal $(n_1 \neq n_2$

or

$$t = \sqrt{n} * \frac{\bar{x} - \bar{y}}{\sqrt{s^2_1 + s^2_2}}$$

if the samples are of the same size $(n_1 = n_2)$

3) <u>The significance test</u>

- <u>in case a)</u>

determine the t value with (n_1+n_2-2) degrees of freedom at the critical value α in the table of the t-distribution (See Table C, in the Appendix).

- in case b)

  determine the t value with (n_1+n_2-2) degrees of freedom at the critical value $\alpha_{/2}$ in the table of the t-distribution.

4) The decision

- in case a)

  if the calculated t value is less than or equal to the t value from the table, the null hypothesis that the two populations have equal means is accepted otherwise the null hypothesis is rejected.

- in case b)

  if the absolute value of the calculated t value is less than or equal to the t value from the table, the null hypothesis of equality of the two populations means is accepted, otherwise the null hypothesis is rejected.

CAUTION !

THE T-TEST SHOULD NOT BE USED IF

1) the obaervations in the two samples are not independent from each other.

2) the two populations are not normally distributed.

3) the variances o^2_I and o^2_I of the two populations are not equal.

Example 3.7

Two fertilizers A and B are being tested for their effect on the height growth of nursery plants. Each of the fertilizers has been applied to 16 different experimental plots. For fertilizer A the calculated mean height is 20.5 cm and the variance $s^2_A = 16.8$. For fertilizer B the calculated mean height is 18.0 cm and the variance $s^2_B = 8.4$.

To test is the hypothesis $u_A = u_B$ against the alternative $u_A > u_B$, using a significance level of $\alpha = 0.05$.

Solution

1) The means and variance are given

2) Since the sizes of the samples are equal the t value is calculated as

$$t = \sqrt{16} * \frac{20.5 - 18}{\sqrt{16.8+8.4}} = \sqrt{16} * \frac{2.5}{\sqrt{25.2}}$$

$= 2$

3) The significance test

The t value with (n_1+n_2-2 = 16+16-2) or 30 degrees of freedom using the significance level $\alpha = 0.05$

4) The decision

Since the calculated t value is greater than the t value from the Table (2 > 1.7) the null hypothesis of equality of the means of the two populations is rejected. The interpretation of this result is:

the observed difference in the height growth is not due to random factors but is likely due to the different effect of fertilizer A and B on the height growth of the plants. Fertilizer A has a significantly greater effect on the height growth of the nursery plants than fertilizer B.

Exercise 3.4

Two varieties of rice are being investigated for their yield potentials. Several experimental plots of the two rice varieties have been established under the same environmental conditions. The recorded yield values are as follows:

Variety 1

Plot No.	Yield in kg
1	12.6
2	10.1
3	11.5
4	9.6
5	13.7
6	7.8
7	8.5
8	9.6
9	13.6
10	12.5
11	10.4
12	15.3
13	16.5
14	12.5
15	15.3
16	13.2
17	7.4
18	9.5
19	13.8
20	11.3

Variety 2

Plot No.	Yield in kg
1	13.4
2	15.3
3	12.3
4	16.4
5	16.8
6	11.9
7	12.3
8	15.8
9	8.1
10	16.1
11	15.4
12	9.6
13	6.9
14	18.3
15	15.4
16	7.4
17	17.9
18	16.2
19	12.8
20	15.1
21	9.9
22	14.3

To be tested is the null hypothesis that the mean yields of the 2 varieties are equal ($u_1 = u_2$) against the alternative hypothesis that the mean yield of variety 2 is greater than the mean yield of variety 1 ($u_2 > u_1$), using a significance level $\alpha = 0.05$.

3.6 SIMPLE LINEAR REGRESSION

This statistical procedure is used for estimating or predicting the value of a dependent random variable y on the basis of a known measurement on an independent controlled variable x. Here the pair of observations (y,x) represents the results of any member of the population. The relationship between y and x is assumed to be linear and can therefore be represented by the equation of a straight line.

$$u_y = \alpha + \beta x$$

The parameter u_y is called the regression line
" " α " " " intercept
" " β " " " regression coefficient

A random sample of size n from the population might be represented by pairs of values (y_i, x_i), for $i=1,2,\ldots,n$. The problem is to use these sample values to estimate the regression line. This is done by estimating the 2 parameters α and β. If the estimate of α is denoted by a and the estimate of β is denoted b the parameter u_y can be estimated by y_x from the sample regression line, that is;

$$\bar{y}_x = a + b_x$$

The formula needed for computing a and b are

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \bar{y} - b\bar{x}$$

Where $\bar{y}$ = mean of the y_i 's

$\bar{x}$ = mean of the x_i 's

3.6.1 The computational procedure

> Given a random sample (y_i, x_i) , $i = 1,2, \ldots\ldots, n$ from a population. To fit a regression line using y as dependent variable and x as independent variable the following procedure can be used:
>
> 1) arrange the raw scores y_i, x_i as well as the sums needed to compute b in a table.
>
> 2) compute the means $\bar{y}$ and $\bar{x}$
>
> 3) compute the regression coefficient b
>
> 4) compute the intercept a
>
> 5) write the equation of the regression line
>
> 6) plot the raw scores to give a scatter diagram
>
> 7) draw the regression line into the scatter diagram

Example 3.8

Given the following 8 pairs of numbers

y	1	2	4	4	5	7	8	9
x	1	3	4	6	8	9	11	14

Fit a regression line to this data, using y as dependent variable and x as independent variable.

Solution

1) the raw scores and sums are arranged in the following table

Table 5

y	x	x^2	xy
1	1	1	1
2	3	9	6
4	4	16	16
4	6	36	24
5	8	64	40
7	9	81	63
8	11	121	88
9	14	196	126
$\sum y = 40$	$\sum x = 56$	$\sum x^2 = 524$	$\sum xy = 364$

2) The means are

$$\bar{y} = \frac{40}{8} = 5$$

$$\bar{x} = \frac{56}{8} = 7$$

3) The regression coefficient b is

$$b = \frac{(8)\ (364)\ -\ (56)\ (40)}{(8)\ (524)\ -\ (56)^2} = 0.636$$

4) The intercept a is

$$a = 5 - (7)\ (.636) = 0.548$$

5) The regression line is

$$\bar{y}_x = 0.548 + 0.636\ x$$

6) 7)

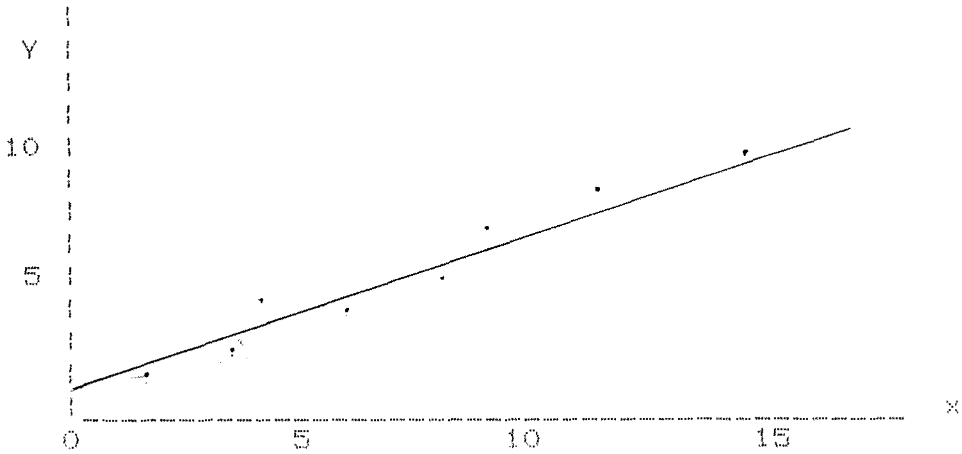


Fig. 4 *Scatter diagram and regression line*

To draw the regression line one needs substitute only two of the given values of x (preferably the smallest and biggest values of x) into the equation, for instance $x_1 = 1$ and $x_0 = 14$ to obtain the ordinates $y_1 = 1.18$ and $y_{14} = 9.45$. Connecting the two ordinates with a straight line that is extended until it touches the y axis gives the regression line. The value x=0 given $y_0 = .548$ (= the intercept a) can be also be used as the first of the two values of x.

Prediction

A regression $\bar{y}_x = a + bx$ may be used to predict values of the parameter u_y for values of x that are not necessarily some of the prechosen values used for fitting the regression line. However, the regression equation should be used for prediction only by substituting x values in the range of the biggest and smallest of the x values that were involved in the fitting of the regression line.

Exercise 3.5

The following data represent measurements on the weight in kg of the body and hindleg of 15 killed elephants

Body (y)	Hindleg (x)
340	20.4
837	48.1
347	18.6
604	47.2
527	25.4
695	36.3
721	43.5
281	11.8
947	47.2
2653	145.1
2394	127.0
1270	67.1
200	14.7
1304	75.0
1928	108.0

1) *Fit a regression line to this data, using y as dependent variable and x as independent variable.*

2) *Use the fitted regression equation to predict the body weight of an elephant of which the weight of the hindleg is 92.5 kg.*

3.7 SIMPLE CORRELATION

A simple correlation problem is similar to a simple linear regression problem in that they both deal with the relationship between two variables, let us say y and x. However they differ in that

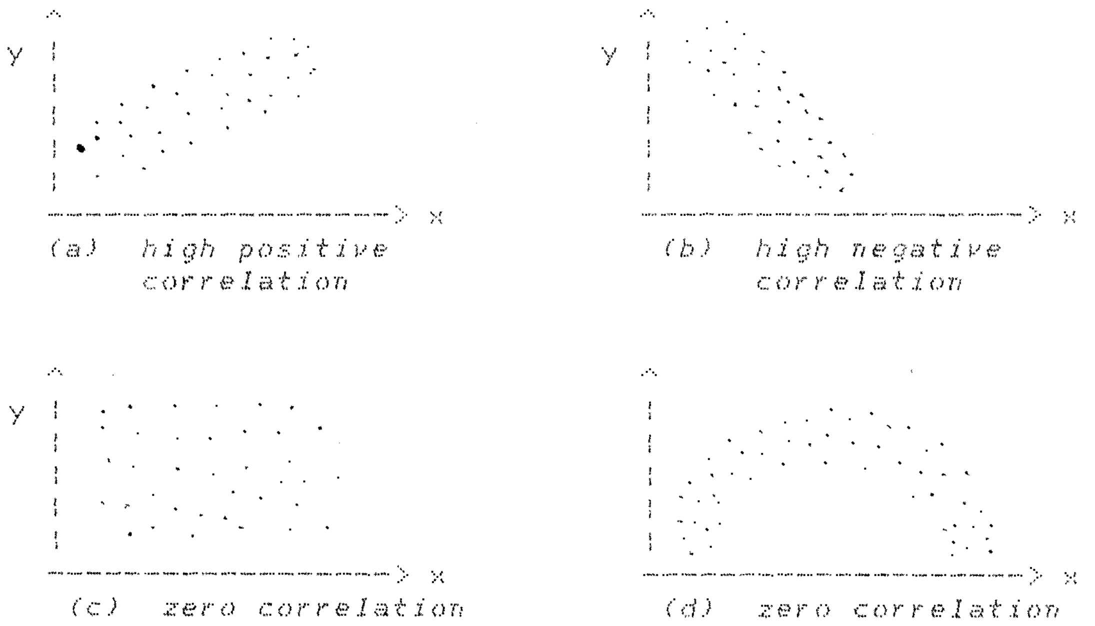
1) *a simple correlation problem is concerned with a measure of the relationship between the two variables y and x, whereas a simple regression problem uses the relationship to predict one of the two variables that is called the dependent variable from a knowledge of the other, the independent variable*

2) *the two variables are random variables in a simple correlation problem, whereas the values of the independent variable are fixed in a simple regression study.*

In a simple correlation problem the relationship between the two random variables y and x is measured by the so called linear correlation coefficient that is denoted by σ

Now, to estimate a linear correlation coefficient a random sample of n pairs of measurements (y_i, x_i) is selected from the population under study. Before correlation computations are performed certain conclusions can be drawn by constructing a scatter diagram for the (y_i, x_i) values as shown in fig. 5

Fig. 5 Scatter diagrams showing various degrees of correlation



Scatter diagram (a)

The points follow closely a straight line with positive slope: a high positive correlation exists between the two variables y and x.

Scatter diagram (b)

The points follow closely a straight line with negative slope: a high negative correlation exists between the two variables y and x.

Scatter diagram (c)

The points follow a random pattern: the correlation is equal to or near 0, which means that there is no relationship between y and x.

Scatter diagram (d)

The points follow the pattern of a strictly quadratic relationship: the correlation is equal to or near 0, that indicates a lack of linearity in the relationship, but not a lack of association of the variables y and x. This is because the correlation coefficient between two variables is only a measure of their linear relationship

The estimate of the linear correlation coefficient or sample correlation coefficient, denoted by r, is given by;

$$r = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i) - (\sum_{i=1}^{n} y_i)}{\sqrt{(n \sum_{i=1}^{n} x^2_i - (\sum_{i=1}^{n} x_i)^2) \quad (n \sum_{i=1}^{n} y^2_i - (\sum_{i=1}^{n} y_i)^2)}}$$

r can take on values from -1 to +1. The closer the value of r to -1 or +1 the closer the relationship between the variables y and x.

The square of r (r^2) is also a useful statistic called coefficient of determination. Multiplying r^2 by 100 gives the percentage of the variation in the values of the variable y that may be accounted for by the linear relationship with the variable x. For instance if $r = 0.8$, $r^2 = 0.64$ and $r^2 * 100 = 64$; thus a correlation of 0.8 means that 64% of the variation of the random variable y is accounted for by differences in the variable x.

3.7.1 The computational procedure

1) *Construct a scatter diagram for the (y_i, x_i) values from the sample*

2) *Arrange the values y_i, x_i as well as the sums needed to compute r in a table*

3) *Compute r, r^2, $(100.r^2)\%$*

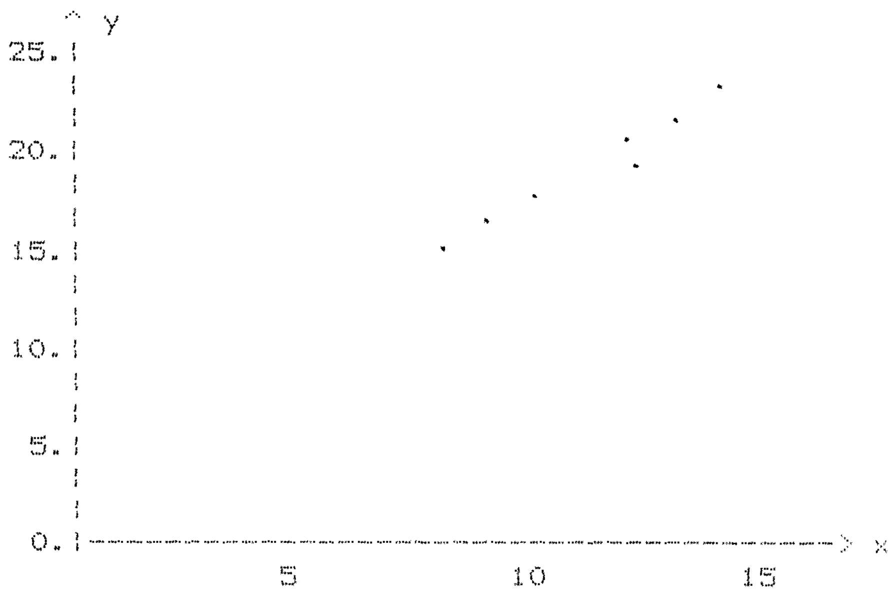
Example 3.9

Compute and interpret the correlation coefficient for the following data

y (weight)	17	18	20	16	22	19	15
x (height)	10	12	13	9	14	12	8

Solution

Fig. 6. Scatter Diagram



y	x	y^2	x^2	xy
17	10	289	100	170
18	12	324	144	216
20	13	400	169	260
16	9	256	81	144
22	14	484	196	304
19	12	361	144	228
15	8	225	64	120
$\sum_{i=1}^{7} y_i = 127$	$\sum_{i=1}^{7} x_i = 78$	$\sum_{i=1}^{7} y^2_i = 2339$	$\sum_{i=1}^{7} x^2_i = 898$	$\sum_{i=1}^{7} x_i y_i = 1442$
$(\sum_{i=1}^{7} y_i)^2 = 16129$	$(\sum_{i=1}^{7} x_i)^2 = 6084$			

Table 6 Table of values needed for the computation of the correlation coefficient r.

3) The correlation coefficient is:

$$r = \frac{(7)(1442) - (78)(127)}{\sqrt{((7)(898) - (6084))((7)(2339) - (16129))}}$$

$$= 0.85$$

$$r^2 = (0.85)^2$$

$$= 0.72$$

$$(100*r^2)\% = 72\%$$

Interpretation of the result

A correlation coefficient of 0.85 indicates a good linear relationship between y and x. Since $(100*r^2)\% = 72\%$, we can say that 72% of the variation in the values of y is accounted for by a linear relationship with x.

Exercise 3.6

Compute and interpret the correlation coefficient and coefficient of determination for the following 10 pairs of data that are height measurements on sample trees from a plantation. The height measurements had been made on the same trees at the age of 3 and 6.

Tree No.	height at 3 in m x	height at 6 in m y
1	4.5	12.8
2	4.6	12.6
3	4.3	12.1
4	4.3	12.0
5	4.2	11.5
6	4.1	11.6
7	3.9	11.0
8	3.8	10.8
9	3.7	10.9
10	3.5	10.0

Exercise 3.7

Case Study

The yield potential of a variety of potato at two homogeneous sites is being investigated. Of particular interest is the influence of the sites on the yield of the variety.

Investigation Procedure

The study includes the following points:

1) Sampling technique

   a) random sampling procedure

   b) size of samples

2) Data collection (see Sample Data below)

3) Data presentation

   a) frequency tables

   b) histograms

4) Data analysis

   a) Basic Statistics: mean, variance, standard deviation

   b) Normality test

   c) Test for difference between the mean of the two populations

5) Interpretation of the result

Sample data

The following yield data is to be used in the analysis:

Yield Measurements of Sample Plots in (kg)

Plot No.	Yield	Plot No.	Yield	Plot No.	Yield
1	20.7	26	26.4	51	12.1
2	18.2	27	7.5	52	13.0
3	19.8	28	15.5	53	6.1
4	20.5	29	24.5	54	22.2
5	15.0	30	19.6	55	17.7
6	20.5	31	19.1	56	18.9
7	21.9	32	19.9	57	19.1
8	18.9	33	21.0	58	10.5
9	19.4	34	17.4	59	31.1
10	22.3	35	10.5	60	17.5
11	10.3	36	22.1	61	21.2
12	10.8	37	22.0	62	23.8
13	18.4	38	12.8	63	11.5
14	20.0	39	17.4	64	20.0
15	16.7	40	18.3	65	14.3
16	16.9	41	22.3	66	19.4
17	13.4	42	14.9	67	19.6
18	22.5	43	23.8	68	11.6
19	18.3	44	11.7	69	16.9
20	14.4	45	22.3	70	19.4
21	19.5	46	13.5	71	15.0
22	17.5	47	13.4	72	20.7
23	13.5	48	28.5	73	24.8
24	19.1	49	24.5	74	19.0
25	20.9	50	10.7	75	22.7
				76	23.8
				77	27.3
				78	28.5
				79	27.1
				80	18.8
				81	30.2
				82	16.5
				83	24.7
				84	12.0
				85	29.6
				86	19.5
				87	25.8
				88	25.9
				89	13.4
				90	18.1

Yield Measurements of Sample Plots in (kg)

Plot No.	Yield	Plot No.	Yield	Plot No.	Yield
1	14.0	26	10.3	51	15.9
2	17.2	27	19.9	52	19.4
3	17.5	28	16.7	53	17.2
4	13.5	29	14.8	54	10.3
5	14.9	30	20.8	55	16.5
6	21.4	31	23.5	56	26.8
7	20.1	32	7.7	57	27.0
8	16.0	33	16.5	58	28.3
9	18.7	34	18.8	59	24.8
10	13.3	35	24.8	60	15.5
11	11.1	36	21.9	61	19.9
12	11.8	37	21.0	62	19.5
13	18.0	38	20.8	63	14.5
14	14.0	39	14.8	64	20.7
15	16.2	40	17.0	65	25.5
16	18.5	41	15.1	66	11.5
17	10.8	42	26.2	67	11.9
18	10.3	43	28.0	68	16.5
19	9.0	44	21.3	69	14.5
20	11.9	45	19.1	70	16.0
21	19.3	46	27.2	71	12.3
22	26.8	47	17.6	72	12.9
23	14.7	48	26.7		
24	12.7	49	12.2		
25	17.7	50	15.5		

3.8 ANALYSIS OF VARIANCE

To test whether the means of two populations are significantly different we use the t-test (see page 25). When more than two population means are to be compared simultaneously for equality a technique, called the analysis of variance, is used.

The analysis of variance is a method for breaking down the total variation of the collected data into meaningful components that measure different sources of variation. The sources of variation are determined by the criteria used to classify the observations, the possible interrelationships between these criteria, the experimental or sampling error in the data.

The classification of observations on the basis of a single criterion or factor such as variety, is called a one-way classification. The classification of observations on the basis of two criteria, such as variety and site is called a two-way classification. If the observations are classified according to three criteria, such as variety, site and fertilizer or spacing we have a so-called three-way classification. In the following the fixed effects or Model I analysis of variance procedure for each of these three classifications will be considered, because Model I is the simplest and the most useful of the three Models, that are usually designated Models I, II (random effects) and III (mixed = fixed + random effects). Fixed effects means that we are studying only some particular levels of the factor(s) of interest.

3.8.1 One-Way Classification

The Problem

Independent random samples of size n_i are selected from each of (a) levels of a factor or populations or treatments.

The a populations are assumed to be normally distributed with means $u_1, u_2, u_3, \ldots\ldots, u_a$ and have equal variance σ^2. We wish to test the hypothesis that all the means are equal against the alternative that at least two of the means are not equal. This is stated as follows:

$$H_0 : u_1 = u_2 = \ldots\ldots = u_a$$

(null hypothesis)

H_1 : At least two of the means are not equal

(alternative hypothesis)

3.8.1.1 <u>The Mathematical Model</u>

Given the following observations y_{ij}, where i=1,.....a indicates the treatments or levels of a factor and j=1,....., n_i denotes the observations for any particular treatment i. Each observation may be written in the form

$$y_{ij} = u + \alpha_i + \varepsilon_{ij}$$

subject to the restriction:

$$n_1\alpha_1 + n_2\alpha_2 + \ldots\ldots\ldots + n_a\alpha_a = 0 ,$$

in order to obtain unique least - squares estimators,

where u = total mean or mean of all treatment means

u_i with

$u_i = u+\alpha_i$

α_i = effect of the i^{th} population

ε_{ij} = deviation of the j^{th} observation of the i^{th} treatment from the corresponding treatment mean u_i

Another way of writing the null hypothesis is

$$H_0 : \alpha_1 = \alpha_2 = \ldots\ldots = \alpha_a = 0$$

and

$$H_1 : \alpha_i \# 0 \text{ for any } i, (i=1,\ldots\ldots,a)$$

The above model implies that each of the <u>a</u> population means, which we want to compare, is arbitrarily divided into two parts. The first part is the total mean and the second part is the difference between the mean of each population and the total mean.

Here again we emphasize the assumptions implied by the model:

a) *The* <u>*a*</u> *populations are normally distributed*

b) *The variance of the* <u>*a*</u> *populations are equal*

c) *The observations are independent*

3.8.1.2 <u>The Computational Procedure</u>

We use the collected data to estimate the values of u, α and ε.

The following table may be used to simplify the computations:

<u>Table 7</u>

	Sample		
Treatment 1	y_{11} y_{12} .... y_{1n_1}	$y_{1.}$	$\bar{y}_{1.}$
Treatment 2	y_{21} y_{22} .... y_{2n_2}	$y_{2.}$	$\bar{y}_{2.}$
.	. . .	.	.
.	. . .	.	.
.	. . .	.	.
.	. . .	.	.
.	. . .	.	.
Treatment a	y_{a1} y_{a2} .... y_{an_a}	$y_{a.}$	$\bar{y}_{a.}$
		$y_{..}$	$\bar{y}_{..}$

where

y_{ij}, (i=1, ... a; j=1, ..., n_i) is the j^{th} observation from the i^{th} treatment

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij},$$ Thats is the sum of observations of treatment i

$\bar{y}_{i.} = \dfrac{y_{i.}}{n_i}$, that is the mean of treatment i

$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}$, that is the sum of all observations

$N = \sum_{i=1}^{a} n_i$, that is the total number of observations

$\bar{y}_{..} = \dfrac{y_{..}}{N}$, that is the total mean. (For explanation of symbols, see Summation and dot Notation in Manual 1 on Basic Computer Mathematics.

Unbiased estimates of the parameters u, α and ε are obtained by using the appropriate entries in the above table, as follows:

u	is estimated by			$\bar{y}_{..}$
α_i	"	"	"	$\bar{y}_{i.} - \bar{y}_{..}$
ε_{ij}	"	"	"	$y_{ij} - \bar{y}_{i.}$

By replacing the unknown parameters by their estimates in the model

$$y_{ij} = u + {}_{i} + {}_{ij} ,$$ we obtain

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

or

$$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

which means that we transform our model to consider the deviation of each observation from the total mean, and divide that deviation into two parts. The first part represents the deviation of the mean of treatment i from the total mean. The second part represents the deviation of that particular observation in treatment i from the mean of treatment i. Taking the squares of both sides of the last identity, we have:

$$(y_{ij} - \bar{y}_{..})^2 = ((\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}))^2$$

$$+ (\bar{y}_{i.} - \bar{y}_{..})^2 + (y_{ij} - \bar{y}_{i.})^2$$

$$+ 2(\bar{y}_{i.} - \bar{y}_{..}) (y_{ij} - \bar{y}_{i.})$$

If we sum over all N observations we have, by using summation notation:

$$\sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$+ \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - y_{i.})$$

since $\sum_{i=1}^{a} \sum_{j=1}^{n_i} (2(\bar{y}_{i.} - \bar{y}_{..}) (y_{ij} - \bar{y}_{i.})) = 0$

Thus the sum of the squared deviations of the observations from the total mean, denoted by SST, equals the sum of squared deviations of the treatment means from the total mean, denoted by SSA, plus the sum of squared deviations of the observations from the treatment means, denoted by SSE. Therefore

$$SST = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

$$SSA = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

In practice one usually evaluates SST and SSA using the following formula that are more suitable for computation:

$$SST = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij} - \frac{y^2_{..}}{N}$$ = total sum of squares, with N-1 degrees of freedom

$$SSA = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y^2_{..}}{N}$$ = sum of squares for treatments with (a-1) degrees of freedom

Then SSE is evaluated as :

SSE = SST - SSA = error sum of squares with (N-a) degrees of freedom

The sum of squares formulas apply to both the case where the sample sizes are equal, that is $n_i = n$ for all i, (i=1,.. ...,a) and the case where the sample sizes are not equal, that is some of the n_i's or all the n_i's are different.

The second case of unequal sample sizes is quite common in agriculture and forestry due to frequent loss of observations in field experiments.

The variance of the grouped data s^2 is obtained by dividing the total sum of squares by the corresponding number of degrees of freedom as follows:

$$s^2 = \frac{SST}{N-1}$$

Other estimates of the variance are the mean square for treatment MST and the error mean square MSE

$$MSA = \frac{SSA}{(a-1)}$$

$$MSE = \frac{SSE}{(N-1)}$$

The value F necessary for testing the hypothesis is then evaluated as:

$$F = \frac{MSA}{MSE}$$

The above results can be summarized in the following table, called analysis of variance table

Table 8

Analysis of Variance Table

One-Way Classification

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed F
Treatment	SSA	a-1	MSA	F
Error	SSE	N-a	MSE	
Total	SST	N-1		

3.8.1.3 The Test

The random variable F has a so called F-distribution with $n_1 = (a-1)$ and $n_2 = (N-a)$ degrees of freedom. Critical values of the F distribution at the levels of significance $\alpha = 0.05$ are contained in Table D in the Appendix. To test the significance of the evaluated F value of the analysis of variance table at the level $\alpha = 0.05$, compare this value to the corresponding value of the Table D that is obtained by using the number of degrees of freedom for MSA, n_1, and the number of degrees of freedom for MSE, n_2, as column and row entries respectively. The table value is usually denoted by $F_\alpha(n_1,n_2)$.

If the evaluated F value is greater than the critical F-value the null hypothesis H_0 of equality of the populations or treatments means is to be rejected, otherwise the null hypothesis is accepted.

Example 3.10 Model I One-way Classification
Equal Sample Size

The data in Table 9 represent 5 random samples, each of size 5, from independent normal distributions with means u_1, u_2, u_3, u_4 and u_5; the common variance is o^2.

To test is the hypothesis:

$$H_0 : u_1 = u_2 = u_3 = u_4 = u_5$$

against the alternative:

H_1 : at least two of the means are not equal,
at the level of significance $\alpha = 0.05$.

Table 9

Sample Data, Sums and Means

	SAMPLE					TOTAL	MEAN
Treatment 1	4	5	7	3	2	21	4.2
Treatment 2	5	6	7	5	4	27	5.4
Treatment 3	2	6	3	4	2	17	3.4
Treatment 4	1	6	5	3	4	19	3.8
Treatment 5	3	2	4	5	3	17	3.4
						101	4.04

a) Evaluation of the Sums of Squares

$$SST = (4^2+5^2+7^2+3^2+2^2+5^2+6^2+7^2+5^2+4^2+2^2+6^2+3^2+4^2+2^2+1^2+6^2+$$

$$+ 5^2+3^2+4^2+3^2+2^2+4^2+5^2+3^2) - \frac{101^2}{25}$$

$$= 473 - \frac{101^2}{25}$$

$$= 473 - 408.04$$

$$= 64.96$$

$$SSA = \frac{1}{5}(21^2+27^2+17^2+19^2+17^2) - \frac{101^2}{25}$$

$$= \frac{1}{5}(2109) - \frac{101^2}{25}$$

$$= 421.8 - 408.04$$

$$= 13.76$$

$SSE = SST - SSA$

$= 64.96 - 13.76$

$= 51.20$

b) Table 10

Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Computed F
Treatment	13.76	4	3.44	1.34375
Error	51.20	20	2.56	
Total	64.96	24		

c) The Test

The critical value of F or $F_{0.05}$ (4,20) is 2.87. Since the computed F value 1.34 is less than the Table value, the null hypothesis of equality of the 5 treatment means is accepted.

Example 3.11 Model I One-Way Classification: Unequal Sample Size

For the data in Table 11 test the hypothesis

$H_0 : u_1 = u_2 = u_3 = u_4$

against the alternative

H_1 : at least two of the means are not equal at the level of significance $\alpha = 0.05$.

Table 11

Sample Data, Sums and Means

	SAMPLE					TOTAL	MEAN
Treatment 1		10	9	7	8	34	8.5
Treatment 2	3	5	6	4	5	23	4.6
Treatment 3		2	4	3		9	3.0
Treatment 4	5	6	4	3	2	20	4.0
						86	5.06

a) Evaluation of the Sums of Squares

$$SST = (10^2+9^2+7^2+8^2+3^2+5^2+6^2+4^2+5^2+2^2+4^2+3^2+5^2+6^2+4^2+3^2+2^2)$$

$$- \frac{86^2}{17}$$

$$= 524 - 435.06$$

$$= 88.94$$

$$SSA = \left(\frac{34^2}{4} + \frac{23^2}{5} + \frac{9^2}{3} + \frac{20^2}{5} \right) - \frac{86^2}{17}$$

$$= 501.80 - 435.06$$

$$= 66.74$$

$$SSE = SST - SSA$$

$$= 88.94 - 66.74$$

$$= 22.20$$

b)

Table 12

Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed F
Treatment	66.74	3	22.247	13.025
Error	22.20	13	1.708	
Total	88.94	16		

c) The Test

Since the computed F, 13.025, is greater than the table value $F_{o.05(3,13)}$ = 3.41, we reject the null hypothesis of equality of the 4 population means.

Exercise 3.8

Four chemicals were used to combat plant lice on potatoes, each chemical being applied to one plot. Twenty leaves were picked from each plot and the number of plant lice on each leaf were recorded. The collected data is contained in the following table:

Table 13

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Chemical 1	12	10	8	5	15	16	11	9	8	7	12	11	13	5	19	12	11	13	9	8
Chemical 2	22	19	18	18	17	15	9	18	15	7	18	20	14	13	17	18	16	11	7	9
Chemical 3	10	8	7	4	3	2	8	9	11	10	8	7	6	2	5	7	8	5	8	7
Chemical 4	11	13	7	20	15	14	13	12	11	9	20	21	8	9	10	11	12	13	18	6

1. Compute the entries in the analysis of variance table

2. Test at a 0.05 level of significance whether differences exist among the chemicals.

3.8.2 Duncan's Multiple Range Test

The analysis of variance provides the test for comparing several population means simultaneously. If the null hypothesis is rejected, that is we conclude that the population means are not equal, we still do not know which of the population means are equal and which are different. A statistical test that is often used to segregate subsets of equal means from a set of significantly different means is the Duncan Multiple Range Test. Two conditions are required to apply the test:

(a) The analysis of variance has led to a rejection of the null hypothesis of homogeneous population means.

(b) The random sample used in the analysis of variance are all of equal size, that is $n_i = n$ for $i = 1, \ldots, a$. The test procedure is as follows:

1. Arrange the sample means in increasing order of magnitude

2. Compute the statistic R_p that is called the least significant range for the p means with

   $p = 2,3,\ldots\ldots,a$, using the formula $R_p =$

   $r_p \sqrt{\frac{s^2}{n}}$, where s^2 is the error mean square

   from the analysis of variance table

   r_p is a quantity called the least significant studentized range, the values of which depend on the used level of significance and the number of degrees of freedom of the error mean square. The values of r_p at the level of significance $\alpha = 0.05$ and for $p = 2,3,\ldots\ldots$ are contained in Table E in the Appendix.

3. Set up the following table to summarize the results of 2).

p	2	....	a
r_p	r_2	....	r_a
R_p	R_2	....	R_a

4. Compare the least significant ranges $R_2,\ldots\ldots$ with the differences in ordered means, proceeding from the right to the left of the set of ordered means. For each comparison: conclude that the two involved means are not significantly different if their difference is less than the appropriate R_p value otherwise conclude that the greater mean is significantly larger than the smaller one.

Example 3.12

Given the following results from the one-way analysis of variance for the problem of exercise 3.8 page 57, with sample size n = 20

$\bar{y}_1$ (Mean of Treatment 1) = 10.7

$\bar{y}_2$ (Mean of Treatment 2) = 15.05

$\bar{y}_3$ (Mean of Treatment 3) = 6.95

$\bar{y}_4$ (Mean of Treatment 4) = 12.65

s^2 = MSE = 14.272

Number of degrees of freedom for the error mean square = 79

Perform the multiple range test procedure to find subsets of homogeneous means from the set of 4 significantly different means.

Solution

1. Means arranged in ascending order

$\bar{y}_3$	$\bar{y}_1$	$\bar{y}_4$	$\bar{y}_2$
6.95	10.7	12.65	15.05

2. From Table E in the Appendix, we read

$r_2 = 2.829$ $r_3 = 2.976$ $r_4 = 3.073$

Thus

$$R_2 = r_2 \sqrt{\frac{s^2}{n}} = 2.829 \sqrt{\frac{14.272}{20}} = 2.389$$

and in similar way

$R_3 = 2.51$, $R_4 = 2.89$

3. The summary table is

p	2	3	4
r_p	2.829	2.976	3.073
R_p	2.389	2.51	2.89

4. Comparing the least significant ranges in the above table with the differences in ordered means leads to the following conclusions:

(a) *Since* $\bar{y}_2 - \bar{y}_4 = 2.4$, $R_2 = 2.389$, *we conclude that* $\bar{y}_2$ *is significantly larger than* $\bar{y}_4$ *and therefore* $u_2 > u_4$. *From this follows that* $u_2 > u_1$ *and* $u_2 > u_3$.

(b) *Since* $y_4 - y_1 = 1.95$, $R_2 = 2.389$, *we conclude that* y_4 *and* y_1 *are not significantly different.*

(c) *Since* $y_4 - y_3 = 5.70$, $R_3 = 2.51$, *we conclude that* y_4 *is significantly larger than* y_3 *and therefore* $u_4 > u_3$.

(d) *Since* $y_1 - y_3 = 3.75$, $R_2 = 2.389$, *we conclude that* y_1 *is significantly larger than* y_3 *and therefore* $u_1 > u_3$.

From these results we would differentiate 3 subsets of population means. The first subset contains one element that is population 2; the second subset contains two elements that are population 4 and population 1; the third subset has one element that is population 3. This can be summarized by writing $u_2 > u_4 \;>= u_1 > u_3$

3.8.3 Bartlett's Test for the Equality of Several Population Variances

When performing an analysis of variance using sample data from a number of a populations we assume that the variances σ^2_i, (i=1,..........,a) of the populations are equal. Departures from this assumption will not affect the F value if the samples are of equal size. This is not the case, however, if unequal sample sizes are involved in the analysis of variance. We should therefore first test the hypothesis of equal population variances before applying the analysis of variance to experimental data with unequal numbers of observations. The null hypothesis to test is:

$$H_0 : \sigma^2_1 = \sigma^2_2 = \ldots\ldots\ldots = \sigma^2_a$$

against the alternative

H_1 : the variances are not equal

The procedure commonly used to test this hypothesis is the Bartlett's test. The steps are as follows:

1. Computation of the a sample variances s^2_1, s^2_2,......., s^2_a from the samples of sizes n_1, n_2 ,........., n_a, with

$$\sum_{i=1}^{a} n_i = N$$

2. Computation of the pooled variance s^2_p that involves the a variances, using the formula

$$s^2_p = \frac{\sum_{i=1}^{a} (n_i-1)\, s^2_i}{N-a}$$

3. Evaluation of the quantities Q and H as

$$Q = (N-a)\log_{10} s^2_p - \sum_{i=1}^{a} (n_i - 1)\log_{10} s^2_i$$

and

$$H = 1 + \frac{1}{3(a-1)} \left(\sum_{i=1}^{a} \frac{1}{n_i - 1} - \frac{1}{N-a} \right)$$

4. Computation of the test statistics B as

$$B = 2.3026 * \frac{Q}{H}$$

5. Comparison of B with $X^2(\alpha, a-1)$ from an appropriate table of the Chi-square distribution (see Table B in the Appendix)

6. Conclusion:

   Reject the hypothesis of equality of the population variances if $B > X^2(\alpha, a-1)$, otherwise do not reject it and conclude that the population variances are equal.

Example 3.13

Test at the level of significance $\alpha = 0.05$ the hypothesis that the variances of the 4 populations in example 3.11 are equal, using the Bartlett's test.

Solution

(1) The sample variances are

$$s^2_1 = \frac{((10)^2 + (9)^2 + (7)^2 + (8)^2 - (34)^2) / 4}{3}$$

$$= \frac{5}{3}$$

$$s^2_2 = \frac{((3)^2 + (5)^2 + (6)^2 + (4)^2 + (5)^2 - (23)^2) / 5}{4}$$

$$= 1.3$$

$$s^2_3 = \frac{((2)^2 + (4)^2 + (3)^2 - (9)^2) / 3}{2} = 1$$

$$s^2_4 = \frac{((5)^2 + (6)^2 + (4)^2 + (3)^2 - (20)^2) / 5}{4}$$

$$= 2.5$$

(2) the pooled variance is

$$s^2_p = \frac{(3)\ (5/3) + (4)\ (1.3) + (2)\ (1) + (4)\ (2.5)}{13}$$

$$= 1.7$$

(3) (a) The value for the quantity Q is

$$Q = 13\log_{10}1.7 - ((3\log_{10}5/3) + 4\log_{10}1.3) + (2\log_{10}1) + (4\log_{10}2.5))$$

$$= 0.1444$$

(b) The value for the quantity H is

$$H = 1 + \frac{1}{9}\ (\frac{1}{3} + {}_{1/4} + 1/2 + {}_{1/4} - \frac{1}{13})$$

$$= 1.1396$$

(4) The value for the statistic B is

$$B = 2.3026 * \frac{0.1444}{1.1396}$$

$$= 0.29$$

(5) $B = 0.29 < X^2(0.05,3) = 7.815$

(6) Conclusion: We accept the hypothesis of homogeneous population variances.

3.8.4 Two-Way Classification Model I

The Problem

In the model I two-way analysis of variance we are interested in analysing the differential fixed effects of certain levels of two factors, let us say A and B. Therefore we include all possible combinations of the interesting levels of the factors in the experimental design. If there are a levels of A and b levels of B then there are a*b possible level combinations obtained by crossing the a and b levels. Now if we denote the levels of A by i, with i=1,...., a and the level of B by j with j=1,......, b then a combination i*j is called a cell, the ij^{th} cell. Each cell may contain one or several observations. By denoting the number of observations in the ij^{th} cell by n_{ij} we might consider three cases:

1) *the n_{ij} are equal but all greater than 1. The experiment is then called "replicated" two-way classification (with interaction)*

2) *the n_{ij} are all equal to 1. We than have a two-way classification with one observation per cell, also called "unreplicated" two-way classification (without interaction)*

3) *the n_{ij} are not equal, due to loss observations. Here we have a two-way classification with missing values.*

3.8.4.1 The "replicated" Two-way Classification

(with interaction) (n observations per cell)

3.8.4.1.1 The Mathematical Model

Given a set of observations yijk, where (i=1,......, a) indicates the levels of factor A, (j=1,...., b) indicates the levels of factor B, (k=1,........., n) indicates the replications in the cell ij. Then the model is:

$$y_{ijk} = u + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where

u	is the total or overall mean
α_i	" " effect of the i^{th} level of factor A
β_j	" " effect of the j^{th} level of factor B
$(\alpha\beta)_{ij}$	is called the interaction between the i^{th} level of factor A and the j^{th} level of factor B
ε_{ijk}	is the deviation of yijk from the mean of the ij^{th} population.

It is assumed that the ε_{ijk} are independent, normally distributed and that they all come from populations with the same variance.

Unique least-square estimators of the parameters are obtained by imposing the following restrictions on the model:

$$\sum_{i=1}^{a} \alpha_i = 0 , \qquad \sum_{j=1}^{b} \beta_j = 0$$

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 , \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$$

Three hypothesis can be tested:

1) H_o : $\alpha_1 = \alpha_2 = \alpha_3 = \ldots\ldots = \alpha_m = 0$

H_1 : at least one of the α_i is not equal to 0.

2) H'_o : $\beta_1 = \beta_2 = \beta_3 = \ldots\ldots = \beta_b = 0$

H'_1 : at least one of the β_j is not equal to 0.

3) H''_o : $(\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{13} = \ldots\ldots (\alpha\beta)_{mb} = 0$

H''_1 : at least one of the $(\alpha\beta)_{ij}$ is not equal to 0.

3.8.4.1.2 The Analysis of Variance

A scheme in the form of Table (14) can be used to ease computation.

Table 14

Two-Way Analysis of Variance, Replicated Experiment

Factor A (rows)	Factor B (columns) 1	2	....	b	Total	Mean
	y_{111}	y_{121}	....	y_{1b1}		
	y_{112}	y_{122}	....	y_{1b2}		
	.	.		.		
1	.	.		.		
	.	.		.		
	y_{11n}	y_{12n}		y_{1bn}		
	$y_{11.}$	$y_{12.}$		$y_{1b.}$	$y_{1..}$	$\bar{y}_{1..}$
	y_{211}	y_{221}		y_{2b1}		
	y_{212}	y_{222}		y_{2b2}		
	.	.		.		
2	.	.	....	.		
.	.	.		.		
.	y_{21n}	y_{22n}		y_{2bn}		
.	$y_{21.}$	$y_{22.}$		$y_{2b.}$	$y_{2..}$	$\bar{y}_{2..}$
.	.	.		.	.	
.	.	.		.	.	
.	.	.		.	.	
.	y_{a11}	y_{a21}		y_{ab1}	.	
.						
.	y_{a12}	y_{a22}		y_{ab2}		
.	.	.		.		
a	.	.	....	.		
	.	.		.		
	y_{a1n}	y_{a2n}		y_{abn}		
	$y_{a1.}$	y_{a2}		$y_{ab.}$	$y_{a..}$	$\bar{y}_{a..}$
Total	$y_{.1.}$	$y_{.2.}$	....	$y_{.b.}$	$y_{...}$	
Mean	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	....	$\bar{y}_{.b.}$		$\bar{y}_{...}$

To derive the sums of squares for the analysis of variance we use the identity

$$y_{ijk} - y_{...} = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

where

$\bar{y}_{...}$	estimates	u
$\bar{y}_{i..} - \bar{y}_{...}$	"	o_i
$\bar{y}_{.j.} - \bar{y}_{...}$	"	$_j$
$\bar{y}_{ij} - \bar{y}_{i..} - y_{.j.} + y_{...}$	"	$(o\quad)_{ij}$
$y_{ijk} - \bar{y}_{ij.}$	"	$_{ijk}$

we then square each term and sum over i, j, k, to obtain:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2$$

$$+\ an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$+\ n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$+\ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$$

For computational purposes the following formulas for the sums of squares can be used:

a) Total Sum of Squares: SST with abn-1 degrees of freedom

$$SST = \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{n} (y_{ijk} - \bar{y}_{...})^2 = \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{n} y^2_{ijk} - \frac{1}{abn} \left(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{n} y_{ijk} \right)^2$$

b) Sum of Squares Due to A SSA with a-1 degrees of freedom

$$SSA = bn \sum_{i}^{a} (\bar{y}_{i..} - \bar{y}_{...})^2 = \frac{1}{bn} \sum_{i}^{a} \left(\sum_{j}^{b} \sum_{k}^{n} yijk \right)^2$$

$$- \frac{1}{abn} \left(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{n} yijk \right)^2$$

c) Sum of Squares Due to B SSB with b-1 degrees of freedom

$$SSB = an \sum_{j=1}^{b} (\bar{y}_{.j.} - \bar{y}_{...})^2 = \frac{1}{an} \sum_{j}^{b} \left(\sum_{i}^{a} \sum_{k}^{n} y_{ijk} \right)^2$$

$$- \frac{1}{abn} \left(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{n} y_{ijk} \right)^2$$

d) Sum of Squares Due to AB SS(AB) with (a-1) (b-1) degrees of freedom

$$SS(AB) = n \sum_{i}^{a} \sum_{j}^{b} (\bar{y}_{ij..} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$= \frac{1}{n} \sum_{i}^{a} \sum_{j}^{b} \left(\sum_{k}^{n} y_{ijk} \right)^2 - \frac{1}{bn} \sum_{i}^{a} \left(\sum_{j}^{b} \sum_{k}^{n} y_{ijk} \right)^2 - \frac{1}{an} \sum_{i=1}^{b} \left(\sum_{j=1}^{a} \sum_{k=1}^{n} y_{ijk} \right)^2$$

$$+ \frac{1}{abn} \left(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{n} y_{ijk} \right)$$

e) <u>Error Sum of Squares: SSE</u> with ab (n-1) degrees of freedom

SSE = SST - SSA - SSB - SS(AB)

The analysis of variance table follows

Table 15 <u>Analysis of Variance Table, Two-Way Classification with Repeated Observations</u>

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Computed F
Factor A	SSA	(a-1)	$MSA = \frac{SSA}{a-1}$	$F_3 = \frac{MSA}{MSE}$
Factor B	SSB	(b-1)	$MSB = \frac{SSB}{b-1}$	$F_2 = \frac{MSB}{MSE}$
Interaction A * B	SS(AB)	(a-1) (b-1)	$MS(AB) = \frac{SS(AB)}{(a-1)(b-1)}$	$F_1 = \frac{MS(AB)}{MSE}$
Error	SSE	ab(n-1)	$MSE = \frac{SSE}{ab(n-1)}$	
Total	SST	abn-1		

3.8.4.1.3 The Tests

Tests are performed by comparing the computed F values with corresponding $F_{\alpha}(n_1, n_2)$ from the table of the F distribution. Frequently the null hypothesis H_o : all $(\alpha\beta)_{ij} = 0$ is tested first. If this test is not significant, the interaction and residual sums of squares SS(AB) and (SSE) may be or may not be added together (pooled) before testing the effects of A and B. If the investigator chooses not to pool SS(AB) and SSE, he will use F_2 and F_3 in the last column of table 15 to test the effect of Factor B and Factor A respectively. But if chooses to pool he will have to replace the demoninator MSE in F_1 and F_2 by MSE_p, that is

$$MSE_p = \frac{SS(AB + SSE}{(a-1)\ (b-1) + ab(n-1)}$$

(pooled sums of squares)

(pooled numbers of degrees of freedom)

3.8.4.1.4 The Meaning of Interactions

It is not always easy to interpret interactions because an apparent interaction may be real or it may be due to experimental error. If the factors under study are quantitative graphing the cells means $(y_{ij.})$ may help in interpreting the interaction. When lines joining the cell means appear roughly parallel as in Fig. 7 there will be no significant interaction. This implies that the differential effects between the levels of factor A for instance are the same at all levels of the Factor B. Non parallel lines like the ones in Fig. 8 would lead us to expect a rather large interaction and suggest that the third level of A interact positively with the third level of B.

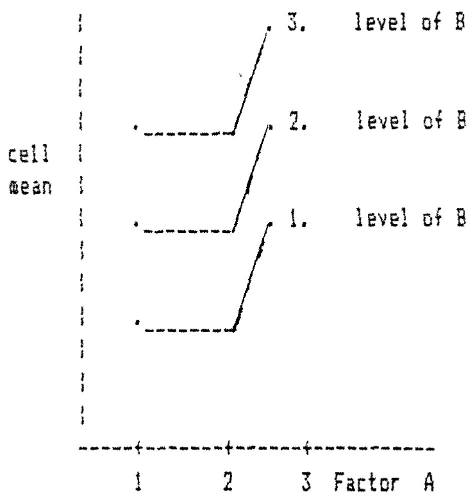

3. level of B
2. level of B
1. level of B
cell
mean
1
2
3 Factor A

FIG. (7)

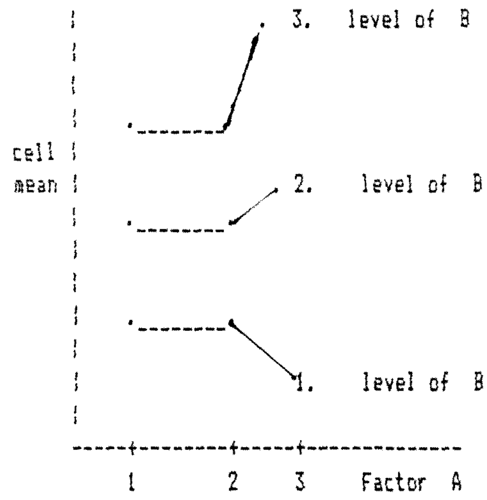

3. level of B
2. level of B
1. level of B
cell
mean
1
2
3
Factor A

FIG. (8)

Example 3.14

Three varieties of jute are being compared for yield. The experiment was conducted by using 12 uniform plots at each of 4 locations. The 3 varieties of jute are each planted on 4 plots selected at random for each location. The hypothetical yields per plot were as follows:

Table 16

Factor A (Location)	Factor B (Variety of jute) $j = 1$	$j = 2$	$j = b = 3$
$i = 1$	9 7 6 5	8 3 4 5	4 2 3 6
$i = 2$	10 9 7 6	7 4 3 6	1 5 4 7
$i = 3$	11 8 9 10	6 8 7 5	2 3 6 5
$i = a = 4$	9 7 10 11	5 6 7 9	5 4 7 8

Use a level of significance = 0.05 to test the hypothesis that

(a) *There is no difference in the average yield of the varieties of jute when planted at different locations:*

H_o *(null hypothesis):* ${}_1 = {}_2 = {}_3 = {}_4 = 0$

H_1 *(alternative hypothesis): at least one of the* ${}_i$ *is not equal to zero*

(b) *There is no difference in the yielding capabilities of the three varieties of jute*

H'_o *(null hypothesis):* $3_1 = 3_2 = 3_3 = 0$

H'_1 *(alternative hypothesis): at least one of the* B_j *is not equal to zero*

(c) *The locations and varieties of jute do not interact*

H''_o *(null hypothesis):* $(\ 3)_{11} = (\ 3)_{12} = (\ 3)_{13} = \ldots = (\ 3)_{43} = 0$

H''_1 *(alternative hypothesis): at least one of the* $(\ 3)_{ij}$ *is not equal to zero*

<u>Solution</u>

From table 16 we construct the following table of totals and means

Table 17 Totals and Means

	VARIETY				
Location	1	1	3	TOTAL	MEAN
1	9 7 6 5 —— 27	8 3 4 5 —— 20	4 2 3 6 —— 15	62	5.167
2	10 9 7 6 —— 32	7 4 3 6 —— 20	1 5 4 7 —— 17	69	5.750
3	11 8 9 10 —— 38	6 8 7 5 —— 26	2 3 6 5 —— 16	80	6.667
4	9 7 10 11 —— 37	5 6 7 9 —— 27	5 4 7 8 —— 24	88	7.333
TOTAL	134	93	72	299	
MEAN	8.375	5.813	4.500		6.229

<u>Evaluation of the Sums of Squares</u>

a) <u>Total Sum of Squares</u>

$$\sum_{i=1}^{4}\sum_{j=1}^{3}\sum_{k=1}^{4} Y^2_{ijk} = 9^2+7^2+6^2+5^2+8^2+3^2+4^2+2^2+3^2+6^2+10^2+9^2+7^2+11^2+8^2+9^2 10^2+6^2+8^2+7^2+5^2+2^2+3^2+6^2+5^2+9^2+7^2+10^2+11^2+5^2+6^2+7^2+9^2+5^2+4^2+7^2+8^2$$

$$= 2147$$

$$\left(\sum_{i=1}^{4}\sum_{j=1}^{3}\sum_{k=1}^{4} Y_{ijk}\right)^2 / abn = (9 + 7 + 6 + \ldots, + 4 + 7 + 8)^2/_{48}$$

$$= \frac{299^2}{48}$$

$$SST = 2147 - \frac{299^2}{48}$$

$$= 2147 - 1862.52$$

$$= 284.48$$

b) <u>Sum of Squares due to A</u>

$$\frac{1}{bn}\sum_{i=1}^{4} Y^2_{i..} = \frac{1}{12}(62^2 + 69^2 + 80^2 + 7744)$$

$$= \frac{1}{12}(3844 + 4761 + 6400 + 7744)$$

$$= \frac{1}{12}(22749)$$

$$= 1895.75$$

$$SSA = 1895.75 - \frac{299^2}{48}$$

$$= 1895.75 - 1862.52$$

$$= 33.23$$

c) <u>Sum of Squares due to B</u>

$$\frac{1}{an}\sum_{j=1}^{3} y_{.j.} = \frac{1}{16}(134^2 + 93^2 + 72^2)$$

$$= \frac{1}{16}(17956 + 8649 + 5184)$$

$$= \frac{1}{16}(31789)$$

$$= 1986.18$$

$$SSB = 1986.81 - 1862.52$$

$$= 124.29$$

d) <u>Sum of Squares due to AB</u>

$$\frac{1}{n}\sum_{i=1}^{4}\sum_{j=1}^{3} y^2_{ij.} = \frac{1}{4}(27^2+20^2+15^2+32^2+20^2+17^2+38^2+26^2+16^2+37^2+27^2+24^2)$$

$$= \frac{1}{4}(8117)$$

$$= 2029.25$$

$$SS(AB) = 2029.25 - 1895.75 - 1986.18 + \frac{299^2}{48}$$

$$= 9.84$$

e) <u>Error Sum of Squares</u>

$$SSE = 284.48 - 33.23 - 124.29 - 9.84$$

$$= 117.12$$

Table 18 Analysis of Variance Table Two-way Classification,

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed F
Factor A	33.23	3	11.077	$F_3 = 3.41$
Factor B	124.29	2	62.15	$F_2 = 19.1$
Interaction AB	9.84	6	1.64	$F_1 = 0.5$
Error	117.24	36	3.25	
Total	284.48	47		

Conclusions

(a) *Since $F_1 = 0.5$ is less than $F_{0.05}(6,36) = 2.38$, do not reject H''_0 and conclude that there is no interaction between the different locations and the different varieties of jute.*

(b) *Since $F_2 = 19.1$ is greater than $F_{0.05}(2,36)$, reject H'_0 and conclude that a difference in the average yields of the three varieties of jute exists.*

(c) *Since $F_3 = 3.41$ is greater than the table value $F_{0.05}(3,36) = 2.9$, reject the null hypothesis H_0 and conclude that a difference in the average yields of jute exists when planted at the different locations.*

If the cell means that have been computed from Table 17, (see Table 19) are plotted, we obtain the three lines in figure 9. We see that the lines show a pattern of parallelism, which is in accordance with the above results that no interaction is present. Thus a change in location does not produce a different change in the average yields of the varieties of jute.

Table 19 Cell Means

	VARIETY		
Location	1	2	3
1	9	6.66	5
2	10.66	6.66	5.66
3	12.66	8.66	5.33
4	12.33	9	8

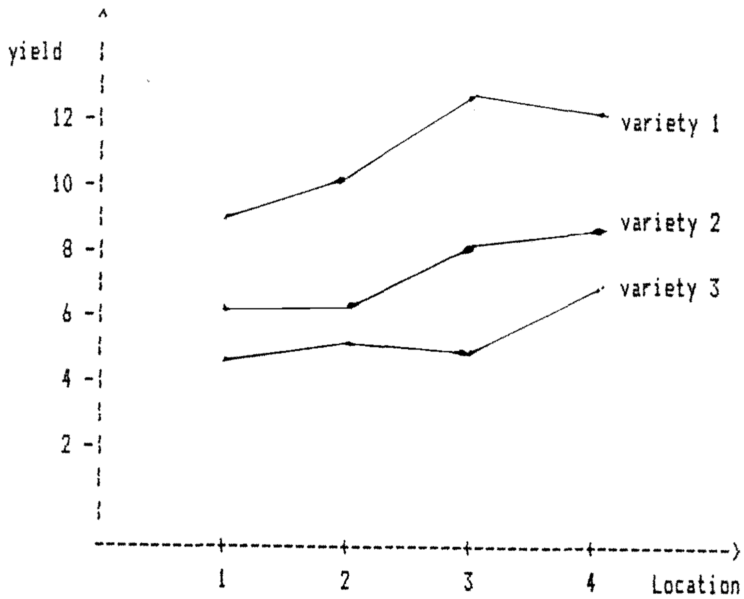


Figure 9

Interaction in the Block Experiment of Example 3.14

3.8.4.2 The "Unreplicated" Two-way Classification

(no interaction)

3.8.4.2.1 The Mathematical Model

Given the following observations y_{ij}, where $(i=1,.........,a)$ indicates the levels of factor A, and $(j=1,........,b)$ indicates the levels of factor B, with $k=1$ in every cell, each observation can be represented by the model

$$y_{ij} = u + \alpha_i + \beta_j + \varepsilon_{ij}$$

where

u is the total mean

α_i is the effect of the i_{th} of the a levels of factor A

β_j is the effect of the j^{th} of the b levels of factor B

ε_{ij} are independently and normally distributed error variables.

Since the experiment is unreplicated the error sum of squares can not be evaluated in the same way as in the replicated case, that is by using the observations within the cells. But it is assumed that the effects due to the interaction of A and B, $(\alpha\beta)_{ij}$, are zero. Then the interaction sum of squares and degrees of freedom from the replicated case is used as the residual or error sum of squares and degrees of freedom.

To obtain unique least-squares estimators for the parameters the following restrictions are imposed on the model

$$\sum_{i=1}^{a} \alpha_i = 0 \quad , \quad \sum_{j=1}^{b} \beta_j = 0$$

Here are two hypothesis can be tested. They are

(a) H_0 : $\alpha_1 = \alpha_2 = \dots\dots\dots = \alpha_a = 0$

H_1 : at least one of the α_i is not equal to 0

(b) H'_0 : $\beta_1 = \beta_2 = \dots\dots\dots = \beta_b = 0$

H'_1 : at least one of the β_j is not equal to 0.

Table 20

Two-way Classification with no Replication

(One observation per cell)

Factor A (rows)	Factor B (Columns) 1	2	......	b	Total	Mean
1	y_{11}	y_{12}	......	y_{1b}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	....	y_{2b}	$y_{2.}$	$\bar{y}_{2.}$
.	.				.	.
.	.				.	.
.	.				.	.
a	y_{a1}	y_{a2}	....	y_{ab}	$y_{a.}$	$\bar{y}_{a.}$
Total	$y_{.1}$	$y_{.2}$	....	$y_{.b}$	$y_{..}$	
Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	....	$\bar{y}_{.b}$		$\bar{y}_{..}$

3.8.4.2.2 The Analysis of Variance

Table (20) simplifies the computation of the parameter estimates and the sum of squares

Since the subscript k is omitted in the model the identity used in deriving the sums of squares is:

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..})$$

$$+ (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}x)$$

We estimate u by $\bar{y}_{..}$

" " α_i " $\bar{y}_i - \bar{y}_{..}$

" " β_j " $\bar{y}_{.j} - \bar{y}_{..}$

" " ε_{ij} " $y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$

By squaring each term of the identity and sum over i,j we obtain

$$\sum_{i}^{a}\sum_{j}^{b}(y_{ij}-\bar{y}_{..})^2 = b\sum_{i}^{a}(\bar{y}_{i}-\bar{y}_{..})^2 + a\sum_{j}^{b}(\bar{y}_{.j}-\bar{y}_{..})^2$$

$$+\sum_{i}^{a}\sum_{j}^{b}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}-\bar{y}_{..})^2$$

The following sum of squares formulas are more convenient for computation:

<u>Total sum of squares</u> SST with (ab-1) degrees of freedom

$$SST = \sum_{i}^{a}\sum_{j}^{b}(y_{ij}-\bar{y}_{..})^2 = \sum_{i}^{a}\sum_{j}^{b}y^2_{ij} - \frac{1}{ab}\left(\sum_{i}^{a}\sum_{j}^{b}y_{ij}\right)^2$$

<u>Sum of squares due to factor A:</u> SSA with a-1 degrees of freedom

$$SSA = b\sum_{i}^{a}(\bar{y}_{i.}-\bar{y}_{..}) = \frac{1}{a}\sum_{i}^{a}\left(\sum_{j}^{b}y_{ij}\right)^2 - \frac{1}{ab}\left(\sum_{i}^{a}\sum_{k}^{b}y_{ij}\right)^2$$

<u>Sum of squares due to factor B:</u> SSB with b-1 degrees of freedom

$$SSB = a \sum_{j}^{b} (y_{i.} - y_{..})^2 = \frac{1}{b} \sum_{j}^{b} \left(\sum_{i}^{a} y_{ij} \right)^2 - \frac{1}{ab} \left(\sum_{i}^{a} \sum_{j}^{b} y_{ij} \right)^2$$

<u>Error sum of squares:</u> SSE with (a-1) (b-1) degrees of freedom

SSE = SST - SSA - SSB

The results are presented in table (21)

Table 21

Analysis of Variance Table

Unreplicated Two-way Classification, Model I

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed F
Factor A	SSA	(a-1)	$MSA = \frac{SSA}{a-1}$	$F_1 = \frac{MSA}{MSE}$
Factor B	SSB	(b-1)	$MSB = \frac{SSB}{b-1}$	$F_2 = \frac{MSB}{MSE}$
Error	SSE	(a-1) (b-1)	$MSE = \frac{SSE}{(a-1)(b-1)}$	

Remark:

In the unreplicated two-way classification we selected particular a, b levels of two factors A and B and assigned the a*b experimental units to the a*b level combinations, whereby one observation is made in each unit. Such an experimental design is usually called a completely randomized design with two factors.

In some cases the investigator might consider the levels of one factor as fixed blocks and the levels of the other factor as treatments that are assigned randomly to each of the blocks. Such designs are usually called randomized complete block designs. Each block is then considered as a replicate.

The main difference between these two designs is that in the two-way completely randomized design both factors are to be studied; whereas in the one-way randomized complete block design one usually wishes to study one factor and eliminate the other.

But the same model can be used for both designs; estimation of parameters and tests are also the same.

Example 3.15

Three varieties of sweet potato are being compared for yield. The experiment was conducted by assigning each variety at random to 3 equal-size plots at each of 5 different locations. The following hypothetical yields per plot were recorded:

Table 22

	Factor B (Variety of Sweet Potato)		
Factor A (Location)	1	2	3
1	12	10	8
2	9	7	7
3	8	6	6
4	10	8	5
5	13	5	9

Use a level of significance = 0.05 to test the following hypotheses

(a) There is no difference in the yielding capabilities of the three varieties of sweet potato

$H'_0 : \beta_1 = \beta_2 = \beta_3 = 0$

H'_1 : *at least one of the* β_j *is not equal to 0.*

(b) There is no significant difference due to location in the average yield of the varieties of sweet potato:

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$

H_1 : at least one of the α_i is not equal to zero.

Table 23

Factor A	Factor B 1	2	3	TOTAL	MEAN
1	12	10	8	30	10.00
2	9	7	7	23	7.67
3	8	6	6	20	6.67
4	10	8	5	23	7.67
5	13	5	9	27	9.00
Total	52	36	35	123	
MEAN	10.40	7.20	7.00		8.20

Solution

Evaluation of the Sums of Squares

$$\sum_{i=1}^{5} \sum_{j=1}^{3} y^2_{ij} =$$

$$12^2+10^2+8^2+9^2+7^2+7^2+8^26^2+6^2+10^2+8^2+5^2+13^2+5^2+9^2$$

$$= 144+100+64+81+49+49+64+36+36+100+64+25+169+25+81$$

$$= 1087$$

$$SST = 1087 - \frac{123^2}{15}$$

$$= 1087 - 1008.6$$

$$= 78.4$$

$$\frac{1}{3}\sum_{i=1}^{5} y^2_{i.} = \frac{1}{3}(30^2+23^2+20^2+23^2+27^2)$$

$$= \frac{1}{3}(900+529+400+529+729)$$

$$= \frac{1}{3}(3087)$$

$$= 1029$$

$$SSA = 1029 - \frac{123^2}{15}$$

$$= 1029 - 1008.6$$

$$= 20.4$$

$$\frac{1}{5}\sum_{j=1}^{3} y^2_{.j} = \frac{1}{5}(52^2+36^2+35^2)$$

$$= \frac{1}{5}(2704+1296+1225)$$

$$= \frac{1}{5}(5225)$$

$$= 1045$$

$$SSB = 1045 - \frac{123^2}{15}$$

$$= 1045 - 1008.6$$

$$= 36.4$$

$$SSE = SST - SSA - SSB$$

$$= 78.4 - 20.4 - 36.4$$

$$= 21.6$$

Table 24 Analysis of Variance Table Two-way Classification, One Observation per cell

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed F
Factor A	20.4	4	5.1	$F_1 = 1.89$
Factor B	36.4	2	18.2	$F_2 = 6.74$
Error	21.6	8	2.7	

Conclusions

(a) Since $F_2 = 6.74$ is greater than the table value $F_{0.05}(2,8) = 4.46$, we reject the null hypothesis H'_o and conclude that a difference in the yielding capabilities of the three varieties of the of sweet potato exists.

(b) Since $F_1 = 1.89$ is less than the table value $F_{0.05}(4,8) = 3.84$, we accept the null hypothesis H_o and conclude that no significant difference in the average yields of the three varieties of sweet potato due to location exists.

Exercise 3.9

Project Title: Weeding Trial with Eucalyptus

Objective: To investigate growth of Eucalyptus establised under different weeding conditions.

Method: Three treatments:
- No weeding (I)
- Line weeding (II)
- Clean weeding (III)

were applied to the 12 plots of a 4 times replicated trial as shown in the following layout. The trial was established with 10 months old Eucalyptus plants. After 4 years the experiment was assessed and among other characteristics the height of 6 randomly selected trees measured in each plot. No interaction is assumed between treatment and blocks.

Block 1 Block 4

II	III	I	II
I	II	III	I
III	I	II	III

Block 2 Block 3

The height data (in m) collected is as follows:

Table 25

Factor A (block)	Factor B (Treatment) A	B	C
I	1.60	0.82	2.10
	1.55	0.95	1.90
	1.50	0.49	4.00
	0.70	0.75	3.90
	0.45	0.96	2.30
	0.50	0.20	1.80
II	1.63	1.50	1.25
	2.40	1.82	1.47
	1.74	2.30	2.00
	1.80	2.72	1.30
	3.66	2.70	1.55
	1.54	1.50	1.75
III	2.09	2.40	3.40
	0.50	2.40	4.60
	1.20	2.04	5.40
	2.30	3.27	3.40
	1.24	1.80	2.80
	1.15	3.60	3.60
IV	1.09	1.85	4.10
	3.00	2.40	3.50
	1.34	1.89	3.60
	2.30	1.30	4.10
	1.72	2.64	2.70
	2.80	1.80	2.30

Carry out the analysis of variance for the data in Table. 25 Test at the significance level $\alpha = 0.05$ whether the 3 treatment means differ significantly and draw conclusions concerning the treatments.

3.8.5 Three-Way Classification Model I ($n > 1$)

3.8.5.1 The problem

Here we are studying the fixed effects of certain levels of 3 factors and their interactions on the response of an experiment. For instance if we want to investigate the various effects of 2 fertilizer levels and 2 water levels on the growth of seedlings from 3 different tree species, we might have included all possible combinations of the levels of the factors in our experiment by crossing them to obtain 12 treatment combinations. Assuming that each treatment combination is repeated 4 times, we have 48 plots of experimental units to be assigned at random to the 12 treatment combinations or cells. Generally, if the factors are called A, B, C and a denotes the number of levels of A, b denotes the number of levels of B and c denotes the number of levels of C being investigated, the experiment will contain a*b*c cells. By using the indices i,j,k to specify the levels of A, B, C, with $i=1,\ldots,a$, $j=1,\ldots,b$ and $k=1,\ldots,c$, each cell is uniquely determined by the three digit number (ijk) corresponding to its level combination. Lastly, if we have n plots within each cell and use the letter l to represent the plot number, with $l = 1,\ldots,n$, each observation or plot value can be represented by the variable y_{ijkl}.

A way of visualizing the three way experimental design is to represent it by a three dimensional lattice in which the index i refers to rows, j to columns, k to arrays; each array containing (a*b) cells, each of which with n observations. (Fig. 10).

We could have as well selected the first or the second index to denote the array number.

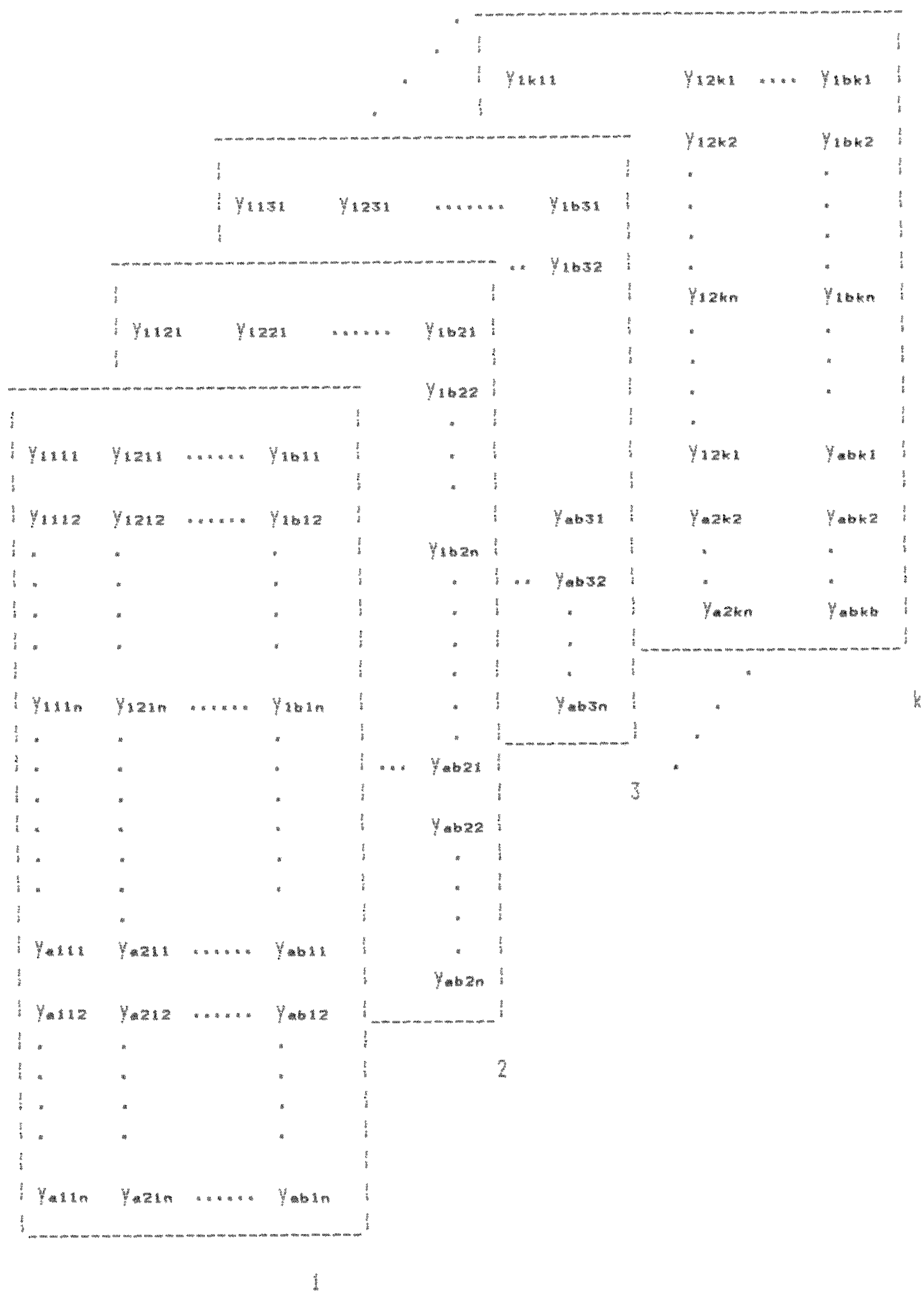

y1k11 y12k1 .... y1bk1
y12k2 y1bk2
y12kn y1bkn
y12k1 yabk1
ya2k2 yabk2
ya2kn yabkb
k
y1131 y1231 .... y1b31
.. y1b32
yab31
.. yab32
yab3n
3
y1121 y1221 ...... y1b21
y1b22
y1b2n
... yab21
yab22
yab2n
2
y1111 y1211 ...... y1b11
y1112 y1212 ...... y1b12
y111n y121n y1b1n
ya111 ya211 ...... yab11
ya112 ya212 ...... yab12
ya11n ya21n yab1n
1

Fig. 10

3.8.5.2 The Mathematical Model

Each observation is expressed as follows

$$y_{ijkl} = u + \alpha_i + \beta_j + \lambda_k + (\alpha\beta)_{ij} + (\alpha\lambda)_{ik} + (\beta\lambda)_{jk} + (\alpha\beta\lambda)_{ijk} + \varepsilon_{ijk}$$

With

$i = 1, \ldots\ldots\ldots, a$

$j = 1, \ldots\ldots\ldots, b$

$k = 1, \ldots\ldots\ldots, c$

$l = 1, \ldots\ldots\ldots, n$

The restrictions on the models are:

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{k=1}^{c} \lambda_k = \sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij}$$

$$= \sum_{i=1}^{a} (\alpha\lambda)_{ik} = \sum_{k=1}^{c} (\alpha\lambda)_{ik} = \sum_{k=1}^{b} (\beta\lambda)_{jk}$$

$$= \sum_{k=1}^{c} (\beta\lambda)_{jk} = \sum_{i=1}^{a} (\alpha\beta\lambda)_{ijk} = \sum_{j=1}^{b} (\alpha\beta\lambda)_{ijk}$$

$$= \sum_{k=1}^{c} (\alpha\beta\lambda)_{ijk} = 0$$

Meaning of the Model Components

u is the total mean

α_i " " mean effect of the i^{th} level of factor A

β_j " " mean effect of the j^{th} level of factor B

λ_k " " mean effect of the k^{th} level of factor C

$(\alpha\beta)_{ij}$ is the mean interaction of the i^{th} level of factor A and the j^{th} level of factor B.

$(\alpha\lambda)_{ik}$ " " interaction of the i^{th} level of factor A and the k^{th} level of factor C.

$(\beta\lambda)_{jk}$ " " interaction of the j^{th} level of factor B and the k^{th} level of factor C

$(\alpha\beta\lambda)_{ijk}$ " " interaction between the i^{th} level of fact-or A, the j^{th} level of factor B and the k^{th} level of factor C.

ϵ_{ijk} " " deviation of the observation y_{ijk} from the mean of the ijk_{th} population.

The ϵ_{ijk} are assumed to be independent, normally distributed variables with mean 0 and variance σ^2.

The Null Hypothesis

Up to 7 hypothesis can be tested: these hypotheses are:

1) H_{10} : $\alpha_1 = \alpha_2 = \ldots\ldots = \alpha_m = 0$

H_{11} : at least one of the α_i is not equal to zero

2) H_{20} : $\beta_1 = \beta_2 = \ldots\ldots = \beta_b = 0$

H_{21} : at least one of the β_j is not equal to zero

3) H_{30} : $\lambda_1 = \lambda_2 = \ldots\ldots = \lambda_c = 0$

H_{31} : at least one of the λ_k is not equal to zero

4) H_{40} : $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots\ldots = (\alpha\beta)_{ab} = 0$

H_{41} : at least one of the $(\alpha\beta)_{ij}$ is not equal to zero

5) H_{50} : $(\alpha\lambda)_{11} = (\alpha\lambda)_{12} = \ldots\ldots = (\alpha\lambda)_{ac} = 0$

H_{51} : at least one of the $(\alpha\lambda)_{ik}$ is not equal to zero

6) H_{60} : $(\beta\lambda)_{11} = (\beta\lambda)_{12} = \ldots\ldots = (\beta\lambda)_{bc} = 0$

H_{61} : (at least one of the $(\beta\lambda)_{jk}$ is not equal to zero

7) H_{70} : $(\alpha\beta\lambda)_{111} = (\alpha\beta\lambda)_{112} = \ldots\ldots = (\alpha\beta\lambda)_{1bc} = 0$

H_{71} : at least one of the $(\alpha\beta\lambda)_{ijk}$ is not equal to zero.

3.8.5.3 <u>The Analysis of Variance</u>

Tables 26,27,28 and 29 simplify the sum of squares computat-ions

<u>Table 26</u> Observations and Cell sums and means for the three-way classification with n replications per cell

	Factor C		1	
	Factor B	1	2 ....b	
Factor A		y_{1111}	y_{1211}	y_{1b11}
		y_{1112}	y_{1212}	y_{1b12}
		. $\bar{y}_{111.}$	. $\bar{y}_{121.}$	. $\bar{y}_{1b1.}$
		y_{111n}	y_{121n}	y_{1b1n}
		$y_{111.}$	$y_{121.}$	$y_{1b1.}$

	Factor C		2	
	Factor B	1	2............b	
		y_{1121}	y_{1221}	y_{1b21}
1		y_{1122}	y_{1222}	y_{1b22}
		. $\bar{y}_{112.}$	. $\bar{y}_{122.}$	. $\bar{y}_{1b2.}$
		y_{112n}	y_{122n}	y_{1b2n}
		$y_{112.}$	$y_{122.}$	$y_{1b2.}$
			.	
			.	

	Factor C		C	
	Factor B	1	2............b	
		y_{11c1}	y_{12c1}	y_{1bc1}
		y_{11c2}	y_{12c2}	y_{1bc2}
		. $\bar{y}_{11c.}$	. $\bar{y}_{12c.}$	. $\bar{y}_{1bc.}$
		y_{11cn}	y_{12cn}	y_{1bcn}
		$y_{11c.}$	$y_{12c.}$	$y_{1bc.}$

Factor C 1

Factor B 1 2 .... b

Factor A 2

	y_{2111}	y_{2211}	y_{2b11}
	y_{2112}	y_{2212}	y_{2b12}
	. $\bar{y}_{211.}$	. $\bar{y}_{221.}$	. $\bar{y}_{2b1.}$
	y_{211n}	y_{221n}	y_{2b1n}
	$y_{211.}$	$y_{221.}$	$y_{2b1.}$

Factor C 2

	y_{2121}	y_{2221}	y_{2b21}
	y_{2122}	y_{2222}	y_{2b22}
	. $\bar{y}_{212.}$	. $\bar{y}_{222.}$	. $\bar{y}_{2b2.}$
	y_{212n}	y_{222n}	y_{2b2n}
	$y_{212.}$	$y_{222.}$	$y_{2b2.}$

.
.
.

Factor C c

	y_{21c1}	y_{22c1}	y_{2bc1}
	y_{21c2}	y_{22c2}	y_{2bc2}
	. $\bar{y}_{21c.}$	. $\bar{y}_{22c.}$	. $\bar{y}_{2bc.}$
	y_{21cn}	y_{22cn}	y_{2bcn}
	$y_{21c.}$	$y_{22c.}$	$y_{2bc.}$

1

Factor C

Y_{a111}	Y_{a211}	Y_{ab11}
Y_{a112}	Y_{a212}	Y_{ab12}
.	.	.
. $\bar{Y}_{a11.}$	. $\bar{Y}_{a21.}$	. $\bar{Y}_{ab1.}$
Y_{a11n}	Y_{a21n}	Y_{ab1n}
$Y_{a11.}$	$Y_{a21.}$	$Y_{ab1.}$

2

Factor C

Y_{a121}	Y_{a221}	Y_{ab21}
Y_{a122}	Y_{a222}	Y_{ab22}
.	.	.
. $\bar{Y}_{a12.}$	. $\bar{Y}_{a22.}$	. $\bar{Y}_{ab2.}$
.	.	.
Y_{a12n}	Y_{a22n}	Y_{ab2n}
$Y_{a12.}$	$Y_{a22.}$	$Y_{ab2.}$

a

.
.
.

C

Factor C

Y_{a1c1}	Y_{a2c1}	Y_{abc1}
Y_{a1c2}	Y_{a2c2}	Y_{abc2}
.	.	.
. $\bar{Y}_{a1c.}$	. $\bar{Y}_{a2c.}$	. $\bar{Y}_{abc.}$
.	.	.
.	.	.
Y_{a1cn}	Y_{a2cn}	Y_{abcn}
$Y_{a1c.}$	$Y_{a2c.}$	$Y_{abc.}$

A \ B	1	2	...	b	
1	$y_{11..}$	$y_{12..}$	....	$y_{1b..}$	$y_{1...}$
	$\bar{y}_{11..}$	$\bar{y}_{12..}$		$\bar{y}_{1b..}$	$\bar{y}_{1...}$
2	$y_{21..}$	$y_{22..}$	....	$y_{2b..}$	$y_{2...}$
	$\bar{y}_{21..}$	$\bar{y}_{22..}$		$\bar{y}_{2b..}$	$\bar{y}_{2...}$
.		.			.
.		.			.
.	$y_{a1..}$	$y_{a2..}$	....	$y_{ab..}$	$y_{a...}$
a	$\bar{y}_{a1..}$	$\bar{y}_{a2..}$		$\bar{y}_{ab..}$	$\bar{y}_{a...}$
	$y_{.1..}$	$y_{.2..}$	....	$y_{.b..}$	$y_{....}$
	$\bar{y}_{.1..}$	$\bar{y}_{.2..}$		$\bar{y}_{.b..}$	$\bar{y}_{....}$

Table 27 Sums and Means from the cell means obtained by summing and averaging over factor C

A \ C	1	2	...	b	
1	$y_{1.1.}$	$y_{1.2.}$	....	$y_{1.c.}$	$y_{1...}$
	$\bar{y}_{1.1.}$	$\bar{y}_{1.2.}$		$\bar{y}_{1.c.}$	$\bar{y}_{1...}$
2	$y_{2.1.}$	$y_{2.2.}$	....	$y_{2.c.}$	$y_{2...}$
	$\bar{y}_{2.1.}$	$\bar{y}_{2.2.}$		$\bar{y}_{2.c.}$	$\bar{y}_{2...}$
.	.				.
.	.				.
.	$y_{a.1.}$	$y_{a.2.}$	....	$y_{a.c.}$	$y_{a...}$
a	$\bar{y}_{a.1.}$	$\bar{y}_{a.2.}$		$\bar{y}_{a.c.}$	$\bar{y}_{a...}$
	$y_{..1.}$	$y_{..2.}$	....	$y_{..c.}$	$y_{...}$
	$\bar{y}_{..1.}$	$\bar{y}_{..2.}$		$\bar{y}_{..c.}$	$\bar{y}_{...}$

Table 28 Sums and means from the cell means obatined by summing and averaging over factor B

B \ C	1	2	...	c	
1	$y_{.11.}$	$y_{.12.}$	....	$y_{.1c.}$	$y_{.1..}$
	$y_{.11.}$	$y_{.12.}$		$y_{.1c.}$	$y_{.1..}$
2	$y_{.21.}$	$y_{.22.}$	....	$y_{.2c.}$	$y_{.2..}$
.	$y_{.21.}$	$y_{.22.}$		$y_{.2c.}$	$y_{.2..}$
.					
.	$y_{.b1.}$	$y_{.b2.}$	....	$y_{.bc.}$	$y_{.b..}$
b	$y_{.b1.}$	$y_{.b2.}$		$y_{.bc.}$	$y_{.b..}$
	$y_{..1.}$	$y_{..2.}$	...	$y_{..c.}$	$y_{....}$
	$y_{..1.}$	$y_{..2.}$		$y_{..c.}$	$y_{...}$

Table 29 Sums and Means from the cell means obtained by summing and averaging over Factor A

With regard to the model components we divide the deviation of each observation from the total mean into 8 parts. We have

$$y_{ijkl} - \bar{y}_{....} =$$

$$(\bar{y}_{i...} - \bar{y}_{....}) + (\bar{y}_{.j..} - \bar{y}_{....})$$

$$+ \quad (\bar{y}_{..k.} - \bar{y}_{....}) + (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....})$$

$$+ \quad (\bar{y}_{i.k.} - \bar{y}_{i...} - \bar{y}_{..k.} + \bar{y}_{....})$$

$$+ \quad (\bar{y}_{.jk.} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....})$$

$$+ \quad (\bar{y}_{ijk.} - \bar{y}_{ij..} - \bar{y}_{i.k.} - \bar{y}_{.jk.} + \bar{y}_{i...} + \bar{y}_{.j..}$$

$$+ \bar{y}_{..k.} - \bar{y}_{...})$$

$$+ \quad (y_{ijkl} - y_{ijk.})$$

The elements of the identity are used to estimate the model parameters as follows:

(1) $\bar{y}_{.....}$ is the mean response for the abc treatments. It estimates the parameter u

(2) $(\bar{y}_{i...} - \bar{y}_{....})$ that is the difference between the mean for the bc treatments involving level i of factor A, estimates the parameter α_i

(3) $(\bar{y}_{.j..} - \bar{y}_{....})$ is the difference between the mean for the ac treatments involving level j of factor B and estimates β_j

(4)	$(\bar{y}_{..k.} - \bar{y}_{....})$	is the difference between the mean for the ab treatments involving level k of factor C and estimates λ_k
(5)	$(\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{....})$	estimates the interaction $(\alpha\beta)_{ij}$. It is the mean for the c treatments that involve the i_{th} level of A and j^{th} level of B minus [(1) + (2) + (3)]
(6)	$(\bar{y}_{i.k} - \bar{y}_{i...} - \bar{y}_{..k} + \bar{y}_{....})$	estimates the interaction $(\alpha\lambda)_{ik}$. It is the mean for the b treatments involving the i^{th} of A and the k^{th} level of C minus [(1) + (2) + (4)]
(7)	$(\bar{y}_{.jk.} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....})$	estimates the interaction $(\beta\lambda)_{jk}$. It is obtained by subtracting [(1) + (3) + (4)] from the mean for the a involving the j_{th} level of B and the k^{th} level of C
(8)	$(\bar{y}_{ijk.} - \bar{y}_{ij..} - \bar{y}_{i.k.} - \bar{y}_{.jk.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} - \bar{y}_{....})$	estimates the interaction $(\alpha\beta\lambda)_{ijk}$. It is equal to the mean for the ijk^{th} cell minus [(1) + (2) + (3) + (4) + (5) + (6) + (7)]
(9)	$(\bar{y}_{ijkl} - \bar{y}_{ijk.})$	estimates ϵ_{ijkl}. It is the difference between the i^{th} observation in the ijk^{th} cell and that cell mean.

We square the terms of the identity and sum over i,j,k,l to obtain

$$\sum_{i}^{a}\sum_{j}^{b}\sum_{k}^{c}\sum_{l}^{n} (y_{ijkl} - \bar{y}_{....})^2 = nbc \sum_{i}^{a} (\bar{y}_{i...} - \bar{y}_{....})^2$$

$$+ \; nac \sum_{j}^{b} (\bar{y}_{.j..} - \bar{y}_{....})^2$$

$$+ \; nab \sum_{k}^{c} (\bar{y}_{..k.} - \bar{y}_{....})^2$$

$$+ \; nc \sum_{i}^{a}\sum_{j}^{b} (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....})^2$$

$$+ \; nb \sum_{i}^{a}\sum_{k}^{c} (\bar{y}_{i.k.} - \bar{y}_{i...} - \bar{y}_{..k.} + \bar{y}_{....})^2$$

$$+ \; na \sum_{j}^{b}\sum_{k}^{c} (\bar{y}_{.jk.} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....})^2$$

$$+\ n \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} (\bar{y}_{ijk.} - \bar{y}_{ij..} - \bar{y}_{i.k.} - \bar{y}_{.jk.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} - \bar{y}_{....})^2$$

$$+ \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} (y_{ijkl} - \bar{y}_{ijk.})^2$$

These sums of squares formulas involving means are inconvenient for calculation. It is more satisfactory to use forms of the sums of squares involving totals instead of means. By manipulation of the above sums of squares we get the following identities:

<u>Total sum of squares</u> SST with abcn-1 degrees of freedom

$$SST = \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} (y_{ijkl} - \bar{y}_{....})^2 = \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} y^2_{ijkl}$$

$$- \frac{1}{abcn} \left(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} y_{ijkl} \right)^2$$

Sum of squares for the main effect for A

$$SS = nbc \sum_{i}^{a} (\bar{y}i\ldots - \bar{y}\ldots.)^2 \quad = \quad \frac{1}{bcn} \sum_{i}^{a} \Big(\sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} yijkl \Big)^2$$

$$= \quad \frac{1}{abcn} \Big(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} yijkl \Big)^2$$

Sum of squares for the main effect for B

$$SSB = nac \sum_{j}^{b} (\bar{y}.j.. - \bar{y}\ldots.)^2 \quad = \quad \frac{1}{acn} \sum_{i}^{b} \Big(\sum_{i}^{a} \sum_{k}^{c} \sum_{l}^{n} yijkl \Big)^2$$

$$- \quad \frac{1}{abcn} \Big(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} yijkl \Big)^2$$

Sum of squares for the main effect for C

$$SSC = nab \sum_{k}^{c} (\bar{y}..k. - \bar{y}\ldots.)^2 \quad = \quad \frac{1}{abn} \sum_{k}^{c} \Big(\sum_{i}^{a} \sum_{j}^{b} \sum_{l}^{n} yijkl \Big)^2$$

$$- \quad \frac{1}{abcn} \Big(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} yijkl \Big)^2$$

Sums of squares for the interaction AB, AC, BC

$$SS(AB) = nc \sum_{i}^{a} \sum_{j}^{b} (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....})^2$$

$$= \frac{1}{cn} \sum_{i}^{a} \sum_{j}^{b} \left(\sum_{k}^{c} \sum_{l}^{n} y_{ijkl} \right)^2 - \frac{1}{abcn} \left(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} y_{ijkl} \right)^2$$

$$- nbc \sum_{i}^{a} (\bar{y}_{i...} - \bar{y}_{....})^2 - nac \sum_{j}^{b} (\bar{y}_{.j..} - \bar{y}_{....})^2$$

$$SS(AC) = nb \sum_{i}^{a} \sum_{k}^{c} (\bar{y}_{i.k.} - \bar{y}_{i...} - \bar{y}_{..k.} + \bar{y}_{....})^2$$

$$= \frac{1}{bn} \sum_{i}^{a} \sum_{k}^{c} \left(\sum_{j}^{b} \sum_{l}^{n} y_{ijkl} \right)^2 - \frac{1}{abcn} \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} y_{ijkl})^2$$

$$- nbc \sum_{i}^{a} (\bar{y}_{i...} - \bar{y}_{....})^2 - nab \sum_{k}^{c} (\bar{y}_{..k.} - \bar{y}_{....})^2$$

$$SS(BC) = na \sum_{j}^{b} \sum_{k}^{c} (\bar{y}_{.jk.} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....})^2$$

$$= \frac{1}{na} \sum_{j}^{b} \sum_{k}^{c} \left(\sum_{i}^{a} \sum_{l}^{n} y_{ijkl} \right)^2 - \frac{1}{abcn} \left(\sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} y_{ijkl} \right)^2$$

$$- nac \sum_{j}^{b} (\bar{y}_{.j..} - \bar{y}_{....})^2 - nab \sum_{k}^{c} (\bar{y}_{..k.} - \bar{y}_{....})^2$$

Error sum of squares

$$SSE = \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} (y_{ijkl} - \bar{y}_{ijk.})^2$$

$$= \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \sum_{l}^{n} y^2_{ijkl} - \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{c} \left(\sum_{l}^{n} y_{ijkl} \right)^2$$

Sum of squares for the interaction ABC

$$SS(ABC) = SST - SSA - SSB - SSC - SS(AB) - SS(AC) - SS(BC) - SSE$$

<u>Table 30</u>

Analysis of Variance Table, Three-Way Classification, Model I

Source of Variation	Sum of Square	Degrees of Freedom	Mean Square	Computed F	Tabled F
A	SSA	$n_1=a-1$	$MSA = \frac{SSA}{n_1}$	$F_1 = \frac{MSA}{MSE}$	$F_{(\alpha;\, n_1, n_8)}$
B	SSB	$n_2=b-1$	$MSA = \frac{SSB}{n_2}$	$F_2 = \frac{MSB}{MSE}$	$F_{(\alpha;\, n_2, n_8)}$
C	SSC	$n_3=c-1$	$MSC = \frac{SSC}{n_3}$	$F_3 = \frac{MSC}{MSE}$	$F_{(\alpha;\, n_3, n_8)}$
AB	SS(AB)	$n_4=n_1*n_2$	$MS(AB) = \frac{SS(AB)}{n_4}$	$F_4 = \frac{MS(AB)}{MSE}$	$F_{(\alpha;\, n_4, n_8)}$
AC	SS(AC)	$n_5=n_1*n_3$	$MS(AC) = \frac{SS(AC)}{n_5}$	$F_5 = \frac{MS(AC)}{MSE}$	$F_{(\alpha;\, n_5, n_8)}$
BC	SS(BC)	$n_6=n_2*n_3$	$MS(BC) = \frac{SS(BC)}{n_6}$	$F_6 = \frac{MS(BC)}{MSE}$	$F_{(\alpha;\, n_6, n_8)}$
ABC	SS(ABC)	$n_7=n_1*n_2*n_3$	$MS(ABC) = \frac{SS(ABC)}{n_7}$	$F_7 = \frac{MS(ABC)}{MSE}$	$F_{(\alpha;\, n_7, n_8)}$
Error	SSE	$n_8=abc(n-1)$	$MSE = \frac{SSE}{n_8}$		
Total	SST	abcn - 1			

3.8.5.4 The Tests

As in the case of the Model I of the two-way classification the tests are performed by comparing the computed F values from the last column of table (30) with the corresponding F_α (n_i ; n_8) values from the table of the F distribution, where i = 1,.......,7 (see Table D) in the Appendix.

A test is significant if the computed F is greater than the corresponding table F, otherwise the test is not significant. A significant test means that the associated null hypothesis is to be rejected. The null hypothesis is accepted when the test is not significant.

3.8.5.5 The Three-Way Classification, Model I with n=1

The analysis of variance for the unreplicated three-way classification is similar to but simpler than the analysis of variance for the replicated three-way classification, because the operations of summing over cells and calculating within cells sum of squares do not arise. The mean square for the three factor interaction AC is used to estimate the variance σ^2, since there is no error mean square MSE.

Example 3.16

Table 31 contains data representing theoretical height measurements in cm for an experiment on seedlings growth involving 3 tree varieties of jute (factor A), 2 fertilizer levels (factor B) and 2 water levels (factor C). Compute the analysis of variance table and test the following hypothesis, using a level of significance $\alpha = 0.05$.

1) H_{10} : $\alpha_1 = \alpha_2 = \alpha_3 = 0$

H_{20} : $\beta_1 = \beta_2 = 0$

3) H_{30} : $\lambda_1 = \lambda_2 = 0$

4) H_{40} : $(\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{22}$

$= (\alpha\beta)_{31} = (\alpha\beta)_{32} = 0$

5) H_{50} : $(\alpha\beta\lambda)_{111} = (\alpha\beta\lambda)_{112} = (\alpha\beta\lambda)_{121} =$

$(\alpha\beta\lambda)_{122} = (\alpha\beta\lambda)_{211} = (\alpha\beta\lambda)_{212} =$

$(\alpha\beta\lambda)_{221} = (\alpha\beta\lambda)_{222} = (\alpha\beta\lambda)_{311} =$

$(\alpha\beta\lambda)_{312} = (\alpha\beta\lambda)_{321} = (\alpha\beta\lambda)_{322} = 0$

Solution The analysis of variance

Table 31

Factor C		LEVEL 1		LEVEL 2		
Factor A	Factor B	level 1	level 2	level 1	level 2	TOTAL
level 1		8.6	2.2	4.3	6.3	
		8.8	2.5	5.9	4.5	
		5.8	1.5	6.2	5.1	
		7.9	3.6	4.8	3.8	
		31.1	9.8	21.2	19.7	81.1
		40.9		40.9		
level 2		9.4	5.7	5.4	2.2	
		8.4	7.8	7.1	3.6	
		8.9	6.9	5.6	2.9	
		8.7	7.8	6.1	4.8	
		35.4	28.2	24.2	13.5	101.3
		63.6		37.7		
level 3		5.9	10.9	9.4	4.5	
		6.5	11.0	11.2	5.2	
		7.2	11.3	10.4	6.8	
		6.5	10.4	10.3	5.9	
		26.1	43.6	41.3	22.4	133.4
		69.7		63.7		
TOTAL		92.6	81.6	86.7	55.6	316.5
		174.2		142.3		

Table 32

Observations and cell means for the Three-Way Classification with n replications per cell

C	LEVEL 1				LEVEL 2					
B	level 1		level 2		Level 1		level 2			
A	SUM	MEAN	SUM	MEAN	SUM	MEAN	SUM	MEAN	TOTAL	MEAN
1	31.1	7.775	9.8	2.450	21.2	5.300	19.7	4.925	81.8	5.113
2	35.4	8.850	28.2	7.050	24.2	6.050	13.5	3.375	101.3	6.331
3	26.1	6.526	43.6	10.900	41.3	10.325	22.4	5.600	133.4	8.338
	92.6	7.717	81.6	6.800	86.7	7.225	55.6	4.63	316.5	6.594

Table 33

Sums and Means from the cell means obtained by summing and averaging over factor C

B / A	1	2	TOTAL	MEAN
1	52.3	29.5	81.8	5.113
2	59.6	41.7	102.3	6.331
3	67.4	66.0	133.4	8.338
TOTAL	179.3	137.2	316.5	
MEAN	7.471	5.717		6.594

Table 34

Sums and Means from the cell Means obtained

by summing and averaging over factor B

C / A	LEVEL 1	LEVEL 2	TOTAL	MEAN
1	40.9	40.9	81.8	5.113
2	63.6	37.7	101.3	6.331
3	69.7	63.7	133.4	8.338
TOTAL	174.2	142.3	316.5	
MEAN	7.258	5.929		6.594

Table 35

Sums and Means from the cell Means obtained

by summing and averaging over factor A

C / B	1	2	TOTAL	MEAN
1	92.6	86.7	179.3	7.471
2	81.6	55.6	137.2	5.717
TOTAL	174.2	142.3	316.5	
MEAN	7.258	5.929		6.594

Total sum of squares SST, with (3*2*2*4)-1) degrees of Freedom

$$SST = \left(\sum_{i=1}^{3} \sum_{j=1}^{2} \sum_{k=1}^{2} \sum_{l=1}^{4} y^2ijkl - \frac{1}{48} \left(\sum_{i=1}^{3} \sum_{j=1}^{2} \sum_{k=1}^{2} \sum_{l=1}^{4} yijkl \right)^2 \right.$$

$$= 2405.11 - \frac{1}{48} (316.5)^2$$

$$= 2405.11 - 2086.922$$

$$= 318.188$$

Sum of squares for the main effect for A

$$\sum_{i=1}^{3} \left(\sum_{j=1}^{2} \sum_{k=1}^{2} \sum_{l=1}^{4} yijkl \right)^2 = (31.1+9.8+21.2+19.7)^2 + (35.4+28.2+24.2+13.5)^2 + (26.1+43.6+41.3+22.4)^2$$

$$= (81.8)^2 + (101.3)^2 + (133.4)^2$$

$$= 34748.49$$

$$SSA = \frac{1}{16} (34748.49) - \frac{1}{48} (316.5)^2$$

$$= 2171.781 - 2086.922$$

$$= 84.859$$

Sum of squares for the main effect for B

$$\sum_{j=1}^{2} \left(\sum_{i=1}^{3} \sum_{k=1}^{2} \sum_{l=1}^{4} yijkl \right)^2 = ((31.1+21.2+35.4) + (24.2+26.1+41.3))^2 + ((9.8+19.7+28.2) + (13.5+43.6+22.6))^2$$

$$= (87.7+91.6)^2 + (57.7+79.7)^2$$

$$= 179.3^2 + 137.2^2$$

$$= 50972.33$$

$$SSB = \frac{1}{24}(50972.33) - \frac{1}{48}(316.5)^2$$

$$= 2123.847 - 2086.922$$

$$= 36.925$$

Sum of squares for the main effects for C

$$\sum_{k=1}^{2} \left(\sum_{i=1}^{3} \sum_{j=1}^{2} \sum_{l=1}^{4} y_{ijkl} \right)^2 = (31.1+9.8+35.4+28.2+26.1+43.6)^2 + (21.2+19.7+24.2+13.5+41.3+22.4)^2$$

$$= (174.2)^2 + (142.3)^2$$

$$= 30345.64 + 20249.29$$

$$= 50594.93$$

$$SSC = \frac{1}{24}(50594.93) - \frac{1}{48}(316.5)^2$$

$$= 2108.122 - 2086.922$$

$$= 21.200$$

Sums squares for the interactions AB, AC, BC

$$\sum_{i=1}^{3} \sum_{j=1}^{2} \left(\sum_{k=1}^{2} \sum_{l=1}^{4} y_{ijkl} \right)^2 = 52.3^2 + 29.5^2 + 59.6^2 + 41.7^2 + 67.4^2 + 66.0^2$$

$$= 17795.35$$

$$\sum_{i=1}^{3} (\bar{y}_{i\ldots} - \bar{y}_{\ldots\ldots})^2 = (5.113 - 6.594)^2 + (6.331 - 6.594)^2 + (8.338 - 6.594)^2$$

$$= 2.193 + 0.069 + 3.042$$

$$= 5.304$$

$$\sum_{j=1}^{2} (\bar{y}_{.j..} - \bar{y}_{....})^2 = (7.471 - 6.594)^2 + (5.717 - 6.594)^2$$

$$= 0.769 + 0.769$$

$$= 1.538$$

$$SS(AB) = \frac{1}{48}(17795.35) - \frac{1}{48}(316.5)^2 - 16(5.304) - 24(1.538)$$

$$= 2224.419 - 2086.922 - 84.864 - 36.912$$

$$= 15.721$$

$$\sum_{k=1}^{3} \sum_{i=1}^{2} (\sum_{j=1}^{2} \sum_{l=1}^{4} yijkl)^2 = 40.9^2 + 40.9^2 + 63.6^2 + 37.7^2 + 69.7^2 + 63.7^2$$

$$= 17727.65$$

$$\sum_{k=1}^{2} (\bar{y}_{..k.} - \bar{y}_{....})^2 = (7.258 - 6.594)^2 + (5.929 - 6.594)^2$$

$$= 0.441 + 0.442$$

$$= 0.883$$

$$SS(AC) = \frac{1}{8}(17727.65) - \frac{1}{48}(316.5)^2 - 16(5.304) - 24(0.883)$$

$$= 2215.956 - 2086.922 - 84.864 - 21.192$$

$$= 22.978$$

$$\sum_{j=1}^{2}\sum_{k=1}^{2}\left(\sum_{i=1}^{3}\sum_{l=1}^{4} y_{ijkl}\right)^2 = 92.6^2 + 86.7^2 + 81.6^2 + 55.6^2$$

$$= 25841.57$$

$$SS(BC) = \frac{1}{12}(25841.57) - \frac{1}{48}(316.5)^2 - 24(1.538) - 24(0.883)$$

$$= 2153.464 - 2086.922 - 36.912 - 21.192$$

$$= 8.438$$

Error sum of squares

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{2}\left(\sum_{l=1}^{4} y_{ijkl}\right)^2 = 31.1^2 + 9.8^2 + 21.2^2 + 19.7^2$$

$$+ 35.4^2 + 28.2^2 + 24.2^2 + 13.5^2$$

$$+ 26.1^2 + 43.6^2 + 41.3^2 + 22.4^2$$

$$= 9506.69$$

$$SSE = 2405.110 - \frac{1}{4}(9506.69)$$

$$= 2405.110 - 2376.673$$

$$= 28.437$$

Sum of squares for the interaction (ABC)

$$SS(ABC) = 318.188 - 84.859 - 36.925 - 21.200 - 15.721 - 22.978 - 8.438 - 28.437$$

$$= 99.63$$

Table 36

Analysis of Variance Table Three-Way Classification, Model I

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Square	Computed F	Table F
A	84.859	$n_1 = 2$	42.430	$F_1 = 53.709$	3.26
B	36.925	$n_2 = 1$	36.925	$F_2 = 46.741$	4.11
C	21.200	$n_3 = 1$	21.200	$F_3 = 26.835$	4.11
AB	15.721	$n_4 = 2$	7.861	$F_4 = 9.951$	3.26
AC	22.978	$n_5 = 2$	11.489	$F_5 = 14.543$	3.26
BC	8.438	$n_6 = 1$	8.438	$F_6 = 10.681$	4.11
ABC	99.630	$n_7 = 2$	49.815	$F_7 = 63.057$	3.26
Error	28.437	$n_8 = 36$	0.790		
Total	318.188	47			

Conclusions

(1) Since $F_1 = 53.709$ is greater than the table value $F_{0.05}^{(2,36)} = 3.26$, we reject the null hypothesis H_{10} and conclude that a difference in the height growth of the seedlings of the 3 varieties of jute exists.

(2) Since $F_2 = 46.741$ is greater than $F_{0.05}^{(1,36)} = 4.11$, we reject the null hypothesis H_{20} and conclude that a difference in the height growth of the seedlings of the varieties of jute due to fertilizer exists.

(3) Since $F_3 = 26.835$ is greater than $F_{0.05}^{(1,36)} = 4.11$, we reject the null hypothesis H_{30} and conclude that there is a difference in the height growth of the seedlings of the 3 varieties of jute due to the different water levels.

(4) Since $F_4 = 9.951$ is greater than $F_{0.05}^{(2,36)} = 3.26$, we reject the null hypothesis H_{40} and conclude that there is an interaction between jute variety and fertilizer on the height growth of the seedlings.

(5) Since $F_7 = 63.057$ $F_{0.05}^{(2,36)}$, we reject the null hypothesis H_{50} and conclude that a significant interact- ion exists between jute variety, fertilizer and water level on the height growth of the seedlings.

Exercise 3.10

3 fertilizers were used in a field experiment which involves 3 varieties of corn, with and without irrigation. The 18 treatment combinations were assigned at random to 18 plots. The hypothetical yields per plot were as follows:

Table 36

Corn yields from three varieties

	no irrigation			irrigation		
Fertilizer / Variety	1	2	3	1	2	3
1	77	80	92	80	81	95
2	89	88	95	91	81	110
3	65	82	89	76	79	92

Analyse the experimental data through the appropriate analysis of variance and null hypothesis tests.

3.9 MULTIPLE LINEAR REGRESSION

3.9.1 The Problem

Multiple regression is a statistical procedure frequently used for the prediction of a dependent or response variable, let us say y, using its relationship to a set of p independent variables, which we denote x_1, x_2, ...,x_p. By independent variables we mean variables that can either be set to desired value or else take values that can be observed or computed from other observed values. For instance, in studying the growth of jackfruit trees we may be interested in predicting the variable tree height from the knowledge of the variable diameter at breast height (DBH). Our intuition or results from previous studies might suggest that there exists a well established relationship between the height and the square of the DBH in addition to the dependence of the height on the DBH. Therefore we would consider the height as dependent variable, the DBH as the first independent variable and the squares of the DBH as the second independent variable. This is a case in which values of an independent variable are computed from values of another independent variable (x_1 = DBH). Now to describe the relationship between dependent and independent variables, a regression model, that expresses the type of the assumed relationship, is used. In the following we will be concerned only with multiple linear regression models, because in many situations a linear relationship can be valuable in summarizing the presumed or observed dependence of one variable on one or several other variables.

3.9.2 The General Regression Model

Given n sets of observations or values on the variables y, x_1, x_2,.......,x_p, the general linear regression model is of the form

$$y_1 = \beta_0 x_{01} + \beta_1 x_{11} + \dots\dots\dots + \beta_p x_{p1} + \varepsilon_1$$

$$y_2 = \beta_0 x_{02} + \beta_1 x_{12} + \dots\dots\dots + \beta_p x_{p2} + \varepsilon_2$$

....

$$y_n = \beta_0 x_{0n} + \beta_1 x_{1n} + \dots\dots\dots + \beta_p x_{pn} + \varepsilon_n$$

or

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \dots\dots\dots + \beta_p x_{pi} + \varepsilon_i$$

Where $i = 1,\dots\dots\dots,n$,

and $x_{0i} = 1$ for all i

or using summation notation

$$y_i = \sum_{j=0}^{p} \beta_j x_{ji} + \varepsilon_i$$

$\beta_0, \beta_1, \dots\dots\dots, \beta_p$ are the unknown regression coefficients

and

$\varepsilon_1, \varepsilon_2, \dots\dots\dots\dots, \varepsilon_n$ are independent normally distributed error variables with mean 0 and variance σ^2.

As already mentioned the independent variables may be functions of other variables or of each other. In particular

if $x_{ji} = x_{ji}^{j}$, then we have the polynomial regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots\dots + \beta_p p_{pi}$$

The term <u>linear</u> refers to linearity in the parameters and not in the independent variables, thus the polynomial regression model is a <u>linear</u> regression model.

3.9.3 The Computational Procedure

As in the case of the simple linear regression the main task in multiple regression analysis is to estimate the regression coefficients $\beta_0, \beta_1, \dots\dots\dots, \beta_p$.

The method of least squares (see Appendix) is used to obtain those estimates which minimize the sums of squares of

deviations $\sum \varepsilon_i^2$ denoted by S, where

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_p x_{pi})^2$$

The estimates are called $b_0, b_1, \dots, b_p$. In addition to these estimates, other quantities necessary to set up the analysis of variance table, test the goodness of the regression equation and construct confidence intervals for the estimated parameters are usually computed.

It is more convenient to present a multiple regression problem and the computational steps necessary to solve it in terms of matrix algebra. The use of matrices has many advantages, one of these relating to the use of computers is that the solution of the problem expressed in matrix terms can be generalized and applied to any regression problem, no matter how many terms there are in the regression equation.

Using matrix representation with capital letters denoting matrices we can write the regression model as

$$Y = XB + \varepsilon$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$ is an (n*1) vector of observations

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix}$$

is an n* (p+1) matrix of known elements

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

is a (p+1) * 1 vector of (coefficients) parameters

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

is an (n*1) vector of errors having a (n*1) zero vector as mean and $I\sigma^2$ as variance, where I denotes the n*n identity matrix

Now to perform the regression analysis using available data we execute the following steps:

1) *Setting up the vector Y and the matrix X*

2) *Evaluate sums and means of the elements of the vector Y and the matrix X*

3) *Estimate the regression coefficients that are the elements of a vector B*

(a) *Compute X'X*

(b) *Compute* $(X'X)^{-1}$ *that is the inverse of (X'X). For some elements of matrix algebra and matrix inversion, see Manual 1 on Basic Computer Mathematics*

(c) *Compute the elements of vector B using the formula*

$$B = \begin{pmatrix} b_0 \\ \\ b_1 \\ . \\ . \\ . \\ . \\ b_p \end{pmatrix} = (X'X)^{-1} X'Y$$

4) Compute the corrected total sum of squares SST

$$SST = Y'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

5) Compute the corrected sum of squares due to regression SSR

$$SSR = B'X'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

6) Compute the error or residual sum of squares SSE

$$SSE = SST - SSR$$

7) Setting up the analysis of variance table

Table 38

Source of variation	Sum of Squares	Degrees of freedom	Mean square
Regression	SSR	p	$MSR = \frac{SSR}{p}$
Error	SSE	n-(p+1)	$MSE = \frac{SSE}{n-(p+1)}$
Total	SST	n-1	

8) Compute the coefficient of determination R^2

$$R^2 = \frac{SSR}{SST}$$

9) Compute the variance-covariance matrix V of the vector B

as $V(B) = (X'X)^{-1}s^2$

where s^2 denotes the residual mean square MSE contained in the analysis of variance table. The variances of the elements of the vector B are the diagonal terms of the variance covariance matrix V.

10) Write the regression equation

$y = b_0 + b_1x + \ldots\ldots + b_px_p$

11) Use the regression equation to predict the response value $\hat{y}_r$ for the vector of mean values $\bar{X} = (1\ \bar{x}_1\ \ldots\ \bar{x}_p)'$ of the elements of matrix X

$$\hat{y}_r = \bar{y} = \bar{X}'B = \bar{B}'X$$

12) Evaluate the variance of the predicted value $\hat{y}_r$

as $V(\hat{y}_r) = (\bar{X}'(X'X)^{-1}\bar{X})_s{}^2$

13) Check the goodness of the regression equation.

3.9.4 <u>Testing the overall regression equation</u>

To test the null hypothesis

$H_0 : \beta_1 = \ldots\ldots = \beta_p = 0$

against the alternative

$H_1 :$ not all $\beta_i = 0$

(a) evaluate the mean squares ratio

$$F = \frac{MSR}{MSE}$$

(b) compare the evaluated F with F_α (p,n-(p+1)) from the table of the F distribution (see F Table in the Appendix). If the mean square ratio is significant, that is, if it exceeds the table value, reject the null hypothesis of all the coefficients β_i being equal to 0.

A significant test merely means that the proportion of the variation observed in the data, which has been accounted for by the equation, is greater than would be expected by chance in a proportion of 100(1-)% of similar samples with the same size n and the same values for the matrix X. A significant test does not necessarily mean that the fitted equation is useful for predictive purposes. Unless the computed F exceeds at least about four time the selected percentage point F (n,n-(p+1)), prediction will often be of no value even though a significant F value has been obtained.

3.9.5 The Examination of residuals

It should be stressed that a prequisite for using the F-test to check the goodness of a regression equation is that the errors are normally distributed. The assumption of normal distribution of the errors can be tested by examining the residuals, using a procedure similar to the normality test, which is described in page 18. The residuals are defined as the differences $e_i = y_i - y_i$ $(i= 1,.......,n)$, where y_i is an observation and y_i is the corresponding regression value obtained by means of the fitted regression equation.

The procedure for testing normality of residuals is as follows:

- compute the values $\frac{e_i}{MSE}$, $(i = 1,....., n)$,

  called the unit normal deviates of the residuals e_i

- evaluate the percentage of unit normal deviates that lie between the limits (-2,+2). If the percentage is equal to or greater than 95%, the hypothesis of normality is accepted, otherwise the hypothesis is rejected.

Example 3.17

Perform the regression analysis for the following data

Y	x_1	x_2
66.0	38	47.5
43.0	41	21.3
36.0	34	36.5
23.0	35	18.0
22.0	31	29.5
14.0	34	14.2
12.0	29	21.0
7.6	32	10.0

Solution

(1) The vector Y and the matrix X are

$$Y_{(8,1)} = \begin{pmatrix} 66.0 \\ 43.0 \\ 36.0 \\ 23.0 \\ 22.0 \\ 14.0 \\ 12.0 \\ 7.6 \end{pmatrix} \qquad X_{(8,3)} = \begin{pmatrix} 1 & 38 & 47.5 \\ 1 & 42 & 21.3 \\ 1 & 34 & 36.5 \\ 1 & 35 & 18.0 \\ 1 & 31 & 29.5 \\ 1 & 34 & 14.2 \\ 1 & 29 & 21.0 \\ 1 & 32 & 10.0 \end{pmatrix}$$

(2) Sums / Means

$$\sum_{i=1}^{8} y_i = 223.6 \qquad \bar{y} = 27.95$$

$$\sum_{i=1}^{8} x_{1i} = 8 \qquad \bar{x}_0 = \frac{\sum_{i=1}^{8} x_{1i}}{8} = 1$$

$$\sum_{i=1}^{8} x_{2i} = 274 \qquad \bar{x}_1 = \frac{\sum_{i=1}^{8} x_{2i}}{8} = 34.25$$

$$\sum_{i=1}^{8} x_{3i} = 198 \qquad \bar{x}_2 = \frac{\sum_{i=1}^{8} x_{3i}}{8} = 24.75$$

(3)

(a

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 38 & 41 & 34 & 35 & 31 & 34 & 29 & 32 \\ 47.5 & 21.3 & 36.5 & 18 & 29.5 & 14.2 & 21 & 10 \end{pmatrix} * \begin{pmatrix} 1 & 38 & 47.5 \\ 1 & 41 & 21.3 \\ 1 & 34 & 36.5 \\ 1 & 35 & 18.0 \\ 1 & 31 & 29.5 \\ 1 & 34 & 14.2 \\ 1 & 29 & 21.0 \\ 1 & 32 & 10.0 \end{pmatrix}$$

$$= \begin{pmatrix} 8 & 274 & 198 \\ 274 & 9488 & 6875.6 \\ 198 & 6875.6 & 5979.08 \end{pmatrix}$$

(b We see that X'X is a symmetric matrix. The inverse $(X'X)^{-1}$, that is also a symmetric matrix, is evaluated by using the following procedure:

$$\text{If } X'X = \begin{pmatrix} a & b & c \\ b & e & f \\ c & f & j \end{pmatrix}$$

Then

$$(X'X)^{-1} = \begin{pmatrix} A & B & C \\ B & E & F \\ C & F & J \end{pmatrix}$$

where

$$A = (ej - f^2)/W \qquad B = -(bj - cf)/W$$

$$C = (bf - ce)/W \qquad E = (aj - c^2)/W$$

$$F = -(af - bc)/W \qquad J = (ae - b^2)/W$$

and

$$W = a(ej - f^2) - b(dj - cf) + c(bf - ce)$$

$$= aej + 2bcf - af^2 - b^2j - c^2e$$

Thus

$$W = (8*9488*5979.08) + (2*274*198*6875.6)$$

$$- (8*(6875.6)^2) - ((274)^2*5979.08) -$$

$$((198)^2*9488))$$

$$= 828232$$

$$A = (9488*5979.8) - (6857.6)^2 / 828232$$

$$= 11$$

$B = -(274*5979.08) - (198*6875.6) / 828232$

$= -0.33677$

$C = -(274*6875.6) - (198*9488) / 828232$

$= 0.00643$

$E = (8*5979.08) - (198)^2 / 828232$

$= 0.01049$

$F = -(8*6875.6) - (274*198) / 828232$

$= -0.00092$

$J = -(8*9488) - (274)^2 / 828232$

$= 0.00101$

and the inverse matrix is

$$(X'X)^{-1} = \begin{pmatrix} 11.50005 & -0.33677 & 0.00643 \\ -0.33677 & 0.01049 & -0.00092 \\ 0.00643 & -0.00092 & 0.00101 \end{pmatrix}$$

c)

$$B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = (X'X)^{-1}X'Y$$

$$= \begin{pmatrix} 11.50005 & -0.33677 & 0.00643 \\ -0.33677 & 0.01049 & -0.00092 \\ 0.00643 & -0.00092 & 0.00101 \end{pmatrix} * \begin{pmatrix} 223.6 \\ 8049.2 \\ 6954.7 \end{pmatrix}$$

$$= \begin{pmatrix} -94.55203 \\ 2.80155 \\ 1.07268 \end{pmatrix}$$

(4)

$$SST = Y'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$= 8911.76 - 6249.62 = 2662.14$$

(5)

$$SSR = B'X'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$= \begin{pmatrix} -94.55203 & 2.80155 & 1.07268 \end{pmatrix} * \begin{pmatrix} 223.6 \\ 8049.0 \\ 6954.7 \end{pmatrix} * -6249.62$$

$$= 2618.97937$$

(6)

$$SSE = SST - SSR$$

$$= 2662.14 - 2618.97937$$

$$= 43.16063$$

(7) Table 29 Analysis of variance table

Source of variation	Sum of Squares	DF	Mean square	F
Regression	2618.97935	2	1309.48968	151.6995
Residual	43.16065	5	8.6321	
Total	2662.14	7		

(8)

The coefficient of determination is:

$$R^2 = \frac{SSR}{SST}$$

$$= \frac{2618.97935}{2662.14}$$

$$= 0.98379$$

(9)

The variance-covariance matrix is:

$$V(B) = \begin{pmatrix} 11.50005 & -0.33677 & 0.00643 \\ -0.33677 & 0.01049 & -0.00092 \\ 0.00643 & -0.00092 & 0.00101 \end{pmatrix} * 8.63213$$

$$\begin{pmatrix} \underline{99.26988} & -2.90704 & 0.0555 \\ -2.90704 & \underline{0.09055} & -0.00794 \\ 0.0555 & -0.00794 & \underline{0.00872} \end{pmatrix}$$

The variance of b_0 is 99.26988

" " " b_1 " 0.09055

" " " b_2 " 0.00872

(10)

The regression equation is

$$y = -94.55203 + 2.80155x_1 + 1.07268x_2$$

(11)

$$\hat{y}_r = \bar{X}'B = \begin{bmatrix} 1 & 34.25 & 24.75 \end{bmatrix} * \begin{pmatrix} -94.55203 \\ 2.80155 \\ 1.07268 \end{pmatrix}$$

$$= 27.95$$

Note that $\hat{y}_r = \bar{y}$

(12)

The variance of $\hat{y}_r$ or $\bar{y}$ is

$$V(\hat{y}_r) = (\bar{X}'(X'X)^{-1}\bar{X})s^2$$

$$= \begin{bmatrix} 1 & 34.25 & 24.75 \end{bmatrix} * \begin{pmatrix} 11.50005 & -0.33677 & 0.00643 \\ -0.33677 & 0.01049 & -0.00092 \\ 0.00643 & -0.00092 & 0.00101 \end{pmatrix}$$

$$* \begin{pmatrix} 1 \\ 34.25 \\ 24.75 \end{pmatrix} * 8.63213$$

$$= 0.9837$$

(13)

The high value for the coefficient of determination $R^2 = 0.98$ and the large F value $F = 151.7 > 4*F_{(.05,2,5)} = 5.79$ indicates that the regression fits the data well.

__N O T E__

Example 3.17 has been computed by means of a pocket calculator and the results are written, using only 5 s ignificant digits. We should therefore expect these results to be less accurate than the ones we would get from the computer. For this reason, one must be careful to carry along as many digits as possible in executing the sequences of operations required in the multiple regression procedure, in order to reduce the magnitude of round-off errors to a minimum.

In certain studies confidence limits for the estimated coefficients and the fitted regression line may be required. Then the results in 9), 10), and 12) are to be used in establishing these confidence intervals.

Exercise 3.11

Use the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

to fit a regression equation to the following wildlife data that represents body proportion measurements in S. lenoensis.

y (Std. length)	x_1 (Body depth)	x_2 (Head length)
18.2	3.1	3.1
27.8	3.6	6.1
27.4	5.8	5.8
19.4	3.5	3.5
29.0	4.8	5.5
27.8	4.7	6.0
26.4	4.1	4.5
28.3	4.8	6.0
20.3	3.1	4.0
26.4	4.4	6.5
26.7	4.9	5.2
22.2	3.4	4.3
19.6	2.9	4.3
18.0	2.7	2.9
27.8	4.9	5.9
18.6	3.3	3.3
26.0	4.9	5.3
19.0	2.5	3.2
25.9	4.5	4.5
28.0	4.1	4.7

THE METHOD OF LEAST-SQUARES

Analysis of variance and regression analysis belong to a class of statistical procedures that involve the construction of some mathematical model to describe the problem to be investigated. In the case of linear regression analysis this mathematical model is used to relate the value of a dependent variable y to one or more independent variables x_j, j=1,.......,p When p=1, we have the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

where y_i represents an observed value for the dependent variable y at a value x_i of the independent variable x, β_0 and β_1 are the parameter coefficients of the function, ε_i is the error term associated with that observation y_i.

To be able to evaluate this function, we need the values of the coefficients β_0 and β_1. But since these are unknown population parameters, we have to determine or estimate them. To do this, sample data is needed for y and x. Suppose that we have collected a sample of n pairs of values (y_1,x_1), (y_2,x_2),........, (y_n,x_n) on y and x.

We can use this data to determine values for the estimates, let us say b_0 and b_1 of the parameters $\beta 0$ and $\beta 1$ respectively. Now, if we use a scatter diagram to represent graphically the n pairs of numbers, as shown in fig. , we easily see that there are many lines that can be fitted to the data points, some of these being better than others. Such a line - an intuitively "good" one - could be the line represented on fig. 11.

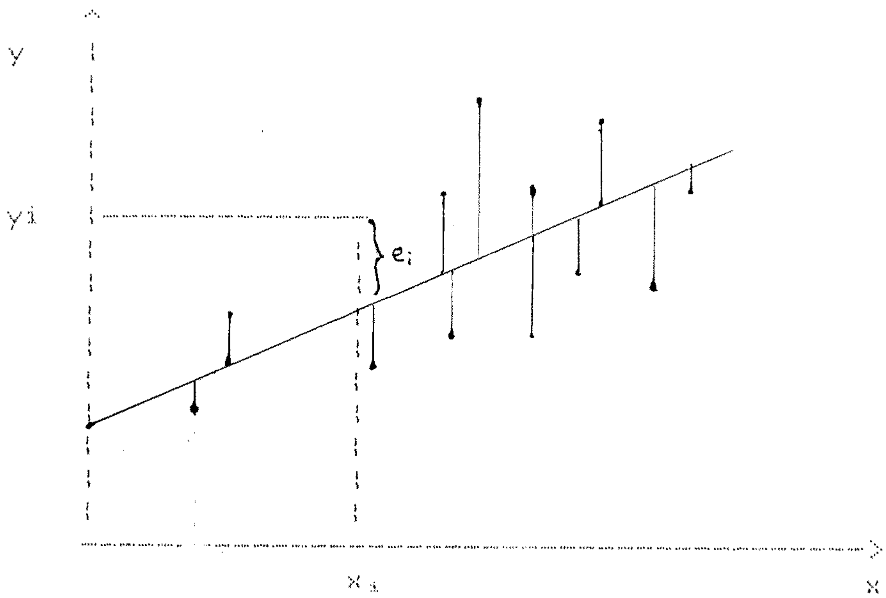


Fig. 11 Graphical representation of a least square line.

But in order to single our one line which provides the "best" fit to the data points, we have to define a criterion on the basis of which the best line can be determined; this is equivalent to finding the best values b_0 and b_1 for the parameters β_0 and β_1. The criterion that is commonly used to define the best fit is known as the <u>method of least-squares</u>. The least-squares method requires that the line is fitted to the data so that the sum of squares of the vertical distances from the points to the line, represented by the solid line segments on Fig. 11 is a minimum.

These vertical distances that are the differences between the observed values and the predicted values of y are usually called <u>errors or deviations</u>, and denoted by e_i, $i=1,\ldots\ldots,n$. The e_i's are the estimates of the true errors ε_i. if the point y_i lies above the line, e_i is positive; if y_i lies below the line, e_i is negative and if y_i lies on the line $e_i=0$. The least-squares problem is to minimize

$$s = \sum_{i=1}^{n} e_i^2$$

where s denotes the error sum of squares of the fitted line.

Note that the sum of the deviations themselves is not to be minimized, this is because $\sum_{i=1}^{n} e_i$ could be 0 even though all the points were far away from the line or the errors are numerically very large.

Now the sum of squares of deviations from the true line, denoted by S, is

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

by considering the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The least-squares estimates b_0 and b_1 are obtained by differentiating the equation $S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$, with respect to β_0 and then with respect to β_1, and setting the results equal to 0. Using the rules of partial differentiation, we have:

$$\frac{d\ S}{d\ \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{d\ S}{d\ \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Setting the results of the differentiation equal to zero and replacing β_0 by b_0 and β_1 by b_1, we obtain

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

From these equations we derive the following ones that are called the <u>normal equations</u>

$$b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

The solution of the normal equations for b_1 and b_0 are

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) / n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

We have already met these 2 formulas in the simple linear regression procedure (see page 32).

The quantities $\sum x_i^2$ are called the *uncorrected sum of squares* of the x's, and $(\sum x_i)^2 / n$ is called *the correction for the mean* of the x's.

The difference $\sum x_i^2 - (\sum x_i)^2 / n$ is called the *corrected sum of squares* of the x's. Similarly $\sum x_i y_i$ is called the uncorrected sum of products, and $(\sum x_i) \; (\sum y_i / n$ is the *correction for the mean*. The difference $\sum x_i y_i - (\sum x_i) (\sum y_i)$ is called the *corrected sum of products*.

The generalization of the method of least-squares to estimate multiple linear coefficients $\beta_0, \beta_1, \ldots\ldots, \beta_p$ is performed by using matrix algebra. In matrix terms the normal equations are given by

$$X'XB = X'Y$$

of which the solution for the vector B of coefficients is

$$B = (X'X)^{-1}X'Y$$

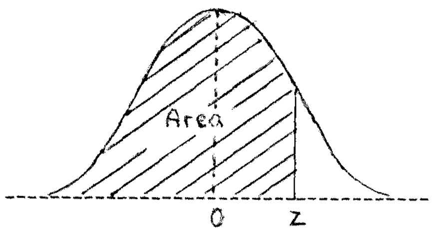
The least formula is used in the computation procedure for the multiple regression (see Section 3.9).

APPENDIX

STATISTICAL TABLES

Table A. Areas under the Normal Curve

z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0020
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559



(1) Reproduced from R.E. Walpole: Introduction to Statistics, 2nd Edition

STATISTICAL TABLES

z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0868	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3400	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
-0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
-0.1	0.5398	0.5438	0.5478	0.5517	0.5558	0.5596	0.5636	0.5675	0.5714	0.5753
-0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
-0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
-0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
-0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
-0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
-0.7	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
-0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
-0.9	0.8259	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

(1) Reproduced from R.E. Walpole: Introduction to Statistics, 2nd Edition

STATISTICAL TABLES

z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9633
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9950	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.7793	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9990	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	029996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

(1) Reproduced from R.E. Walpole: Introduction to Statistics, 2nd Edition

STATISTICAL TABLES

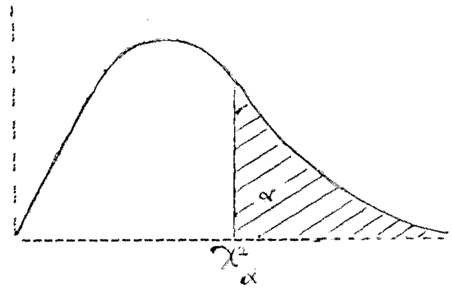


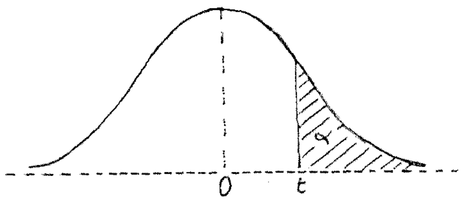
Table B. Critical Values of the Chi-Squares Distribution

d.f	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.04393	0.03157	0.03982	0.02893	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.736	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	8.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	7.564	8.672	27.587	30.191	33.409	35.718
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.410
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.795	50.892	53.672

(1) Reproduced from R.E. Walpole: Introduction to Statistics, 2nd Edition

STATISTICAL TABLES

Table C. Critical values of the t Distribution

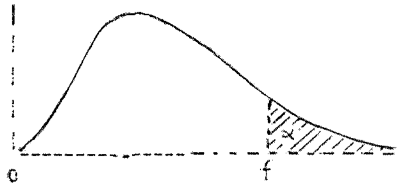


d.f	a: 0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.571	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.813
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.576
inf.					

(1) Reproduced from R.E. Walpole: Introduction to Statistics, 2nd Edition

STATISTICAL TABLES

Table D Critical Values of the Distribution $F_{0.05(n_1,n_2)}$



n_2	n_1 = 1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.61	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.55	2.46	2.40
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

(1) Reproduced from R.E. Walpole: Introduction to Statistics, 2nd Edition

STATISTICAL TABLES

Table D [1] Critical Values of the F Distribution (continued)

n_2	n_1 = 10	12	15	20	24	30	40	60	120	
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.79	8.74	8.70	8.65	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.30	3.27	3.23
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.13	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.74
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.69	1.53	1.47	1.39
120	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

(1) Reproduced from R.E. Walpole: Introduction to Statistics, 2nd Edition

STATISTICAL TABLES

Table E

Least Significant Studentized Ranges r_p

= 0.05

	p								
v	2	3	4	5	6	7	8	9	10
1	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97
2	6.085	6.085							
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.927	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.749	3.749	3.814	3.814	3.814	3.814	3.814
6	3.461	3.587	3.649	3.680	3.694	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.626	3.626	3.626
8	3.261	3.399	3.475	3.521	3.549	3.566	3.575	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522
11	3.113	3.256	3.342	3.397	3.435	3.462	3.480	3.493	3.501
12	3.082	3.225	3.313	3.370	3.410	3.439	3.459	3.474	3.484
13	3.055	3.200	3.289	3.384	3.389	3.419	3.442	3.458	3.470
14	3.033	3.178	3.268	3.329	3.372	3.403	3.426	3.444	3.457
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.412	3.446
16	2.998	3.144	3.235	3.298	3.343	3.376	3.402	3.422	3.437
17	2.984	3.130	3.222	3.285	3.331	3.366	3.392	3.412	3.429
18	2.971	3.118	3.210	3.274	3.321	3.356	3.383	3.405	3.421
19	2.960	3.107	3.199	3.264	3.311	3.347	3.375	3.397	3.415
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.391	3.409
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371
40	2.858	3.006	3.102	3.171	3.224	3.266	3.300	3.328	3.352
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.287	3.314
oo	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294

(1) Reproduced from R.E. Walpole: Introduction to Statistics, 2nd Edition

REFERENCES

1. Brownlee, K.A., — Statistical Theory and Methodology in Science and Engineering, 2nd Edition, John Wiley and Sons, New york, 1965

2. Draper, N. and Smith, H. — Applied Regression Analysis, 2nd Edition, John Wiley and Sons, New york, 1981

3. Dunn, O.J. and Clark, V.A., — Applied Statistics: Analysis of Variance and Regression, John Wiley and Sons, New York, 1974

4. Freund, J. E., — Modern Elementary Statistics, 4th Edition, Prentice/Hall International, London, 1974

5. Hoel, P.G., — Elementary Statistics, 2nd Edition, John Wiley and Sons, New-York, 1968

6. Pearson, E.S. and Hartley H.O., — Biometrika Tables for Statisticians, Volume 1 and 2, Cambridge University Press, 1972

7. Spiegel, R.M. — Theory and Problems of Statistics Asian Student Edition, Schaum's Outline Series, 198

8. Walpole, R.E., — Introduction to Statistics, 2nd Edition, Macmillan Publishing Co., New york, London, 1974.