

# **ELEMENTARY STATISTICAL METHODS FOR FORESTERS**

**Frank Freese**

**Statistician**

**Forest Products Laboratory**

**(Maintained by the Forest Service at Madison, Wis., in cooperation  
with the University of Wisconsin)**

**AGRICULTURE HANDBOOK 317**

**U.S. Department of Agriculture**

**Forest Service**

**January 1967**

---

For sale by the Superintendent of Documents, U.S. Government Printing Office  
Washington, D.C. 20402 — Price 35 cents

## ACKNOWLEDGMENTS

Professor George W. Snedecor and the Iowa State University Press have generously allowed me to republish material from *Statistical Methods*, fifth edition, in tables 1, 3-7 of this handbook. The editor and trustees of *Biometrika* concurred in the reprinting of table 4. I wish also to thank Dr. C. I. Bliss, of the Connecticut Agricultural Experiment Station, who originally prepared the data in table 6.

I am indebted to the literary executor of the late Sir Ronald A. Fisher, F.R.S., Cambridge, to Dr. Frank Yates, F.R.S., Rothamsted, and to Oliver and Boyd, Ltd., Edinburgh, for their permission to reprint present table 2 from *Statistical Tables for Biological, Agricultural and Medical Research*; and portions of present tables 3 and 7 from *Statistical Methods for Research Workers*.

Thanks are also due to those who reviewed the manuscript and contributed to it through their suggestions, particularly Thomas Evans, Virginia Polytechnic Institute; Kenneth Ware, Iowa State University; and Donald Kulow, West Virginia University.

Frank Freese  
Forest Products Laboratory

## PREFACE

This handbook was written under the assumption that forest researchers want and should have a working knowledge of the simpler statistical methods, and that most of them lack the time to extract this information from the comprehensive texts. It defines some basic terms and shows the computational routine for the statistical methods that have been found most useful in forestry. The meaning of various statistical quantities is discussed to a very limited degree. The general approach is based on the observation that most researchers have difficulty in learning the meanings and derivations of statistics (and have some reluctance to do so) until they have mastered the computational details.

The purpose, then, is to give the reader a handy reference for useful basic techniques and also to convince him that statistical methods can be learned. Having absorbed this minimal dose without great pain, he may be inclined to make a more thorough study of the subject as presented in the standard textbooks.

This handbook is an extensive revision and expansion of *Guidebook for Statistical Transients*, an informal release by the same author first issued in 1956 by the Southern Forest Experiment Station and reissued in 1963. The revision was completed after the author's assignment to the Forest Products Laboratory.

A complementary publication, Agriculture Handbook 232, *Elementary Forest Sampling*, by the same author, covers sampling methods and procedures in detail.

# CONTENTS

Preface.....	Page iii
General concepts.....	1
Statistics—what for?.....	1
Probability and statistics.....	2
Some basic terms and calculations.....	3
The mean.....	3
Standard deviation.....	4
Coefficient of variation.....	5
Standard error of the mean.....	5
Covariance.....	6
Simple correlation coefficient.....	7
Variance of a linear function.....	9
Sampling—measurement variables.....	10
Simple random sampling.....	10
Standard errors.....	11
Confidence limits.....	11
Sample size.....	12
Stratified random sampling.....	14
Sample allocations.....	15
Sample size.....	16
Sampling—discrete variables.....	16
Random sampling.....	16
Sample size.....	17
Cluster sampling for attributes.....	18
Transformations.....	19
Chi-square tests.....	20
Test of independence.....	22
Test of a hypothesized count.....	22
Bartlett's test of homogeneity of variance.....	24
Comparing two groups by the t test.....	24
The t test for unpaired plots.....	24
Sample size.....	26
The t test for paired plots.....	27
Number of replicates.....	28
Comparison of two or more groups by analysis of variance.....	28
Complete randomization.....	31
Multiple comparisons.....	31
F test with single degree of freedom.....	31
Scheffé's test.....	32
Unequal replication.....	33
Randomized block design.....	34
Latin square design.....	37
Factorial experiments.....	39
The split plot design.....	45
Missing plots.....	50
Regression.....	51
Simple linear regression.....	51
How well does the regression line fit the data?.....	54
Coefficient of determination.....	55
Confidence intervals.....	55
Multiple regression.....	57
Tests of significance.....	61
Coefficient of multiple determination.....	63
The c-multipliers.....	63
Curvilinear regressions and interactions.....	66
Group regressions.....	68
Analysis of covariance in a randomized block design.....	70
References for further reading.....	75
<b>APPENDIX</b>	
Tables.....	Page 76
1. Ratio of standard deviation to range for simple random samples of size $n$ from normal populations.....	76
2. Distribution of $t$ .....	77
3. Distribution of $F$ .....	78
4. Accumulative distribution of chi-square.....	82
5. Confidence intervals for binomial distribution.....	83
6. Arc sine transformation.....	85
7. Significance of correlation coefficients.....	87

## GENERAL CONCEPTS

### Statistics—What For?

To the uninitiated it may often appear that the statistician's primary function is to prevent or at least impede the progress of research. And even those who suspect that statistical methods may be more boon than bane are at times frustrated in their efforts to make use of the statistician's wares.

Much of the difficulty is due to not understanding the basic objectives of statistical methods. We can boil these objectives down to two:

1. The estimation of population parameters (values that characterize a particular population).
2. The testing of hypotheses about these parameters.

A common example of the first is the estimation of the coefficients  $a$  and  $b$  in the linear relationship,  $Y = a + bX$ , between the variables  $Y$  and  $X$ . To accomplish this objective one must first define the population involved and specify the parameters to be estimated. This is primarily the research worker's job. The statistician helps devise efficient methods of collecting the data and calculating the desired estimates.

Unless the whole population is examined, an estimate of a parameter is likely to differ to some degree from the population value. The unique contribution of statistics to research is that it provides ways of evaluating how far off the estimate may be. This is ordinarily done by computing confidence limits, which have a known probability of including the true value of the parameter. Thus, the mean diameter of the trees in a pine plantation may be estimated from a sample as 9.2 inches, with 95-percent confidence limits of 8.8 and 9.6 inches. These limits (if properly obtained) tell us that, unless a one-in-twenty chance has occurred in sampling, the true mean diameter is somewhere between 8.8 and 9.6 inches.

The second basic objective in statistics is to test some hypothesis about the population parameters. A common example is a test of the hypothesis that the regression coefficient  $b$  in the linear model

$$Y = a + bX$$

has some specified value (say zero). Another example is a test of the hypothesis that the difference between the means of two populations is zero.

Again, it is the research worker who should formulate meaningful hypotheses to be tested, not the statistician. This task can be tricky. The beginner would do well to work with the statistician to be sure that the hypothesis is put in a form that can be tested. Once the hypothesis is set, it is up to the statistician to work out ways of testing it and to devise efficient procedures for obtaining the data.

This handbook describes some of the methods of estimating certain parameters and testing some of the more common hypotheses.

## Probability and Statistics

It is fairly well known that statisticians work with probabilities. They are supposed to know, for example, the likelihood of tossing coins heads up six times in a row, or the chances of a craps shooter making seven consecutive winning throws ("passes"), and many other such useful bits of information. (This is assumed to give them an edge in games of chance, but often other factors enter in there.)

Despite this recognized association of statisticians and probability, the fundamental role of probability in statistical activities is often not appreciated. In putting confidence limits on an estimated parameter, the part played by probability is fairly obvious. Less apparent to the neophyte is the operation of probability in the testing of hypotheses. Some of them say with derision, "You can prove anything with statistics." The truth is, you can prove nothing; you can at most compute the probability of something happening and let the researcher draw his own conclusions.

To illustrate with a very simple example, in the game of craps the probability of the shooter winning (making a pass) is approximately 0.493—assuming, of course, a perfectly balanced set of dice and an honest shooter. Suppose now that you run up against a shooter who picks up the dice and immediately makes seven passes in a row! It can be shown that if the probability of making a single pass is really 0.493, then the probability of seven or more consecutive passes is about 0.007 (or 1 in 141). This is where statistics ends; you draw your own conclusions about the shooter. If you conclude that the shooter is pulling a fast one, then in statistical terms you are rejecting the hypothesis that the probability of the shooter making a single pass is 0.493.

Most statistical tests are of this nature. A hypothesis is formulated and an experiment is conducted or a sample is selected to test it. The next step is to compute the probability of the experimental or sample results occurring by chance if the hypothesis is true. If this probability is less than some preselected value (perhaps 0.05 or 0.01), the hypothesis is rejected. Note that nothing has been proved—we haven't even proved that the hypothesis is false. We merely inferred this because of the low probability associated with the experiment or sample results.

Obviously our inferences may be wrong if we are given inaccurate probabilities. Reliable computation of these probabilities requires a knowledge of how the variable we are dealing with is distributed (that is, what the probability is of the chance occurrence of different values of the variable). Thus, if we know that the number of beetles caught in light traps follows what is called the Poisson distribution we can compute the probability of catching  $X$  or more beetles. But, if we assume that this variable follows the Poisson when it actually follows the negative binomial distribution, our computed probabilities may be in error.

Even with reliable probabilities, statistical tests can lead to the wrong conclusions. We will sometimes reject a hypothesis that is true. If we always test at the 0.05 level, we will make this mistake on the average of 1 time in 20. We accept this degree of risk when we select the 0.05 level of testing. If we're willing to take a bigger risk, we can test at the 0.10 or the 0.25 level. If we're not willing to take this much risk, we can test at the 0.01 or 0.001 level.

The fellow who always wears both a belt and suspenders might, at this point, conclude that he should always test at the 0.00001 level. Then he'd be wrong only 1 time in 100,000. But a researcher can make more than one kind of error. In addition to rejecting a hypothesis that is true

(called a Type I error), he can make the mistake of not rejecting a hypothesis that is false (called a Type II error). In craps shooting, it is a mistake to accuse an honest shooter of cheating (Type I error—rejecting a true hypothesis), but it is also a mistake to trust a dishonest shooter (Type II error—failure to reject a false hypothesis).

The difficulty is that for a given set of data, reducing the risk of one kind of error increases the risk of the other kind. If we set 15 straight passes as the critical limit for a craps shooter, then we greatly reduce the risk of making a false accusation (probability about 0.00025). But in so doing we have dangerously increased the probability of making a Type II error—failure to detect a phony. A critical step in designing experiments is the attainment of an acceptable level of probability for each type of error. This is usually accomplished by specifying the level of testing (i.e., probability of an error of the first kind) and then making the experiment large enough to attain an acceptable level of probability for errors of the second kind.

It is beyond the scope of this handbook to go into basic probability computations, distribution theory, or the calculation of Type II errors. But anyone who uses statistical methods should be fully aware that he is dealing primarily with probabilities and not with immutable absolutes. The results of a  $t$ ,  $F$ , or chi-square test must be interpreted with this in mind. It is also well to remember that one-in-twenty chances do actually occur—about one time out of twenty.

## SOME BASIC TERMS AND CALCULATIONS

### The Mean

One of the most familiar and commonly estimated population parameters is the *mean*. Given a simple random sample, the population mean is estimated by

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

where:  $\bar{X}_i$  = The observed value of the  $i^{\text{th}}$  unit in the sample.

$n$  = The number of units in the sample.

$\sum_{i=1}^n X_i$ ; means to sum up all  $n$  of the  $X$ -values in the sample.

If there are  $N$  units in the population, the total of the  $X$ -values over all units in the population would be estimated by

$$T = N\bar{X}$$

The circumflex ( $\bar{\phantom{x}}$ ) over the  $T$  is frequently used to indicate an estimated value as opposed to the true but unknown population value.

It should be noted that this estimate of the mean is used for a simple random sample. It may not be appropriate if the units included in the sample are not selected entirely at random.

Methods of computing confidence limits for the mean are discussed in the section on sampling (see p. 11).

## Standard Deviation

Another commonly estimated population parameter is the *standard deviation*. The standard deviation characterizes dispersion of individuals about the mean. It gives us some idea whether most of the individuals in a population are close to the mean or spread out. The standard deviation of individuals in a population is frequently symbolized by  $\sigma$  (sigma). On the average, about two-thirds of the unit values of a normal population will be within 1 standard deviation of the mean. About 95 percent will be within 2 standard deviations and about 99 percent within 2.6 standard deviations.

We will seldom know or be able to determine  $\sigma$  exactly. However, given a sample of individual values from the population we can often make an estimate of  $\sigma$ , which is commonly symbolized by  $s$ . For a simple random sample of  $n$  units, the estimate is

$$s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}$$

where  $\sum X^2$  = the sum of squared values of all individual measurements,  $(\sum X)^2$  = the square of the sum of all measurements.

This is equivalent to the formula

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

where  $\bar{X}$  = the arithmetic mean =  $\frac{\sum X}{n}$

$(X - \bar{X})$  = the deviation of an individual measurement from the mean of all measurements.

Here is an example: Ten individual trees in a loblolly pine plantation were selected at random and measured. Their diameters were 9, 9, 11, 9, 7, 10, 8, 9, and 11 inches. Based on this sample, what is the arithmetic mean diameter and the standard deviation? Tabulating the measurements and squaring each of them:

$\bar{X}$	$X^2$
9	81
9	81
11	121
9	81
7	49
7	49
10	100
8	64
9	81
11	121
Sums	90      828

The mean

$$\bar{X} = \frac{\sum X}{n} = \frac{90}{10} = 9.0$$

The standard deviation

$$s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}} = \sqrt{\frac{828 - \frac{90^2}{10}}{9}} = \sqrt{\frac{18}{9}} = 1.414$$

Statisticians often speak in terms of the *variance* rather than standard deviation. The variance is simply the square of the standard deviation. The population variance is symbolized by  $\sigma^2$  and the sample estimate of the variance by  $s^2$ .

*Using the sample range to estimate the standard deviation.*—The standard deviation of the sample is an estimate of the standard deviation ( $\sigma$ ) of the population. The sample range ( $R$ ) may also be used to estimate the population standard deviation. Table 1 (Appendix, p. 76)<sup>1</sup> shows the ratio of the population standard deviation to the range for simple random samples of various sizes. In the example we've been using, the range is  $11 - 7 = 4$ . For a sample of size 10, the table gives the value of the ratio  $\frac{\sigma}{R}$  as 0.325. Therefore,  $\frac{\sigma}{4} = 0.325$  and  $\sigma = 1.3$  is an estimate of the true population standard deviation. Though easy to compute, this is an efficient estimator of  $\sigma$  only for very small samples (say less than 7 observations).

## Coefficient of Variation

In nature, populations with large means often show more variation than populations with small means. The *coefficient of variation* ( $C$ ) facilitates comparison of variability about different sized means. It is the ratio of the standard deviation to the mean. A standard deviation of 2 for a mean of 10 indicates the same relative variability as a standard deviation of 16 for a mean of 80. The coefficient of variation would be 0.20 or 20 percent in each case.

In the problem discussed in the previous section the coefficient of variation would be estimated by

$$C = \frac{s}{\bar{X}} = \frac{1.414}{9.0} = 0.157, \text{ or } 15.7 \text{ percent}$$

## Standard Error of the Mean

There is usually variation among the individual units of a population. The standard deviation is a measure of this variation.

Since the individual units vary, variation may also exist among the means (or any other estimates) computed from samples of these units. Take, for example, a population with a true mean of 10. If we were to select four units at random, they might have a sample mean of 8. Another sample of four units from the same population might have a mean of 11, another 10.5, and so forth. Clearly it would be desirable to know the variation likely to be encountered among the means of samples from this population. A measure of the variation among sample means is the standard error of the mean. It can be thought of as a standard deviation

<sup>1</sup> All tables referred to are in Appendix

among sample means; it is a measure of the variation among sample means, just as the standard deviation is a measure of the variation among individuals. As will be described in the section on simple random sampling, the standard error of the mean may be used to compute confidence limits for a population mean.

The computation of the standard error of the mean (often symbolized by  $s_{\bar{x}}$ ) depends on the manner in which the sample was selected. For simple random sampling without replacement (i.e., a given unit cannot appear in the sample more than once) from a population having a total of  $N$  units the formula for the estimated standard error of the mean is

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)}$$

In the problem discussed on page 4 we had  $n=10$  and found that  $s=1.414$  or  $s^2=2$ . If the population contained 1,000 trees, the estimated mean diameter ( $\bar{X}=9.0$  inches) would have a standard error of

$$s_{\bar{x}} = \sqrt{\frac{2}{10} \left(1 - \frac{10}{1,000}\right)} = \sqrt{0.198} \\ = 0.445$$

The term  $\left(1 - \frac{n}{N}\right)$  is called the *finite population correction* or *fpc*. If sampling is with replacement (not too common) or if the sampling fraction  $\left(\frac{n}{N}\right)$  is very small (say less than  $1/20$ ), the fpc may be omitted and the standard error of the mean for a simple random sample is simply

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

The variance of the sample mean is simply the square of the standard error of the mean.

$$s_{\bar{x}}^2 = \frac{s^2}{n} \left(1 - \frac{n}{N}\right)$$

### Covariance

Very often, each unit of a population will have more than a single characteristic. Trees, for example, may be characterized by their height, diameter, and form class. The *covariance* is a measure of the association between the magnitudes of two characteristics. If there is little or no association, the covariance will be close to zero. If the large values of one characteristic tend to be associated with the small values of another characteristic, the covariance will be negative. If the large values of one characteristic tend to be associated with the large values of another characteristic, the covariance will be positive. The population covariance of  $X$  and  $Y$  is often symbolized by  $\sigma_{xy}$ ; the sample estimate by  $s_{xy}$ .

Suppose that the diameter (inches) and age (years) have been obtained for a number of randomly selected trees. If we symbolize diameter by  $Y$  and age by  $X$ , the sample covariance of diameter and age is given by

$$s_{xy} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{(n-1)}$$

This is equivalent to the formula

$$s_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n-1)}$$

If  $n=12$  and the  $Y$  and  $X$  values were as follows:

Y	4	9	7	7	5	10	9	6	8	6	4	11	Sums
X	20	40	30	45	25	45	30	40	20	35	25	40	86

then

$$s_{xy} = \frac{(4)(20) + (9)(40) + \dots + (11)(40) - \frac{(86)(395)}{12}}{12-1} \\ = \frac{2,960 - 2,830.83}{11} = 11.74$$

The positive covariance is consistent with the well known and economically unfortunate fact that the larger diameters tend to be associated with the older ages.

### Simple Correlation Coefficient

The magnitude of the covariance, like that of the standard deviation, is often related to the size of the variables themselves. Units with large  $X$  and  $Y$  values tend to have larger covariances than units with small  $X$  and  $Y$  values. Also, the magnitude of the covariance depends on the scale of measurement; in the previous example, had diameter been expressed in millimeters instead of inches, the covariance would have been 298.196 instead of 11.74.

The *simple correlation coefficient*, a measure of the degree of linear association between two variables, is free of the effects of scale of measurement. It can vary from  $-1$  to  $+1$ . A correlation of 0 indicates that there is no linear association (there may be a very strong nonlinear association, however). A correlation of  $+1$  or  $-1$  would suggest a perfect linear association. As for the covariance, a positive correlation implies that the large values of  $X$  are associated with the large values of  $Y$ . If the large values of  $X$  are associated with the small values of  $Y$ , the correlation is negative.

The population correlation coefficient is commonly symbolized by  $\rho$  (rho), and the sample-based estimate by  $r$ . The population correlation coefficient is defined to be

$$\rho = \frac{\text{Covariance of } X \text{ and } Y}{\sqrt{(\text{Variance of } X)(\text{Variance of } Y)}}$$

For a simple random sample, the sample correlation coefficient is computed as follows:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}}$$

where:  $s_{xy}$  = Sample covariance of  $X$  and  $Y$   
 $s_x$  = Sample standard deviation of  $X$   
 $s_y$  = Sample standard deviation of  $Y$   
 $\sum xy$  = Corrected sum of  $XY$  products

$$= \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

$\sum x^2$  = Corrected sum of squares for  $X$

$$= \sum X^2 - \frac{(\sum X)^2}{n}$$

$\sum y^2$  = Corrected sum of squares for  $Y$

$$= \sum Y^2 - \frac{(\sum Y)^2}{n}$$

For the values used to illustrate the covariance we have:

$$\sum xy = (4)(20) + (9)(40) + \dots + (11)(40) - \frac{(86)(395)}{12} = 129.1667$$

$$\sum x^2 = 4^2 + 9^2 + \dots + 11^2 - \frac{86^2}{12} = 57.6667$$

$$\sum y^2 = 20^2 + 40^2 + \dots + 40^2 - \frac{395^2}{12} = 922.9167$$

$$So, \quad r = \frac{129.1667}{\sqrt{(57.6667)(922.9167)}} = \frac{129.1667}{230.6980} = .56$$

*Correlation or chance.*—The computed value of a statistic such as the correlation coefficient depends on which particular units were selected for the sample. Such estimates will vary from sample to sample. More important, they will usually vary from the population value which we try to estimate.

In the above example, the sample correlation coefficient was 0.56. Does this mean that there is a real linear association between  $Y$  and  $X$ ? Or could we get a value as large as this just by chance when sampling a population in which there is no linear association between  $Y$  and  $X$  (i.e., a population for which  $\rho = 0$ )?

This can be tested by referring to table 7 (App.). The column headed "Degrees of freedom" refers to the sample size. A correlation coefficient estimated from a simple random sample of  $n$  units will have  $(n-2)$  degrees of freedom. Looking in the row for 10 degrees of freedom we find in the column headed "5%", a value of 0.576. This says that in sampling from a population for which  $\rho = 0$  we would get a sample value as large as 0.576 just by chance about 5 percent of the time. Sample values smaller than 0.576 could occur more often than this. Thus we might conclude that our sample  $r = 0.56$  could have been obtained by chance in sampling from a population with a true correlation of zero.

This test result is usually summarized by saying that the sample correlation coefficient is not significant at the 0.05 level. In statistical terms, we tested the hypotheses that  $\rho = 0$  and failed to reject the hypothesis at the 0.05 level. This is not exactly the same as saying that we reject the hypothesis or that we have proved that  $\rho = 0$ . The distinction is subtle but real.

For a sample correlation larger than 0.576 we might decide that the departure from a value of zero is larger than we would expect by chance. Statistically we would reject the hypothesis that  $\rho = 0$ .

### Variance of a Linear Function

Quite often we will want to combine variables or population estimates in a linear function. For example, if the mean timber volume per acre has been estimated as  $\bar{X}$ , then the total volume on  $M$  acres will be  $M\bar{X}$ ; the estimate of total volume is a linear function of the estimated mean volume. If the estimate of cubic volume per acre in sawtimber is  $\bar{X}_1$  and of pulpwood above the sawtimber top is  $\bar{X}_2$ , then the estimate of total cubic foot volume per acre is  $\bar{X}_1 + \bar{X}_2$ . If on a given tract the mean volume per half-acre is  $\bar{X}_1$  for spruce and the mean volume per quarter-acre is  $\bar{X}_2$  for yellow birch, then the estimated total volume per quarter-acre and birch would be  $2\bar{X}_1 + 4\bar{X}_2$ .

In general terms, a linear function of three variables (say  $X_1, X_2$ , and  $X_3$ ) can be written as

$$L = a_1\bar{X}_1 + a_2\bar{X}_2 + a_3\bar{X}_3$$

where  $a_1, a_2$ , and  $a_3$  are constants.

If the variances are  $s_1^2, s_2^2$ , and  $s_3^2$  (for  $X_1, X_2$ , and  $X_3$  respectively) and the covariances are  $s_{1,2}, s_{1,3}$ , and  $s_{2,3}$ , then the variance of  $L$  is given by

$$s_L^2 = a_1^2s_1^2 + a_2^2s_2^2 + a_3^2s_3^2 + 2(a_1a_2s_{1,2} + a_1a_3s_{1,3} + a_2a_3s_{2,3})$$

The standard deviation (or standard error) of  $L$  is simply the square root of this.

The extension of the rule to cover any number of variables should be fairly obvious.

*Some examples.*—I. The sample mean volume per acre for a 10,000-acre tract is  $\bar{X} = 5,680$  board feet with a standard error of  $s_{\bar{X}} = 632$  ( $S_o s_{\bar{X}}^2 = 399,424$ ). The estimated total volume is

$$L = 10,000(\bar{X}) = 56,800,000 \text{ board feet.}$$

The variance of this estimate would be

$$s_L^2 = (10,000)^2(s_{\bar{X}}^2) = 39,942,400,000,000$$

Since the standard error of an estimate is the square root of its variance, the standard error of the estimated total is

$$s_L = \sqrt{s_L^2} = 6,320,000$$

II. In 1955 a random sample of 40 one-quarter-acre circular plots was used to estimate the cubic foot volume of a stand of pine. Plot centers were monumented for possible relocation at a later time. The mean volume per plot was  $\bar{X}_1 = 225$  cubic feet. The plot variance was  $s_{\bar{X}_1}^2 = 8,281$  so that the variance of the mean was  $s_{\bar{X}_1}^2 = 8,281/40 = 207.025$ .

In 1960 a second inventory was made using the same plot centers. This time, however, the circular plots were only one-tenth acre. The mean volume per plot was  $\bar{X}_2 = 122$  cubic feet. The plot variance was  $s_{\bar{X}_2}^2 = 6,084$ , so the variance of the mean was  $s_{\bar{X}_2}^2 = 152.100$ . The covariance of initial and final plot volumes was  $s_{\bar{X}_1, \bar{X}_2} = 4,259/40 = 106.475$ . The net periodic growth per acre would be estimated as

$$G = 10\bar{X}_2 - 4\bar{X}_1 = 10(122) - 4(225) = 320 \text{ cubic feet per acre.}$$

By the rule for linear functions the variance of  $G$  would be

$$s_G^2 = (10)^2 s_{z_2}^2 + (-4)^2 s_{z_1}^2 + 2(10)(-4) s_{z_1, z_2} \\ = 100(152.100) + 16(207.025) - 80(106.475) \\ = 10,004.4$$

In this example there was a statistical relationship between the 1960 and 1955 means because the same plot locations were used in both samples. The covariance of the means ( $s_{z_1, z_2}$ ) is a measure of this relationship. If the 1960 plots had been located at random rather than at the 1955 locations, the two means would have been considered statistically independent and their covariance would have been set at zero. In this case the equation for the variance of the net periodic growth per acre ( $G$ ) would reduce to

$$s_G^2 = (10)^2 s_{z_2}^2 + (-4)^2 s_{z_1}^2 \\ = 100(152.100) + 16(207.025) = 18,522.4$$

## SAMPLING—MEASUREMENT VARIABLES

### Simple Random Sampling

Most foresters are familiar with *simple random sampling*. As in any sampling system, the aim is to estimate some characteristic of a population without measuring all of the population units. In a simple random sample of size  $n$ , the units are selected so that every possible combination of  $n$  units has an equal chance of being selected. If sampling is with replacement, then at each stage of the sampling all units should have an equal chance of being selected. If sampling is without replacement, then at any stage of the sampling each unused unit should have an equal chance of being selected.

*Sample estimates of the population mean and total.*—From a population of  $N=100$  units,  $n=20$  units were selected at random and measured. Sampling was without replacement—once a unit had been included in the sample it could not be selected again. The unit values were:

10	9	10	9	11
16	11	7	12	12
11	3	5	11	14
8	13	12	20	10

Sum of all 20 random units = 214

From this sample we estimate the population mean as

$$\bar{X} = \frac{\sum X}{n} = \frac{214}{20} = 10.7$$

A population of  $N=100$  units having a mean of 10.7 would then have an estimated total of

$$\hat{T} = N\bar{X} = 100(10.7) = 1,070$$

### Standard Errors

The first step in calculating a standard error is to obtain an estimate of the population variance ( $\sigma^2$ ) or standard deviation ( $\sigma$ ). As noted in a previous section, the standard deviation for a simple random sample is estimated by

$$s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}} = \sqrt{\frac{10^2 + 16^2 + \dots + 10^2 - \frac{214^2}{20}}{19}} \\ = \sqrt{13.4842} = 3.672$$

For sampling without replacement, the standard error of the mean is

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)} = \sqrt{\frac{13.4842}{20} \left(1 - \frac{20}{100}\right)} \\ = \sqrt{0.539368} = 0.734$$

From the formula for the variance of a linear function we find that the variance of the estimated total is

$$s_{\hat{T}}^2 = N^2 s_{\bar{x}}^2$$

The standard error of the estimated total is the square root of this, or

$$s_{\hat{T}} = N s_{\bar{x}} = 100(0.734) = 73.4$$

### Confidence Limits

We have it on good authority that "you can fool all of the people some of the time." The oldest and simplest device for misleading folks is the barefaced lie. A method that is nearly as effective and far more subtle is to report a sample estimate without any indication of its reliability.

Sample estimates are subject to variation. How much they vary depends primarily on the inherent variability of the population ( $\sigma^2$ ) and on the size of the sample ( $n$ ) and of the population ( $N$ ).

The statistical way of indicating the reliability of an estimate is to establish *confidence limits*. For estimates made from normally distributed populations, the confidence limits are given by

$$(\text{Estimate}) \pm (t)(\text{Standard Error})$$

For setting confidence limits on the mean and total we already have everything we need except for the value of  $t$ , and that can be obtained from the table of the  $t$  distribution (table 2 in the appendix). In this table, the column headed df (degrees of freedom) refers to the size of the sample. For the mean (or total) of a simple random sample we would select a  $t$  value with  $(n-1)$  degrees of freedom. The columns labeled "Probability" refer to the kind of odds we demand. If we want to say that the true mean (or total) falls within certain limits unless a one-in-twenty chance has occurred, we use the  $t$  value in the column headed .05. If we want to say that the true value lies within a set of limits unless a one-in-one hundred chance has occurred, we select  $t$  from the column headed .01.

In the previous example the sample of  $n=20$  units had a mean of  $\bar{X}=10.7$  and a standard error of  $s_{\bar{x}}=0.734$ . For 95-percent confidence limits on the mean we would use a  $t$  value from the .05 column and the row for 19 degrees of freedom. As  $t_{.05}=2.093$ , the confidence limits are given by

$$\bar{X} \pm (t)(s_{\bar{x}}) = 10.7 \pm (2.093)(0.734) = 9.16 \text{ to } 12.24$$

This says that unless a one-in-twenty chance has occurred in sampling, the population mean is somewhere between 9.16 and 12.24. It does not



say where the mean of future samples from this population might fall. Nor does it say where the mean may be if mistakes have been made in the measurements.

For 99 percent confidence limits we find  $t_{.01} = 2.861$  (with 19 degrees of freedom), and so the limits are

$$10.7 \pm (2.861)(0.734) = 8.6 \text{ to } 12.8.$$

These limits are wider, but they are more likely to include the true population mean.

For the population total the confidence limits are:

$$95\text{-percent limits} = 1,070 \pm (2.093)(73.4) = 916 \text{ to } 1,224$$

$$99\text{-percent limits} = 1,070 \pm (2.861)(73.4) = 860 \text{ to } 1,280$$

For large samples ( $n > 60$ ) the 95-percent limits are closely approximated by

$$\text{Estimate} \pm (2)(\text{Standard Error})$$

and the 99-percent limits by

$$\text{Estimate} \pm (2.6)(\text{Standard Error})$$

### Sample size

Samples cost money. So do errors. The aim in planning a survey should be to take enough observations to obtain the desired precision—no more, no less.

The number of observations needed in a simple random sample will depend on the precision desired and the inherent variability of the population being sampled. Since sampling precision is often expressed in terms of confidence interval on the mean, it is not unreasonable in planning a survey to say that in the computed confidence interval

$$\bar{X} \pm ts_z$$

we would like to have the  $ts_z$  equal to or less than some specified value  $E$ , unless a one-in-twenty (or one-in-one hundred) chance has occurred in sampling. That is, we want

$$ts_z = E$$

or, since  $s_z = \frac{s}{\sqrt{n}}$ , we want

$$t \left( \frac{s}{\sqrt{n}} \right) = E$$

Solving this for  $n$  gives the desired sample size.

$$n = \frac{t^2 s^2}{E^2}$$

To apply this equation we need to have an estimate ( $s^2$ ) of the population variance and a value for students  $t$  at the appropriate level of probability.

The variance estimate can be a real problem. One solution is to make the sample survey in two stages. In the first stage,  $n_1$  random observations are made and from these an estimate ( $s^2$ ) of the variance is computed. Then this value is plugged into the sample size equation

$$n = \frac{t^2 s^2}{E^2}$$

where:  $t$  has  $n_1 - 1$  degrees of freedom and is selected from table 2 of the appendix. The computed value of  $n$  is the total size of sample needed. As we have already observed  $n_1$  units, this means that we will have to observe  $(n - n_1)$  additional units.

If pre-sampling as described above is not feasible then it will be necessary to make a guess at the variance. Assuming our knowledge of the population is such that the guessed variance ( $s^2$ ) can be considered fairly reliable, then the size of sample ( $n$ ) needed to estimate the mean to within  $\pm E$  units is approximately

$$n = \frac{4s^2}{E^2}$$

for 95 percent confidence and

$$n = \frac{20(s^2)}{3E^2}$$

for 99 percent confidence.

Less reliable variance estimates could be doubled (as a safety factor) before applying these equations. In many cases the variance estimate may be so poor as to make the sample size computation just so much statistical window dressing.

When sampling is without replacement (as it is in most forest sampling situations) the sample size estimates given above apply to populations with an extremely large number ( $N$ ) of units so that the sampling fraction ( $n/N$ ) is very small. If the sampling fraction is not small (say  $n/N \geq .05$ ) then the sample size estimates should be adjusted. This adjusted value of  $n$  is

$$n_a = \frac{n}{1 + \frac{n}{N}}$$

**Warning!** It is important that the specified error ( $E$ ) and the estimated variance ( $s^2$ ) be on the same scale of measurement. We could not, for example, use a board-foot variance in conjunction with an error expressed in cubic feet. Similarly, if the error is expressed in volume per acre, the variance must be put on a per-acre basis.

Suppose that we plan to use quarter-acre plots in a survey and estimate the variance among plot volumes to be  $s^2 = 160,000$ . If the error limit is  $E = 500$  feet per acre, we must convert the variance to an acre basis or the error to a quarter-acre basis. To convert a quarter-acre volume to a per-acre basis we multiply by 4, and to convert a quarter-acre variance to an acre variance we multiply by 16. Thus, the variance would be 2,560,000 and the sample-size formula would be

$$n = \frac{t^2(2,560,000)}{(500)^2} = t^2(10.24)$$

Alternatively, we can leave the variance alone and convert the error statement from an acre to a quarter-acre basis; i.e.,  $E = 125$ . Then the sample-size formula is

$$n = \frac{t^2(160,000)}{(125)^2} = t^2(10.24), \text{ as before.}$$

The problem of units of measure is not difficult, but the unwary can easily go astray.

### Stratified Random Sampling

In *stratified sampling*, a population is divided into subpopulations (strata) of known size, and a simple random sample of at least two units is selected in each subpopulation. This approach has several advantages. For one thing, if there is more variation between subpopulations than within them, the estimate of the population mean will be more precise than that given by a simple random sample of the same size. Also, it may be desirable to have separate estimates for each subpopulation (e.g., in timber types or administrative subunits). And it may be administratively more efficient to sample by subpopulations.

Example:

A 500-acre forested area was divided into three strata on the basis of timber type. A simple random sample of 0.2-acre plots was taken in each stratum, and the means, variances, and standard errors were computed by the formulae for a simple random sample. These results, along with the size ( $N_h$ ) of each stratum (expressed in number of 0.2-acre plots), are:

Type	Stratum number ( $h$ )	Stratum size ( $N_h$ )	Sample size ( $n_h$ )	Stratum mean ( $\bar{X}_h$ )	Within-stratum variance ( $s_h^2$ )	Squared standard error of the mean ( $s_{\bar{x}_h}^2$ )
Pine	1	1,350	30	251	10,860	353.96
Upland hardwoods	2	700	15	164	9,680	631.50
Bottom-land hardwoods	3	450	10	110	3,020	295.29
	Sum	2,500				

The squared standard error of the mean for stratum  $h$  is computed by the formula given for the simple random sample

$$s_{\bar{x}_h}^2 = \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

Thus, for stratum 1 (pine type),

$$s_{\bar{x}_1}^2 = \frac{10,860}{30} \left(1 - \frac{30}{1,350}\right) = 353.96$$

Where the sampling fraction ( $n_h/N_h$ ) is small, the fpc can be omitted.

With this data, the population mean is estimated by

$$\bar{X}_{st} = \sum \frac{N_h \bar{X}_h}{N}$$

where  $N = \sum N_h$

For this example we have

$$\begin{aligned} \bar{X}_{st} &= \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2 + N_3 \bar{X}_3}{N} = \frac{1,350(251) + 700(164) + 450(110)}{2,500} \\ &= 201.26 \end{aligned}$$

The formula for the standard error of the stratified mean is cumbersome but not complicated.

$$\begin{aligned} s_{\bar{x}_{st}} &= \sqrt{\frac{1}{N^2} \left[ \sum N_h^2 s_{\bar{x}_h}^2 \right]} \\ &= \sqrt{\frac{(1350)^2(353.96) + (700)^2(631.50) + (450)^2(295.29)}{(2,500)^2}} \\ &= 12.74 \end{aligned}$$

If the sample size is fairly large, the confidence limits on the mean are given by

$$95\text{-percent confidence limits} \doteq \bar{X}_{st} \pm 2s_{\bar{x}_{st}}$$

$$99\text{-percent confidence limits} \doteq \bar{X}_{st} \pm 2.6s_{\bar{x}_{st}}$$

There is no simple way of computing the confidence limits for small samples.

### Sample allocations

If a sample of  $n$  units is taken, how many units should be selected in each stratum? Among several possibilities, the most common procedure is to allocate the sample in proportion to the size of the stratum; in a stratum having two-fifths of the units of the population we would take two-fifths of the samples. In the population discussed in the previous example the proportional allocation of the 55 sample units would have been (and was) as follows:

Stratum	Relative size ( $N_h/N$ )	Sample allocation
1	0.54	29.7 or 30
2	0.28	15.4 or 15
3	0.18	9.9 or 10
Sums	1.00	55

For proportional allocation the number of sample units to be selected in stratum  $h$  is

$$n_h = \left(\frac{N_h}{N}\right)n$$

Some other possibilities are equal allocation, allocation proportional to estimated value, and optimum allocation. In optimum allocation an attempt is made to get the smallest standard error (of  $\bar{X}_{st}$ ) possible for a sample of  $n$  units. This is done by sampling more heavily in the strata having a larger variation. The equation for optimum allocation is

$$n_h = \left(\frac{N_h s_h}{\sum N_h s_h}\right)n$$

Optimum allocation obviously requires estimates of the within-stratum variances—information that may be difficult to obtain.

A refinement of optimum allocation is to take sampling cost differences into account and allocate the sample so as to get the most information per dollar. If the cost per sampling unit in stratum  $h$  is  $c_h$ , the equation is

$$n_h = \left( \frac{N_h s_h}{\sqrt{C_h}} \right) \left( \frac{n}{\sum \left( \frac{N_h s_h}{\sqrt{C_h}} \right)} \right)$$

### Sample size

To estimate the size of sample to take for a specified error at a given level of confidence, it is first necessary to decide on the method of allocation. Ordinarily, proportional allocation is the simplest and perhaps the best choice. With proportional allocation, the size of sample needed to be within  $\pm E$  units of the true value at the 0.05 probability level can be approximated by

$$n = \frac{N(\sum N_h s_h^2)}{4 - \sum N_h s_h^2}$$

For the 0.01 probability level, use 6.76 in place of 4.

To illustrate, assume that prior to sampling the 500-acre forest, we had decided that we wish to estimate the mean volume per acre to within  $\pm 100$  cubic feet per acre unless a 1-in-20 chance occurs in sampling. As we plan to sample with 0.2-acre plots, the error specification should be put on a 0.2-acre basis. Therefore,

$$E = 20$$

From previous sampling the stratum variances for 0.2-acre volumes are estimated to be

$$s_1^2 = 8,000 \quad s_2^2 = 10,000 \quad s_3^2 = 5,000$$

The stratum sizes are known to be as previously shown

$$N_1 = 1,350, \quad N_2 = 700, \quad N_3 = 450, \quad N = 2,500$$

Therefore,

$$n = \frac{(2,500)^2 (20)^2}{4} + [(1,350)(8,000) + (700)(10,000) + (450)(5,000)] = 77.7 \text{ or } 78$$

The 78 sample units would now be allocated to the strata by the formula

$$n_h = \left( \frac{N_h}{N} \right) n$$

giving

$$n_1 = 42, \quad n_2 = 22, \quad \text{and} \quad n_3 = 14$$

## SAMPLING—DISCRETE VARIABLES

### Random Sampling

The sampling methods discussed in the previous sections apply to data that are on a continuous or nearly continuous scale of measurement. These methods may not be applicable if each unit observed is classified as alive or dead, germinated or not germinated, infected or not infected. Data of this type may follow what is known as the binomial distribution. They require slightly different statistical techniques.

As an illustration, suppose that a sample of 1,000 seeds was selected at random and tested for germination. If 480 of the seeds germinated, the estimated viability for the lot would be

$$\bar{p} = \frac{480}{1,000} = 0.48, \text{ or } 48 \text{ percent}$$

Confidence limits for the population viability are easily obtained from appendix table 5: look in the "fraction observed" column for 0.48, and then move crosswise to the column for a sample of size 1,000. The figures in this column of the 95-percent side of the table are 45 and 51. Thus, unless a one-in-twenty chance has occurred in sampling, the germination percent for the population is between 45 and 51. The 99-percent confidence limits, obtained in the same manner, are 44 and 52.

If the sample size is  $n = 10, 15, 20, 30,$  or  $50,$  it will be necessary to look in the far left column for the number actually observed (rather than the fraction observed). Then in the appropriate sample-size column will be found the confidence limits for the fraction observed. Thus, for a germination of 24 seeds in a sample of 50 (so  $\bar{p} = 0.48$ ) the 95-percent confidence limits would be 0.34 and 0.63.

For large samples (say  $n > 250$ ) with proportions greater than 0.20 but less than 0.80, approximate confidence limits can be obtained another way. First we compute the standard error of  $\bar{p}$  by the equation

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n-1} \left( \frac{n}{1-N} \right)}$$

Then, the 95-percent confidence limits are given by

$$95\text{-percent confidence interval} = \bar{p} \pm \left[ 2(s_{\bar{p}}) + \frac{1}{2n} \right]$$

Applying this to the above example we get

$$s_{\bar{p}} = \sqrt{\frac{(0.48)(0.52)}{999}} \quad (\text{fpc ignored}) \\ = 0.0158$$

And,

$$95\text{-percent confidence interval} = .48 \pm \left[ 2(0.0158) + \frac{1}{2(1,000)} \right] \\ = .448 \text{ to } .512$$

The 99-percent confidence limits are approximated by

$$99\text{-percent confidence interval} = \bar{p} \pm \left[ 2.6s_{\bar{p}} + \frac{1}{2n} \right]$$

### Sample size

Table 5 can also be used to estimate the number of units that would have to be observed in a simple random sample in order to estimate a population proportion with some specified precision.

Suppose, for example, that we wanted to estimate the germination percent for a population to within plus or minus 10 percent (or 0.10) at the 95-percent confidence level. The first step is to guess about what the proportion of seed germinating will be. If a good guess is not possible, then the safest course is to guess  $\bar{p} = 0.50$  as this will give the maximum sample size.

Next, pick any of the sample sizes given in the table (10, 15, 20, 30, 50, 100, 250, and 1,000) and look at the confidence interval for the specified value of  $\bar{p}$ . Inspection of these limits will tell whether or not the precision will be met with a sample of this size or if a larger or smaller sample would be more appropriate.

Thus, if we guess  $\bar{p} = 0.2$ , then in a sample of  $n = 50$  we would expect to observe  $(0.2)(50) = 10$ , and the table says that the 95-percent confidence limits on  $\bar{p}$  would be 0.10 and 0.34. Since the upper limit is not within 0.10 of  $\bar{p}$ , a larger sample would be needed. For a sample of  $n = 100$  the limits are 0.13 to 0.29. Since both of these values are within 0.10 of  $\bar{p}$ , a sample of 100 would be adequate.

If the table indicates the need for a sample of over 250, the size can be approximated by

$$n = \frac{4(\bar{p})(1-\bar{p})}{E^2}, \text{ for 95-percent confidence}$$

or,

$$n = \frac{20(\bar{p})(1-\bar{p})}{3E^2}, \text{ for 99-percent confidence}$$

where:  $E$  = The precision with which  $\bar{p}$  is to be estimated (expressed in same form as  $\bar{p}$ , either percent or decimal).

### Cluster Sampling for Attributes

Simple random sampling of discrete variables is often difficult or impractical. In estimating plantation survival, for example, we could select individual trees at random and examine them, but it wouldn't make much sense to walk down a row of planted trees in order to observe a single member of that row. It would usually be more reasonable to select rows at random and observe all of the trees in the selected row.

Seed viability is often estimated by randomly selecting several lots of 100 or 200 seeds each and recording for each lot the percentage of the seeds that germinate.

These are examples of cluster sampling; the unit of observation is the cluster rather than the individual tree or single seed. The value attached to the unit is the proportion having a certain characteristic rather than the simple fact of having or not having that characteristic.

If the clusters are large enough (say over 100 individuals per cluster) and nearly equal in size, the statistical methods that have been described for measurement variables can often be applied. Thus, suppose that the germination percent of a seedlot is estimated by selecting  $n = 10$  sets of 200 seed each and observing the germination percent for each set. If the results were

Set	1	2	3	4	5	6	7	8	9	10	Sum
Germination percent ( $p$ )	78.5	82.0	86.0	80.5	74.5	78.0	79.0	81.0	80.5	83.5	803.5

then the mean germination percent is estimated by

$$\bar{p} = \frac{\sum p}{n} = \frac{803.5}{10} = 80.35 \text{ percent}$$

The standard deviation of  $p$  is

$$s_p = \sqrt{\frac{\sum p^2 - \frac{(\sum p)^2}{n}}{n-1}} = \sqrt{\frac{78.5^2 + \dots + 83.5^2 - \frac{(803.5)^2}{10}}{9}}$$

$$= \sqrt{10.002778} = 3.163$$

And the standard error for  $\bar{p}$  is

$$s_{\bar{p}} = \sqrt{\frac{s_p^2 \left(1 - \frac{n}{N}\right)}{n}}$$

$$= \sqrt{\frac{10.002778}{10}} = 1.000 \text{ (fpc ignored)}$$

Note that  $n$  and  $N$  in these equations refer to the number of clusters, not to the number of individuals.

The 95-percent confidence interval, computed by the procedure for continuous variables:

$$= \bar{p} \pm (t_{.05})(s_{\bar{p}}), \text{ (} t \text{ has } (n-1) = 9 \text{ df)}$$

$$= 80.35 \pm 2.262(1.000) = 78.1 \text{ to } 82.6$$

### Transformations

The above method of computing confidence limits assumes that the individual percentages follow something close to a normal distribution with homogeneous variance (i.e., same variance regardless of the size of the percent). If the clusters are small (say less than 100 individuals per cluster) or some of the percentages are greater than 80 or less than 20, the assumptions may not be valid and the computed confidence limits will be unreliable.

In such cases it may be desirable to compute the transformation

$$y = \text{arc sine } \sqrt{\text{percent}}$$

and to analyze the transformed variable. The transformation is easily made by means of table 6. Thus in the previous example we would have

percent	78.5	82.0	86.0	80.5	74.5	78.0	79.0	81.0	80.5	83.5	Sum
arc sine $\sqrt{\%}$	62.4	64.9	68.0	63.8	59.7	62.0	62.7	64.2	63.8	66.0	637.5

Then working with the transformed variables,

$$\bar{y} = \frac{637.5}{10} = 63.75, \text{ corresponding to a mean percentage of } 80.4$$

The variance of  $y$  is

$$s_y^2 = \frac{62.4^2 + \dots + 66.0^2 - \frac{637.5^2}{10}}{9} = 5.227222$$

And the standard error of  $\bar{y}$  is

$$s_{\bar{y}} = \sqrt{\frac{5.227222}{10}} = 0.723$$

The 95-percent confidence interval on mean  $y$  is

$$\bar{y} \pm (t_{.05})(s_{\bar{y}}) = 63.75 \pm (2.262)(0.723) \\ = 62.11 \text{ to } 65.39$$

These limits correspond to percentages of 78.1 to 82.7.

Because the clusters are fairly large and the value of  $\bar{p}$  close to .50, the transformation did not have much effect in this case.

## CHI-SQUARE TESTS

### Test of Independence

Individuals are often classified according to two (or more) distinct systems. A tree can be classified as to species and at the same time according to whether it is or is not infected with some disease. A milacre plot can be classified as to whether or not it is stocked with adequate reproduction and whether it is shaded or not shaded. Given such a cross-classification, it may be desirable to know whether the classification of an individual according to one system is independent of its classification by the other system. In the species-infection classification, for example, independence of species and infection would be interpreted to mean that there is no difference in infection rate between species (i.e., infection rate does not depend on species).

The hypothesis that two or more systems of classification are independent can be tested by *chi-square*. The procedure can be illustrated by a test of three termite repellents. A batch of 1,500 wooden stakes was divided at random into three groups of 500 each, and each group received a different termite-repellent treatment. The treated stakes were driven into the ground, with the treatment at any particular stake location being selected at random. Two years later the stakes were examined for termites. The number of stakes in each classification is shown in the following 2 by 3 (two rows and three columns) contingency table:

	Group I	Group II	Group III	Subtotals
Attacked by termites.....	193	148	210	551
Not attacked.....	307	352	290	949
Subtotals.....	500	500	500	1,500

If the data in the table be symbolized as shown below:

	I	II	III	
Attacked.....	$a_1$	$a_2$	$a_3$	A
Not attacked.....	$b_1$	$b_2$	$b_3$	B
	$T_1$	$T_2$	$T_3$	G

the test of independence is made by computing

$$\chi^2 = \frac{1}{(A)(B)} \sum_{i=1}^3 \left( \frac{[a_i B - b_i A]^2}{T_i} \right) \\ = \frac{1}{(551)(949)} \left[ \frac{((193)(949) - (307)(551))^2}{500} + \dots \right]$$

$$+ \frac{((210)(949) - (290)(551))^2}{500} \\ = 17.66$$

This result is compared to the tabular value of  $\chi^2$  (table 4) with  $(c-1)$  degrees of freedom, where  $c$  is the number of columns in the table of data. If the computed value exceeds the tabular value given in the 0.05 column, the difference among treatments is said to be significant at the 0.05 level (i.e., we reject the hypothesis that attack classification is independent of treatment classification).

In this example, the computed value of 17.66 (2 degrees of freedom) exceeds the tabular value in the 0.01 column, and so the difference in rate of attack among treatments is significant at the 1-percent level. Examination of the data suggests that this is primarily due to the lower rate of attack on the Group II stakes.

*The r by c contingency table.*—The above example is a simple case of the chi-square test of independence in an  $r$  by  $c$  table (i.e.,  $r$  rows and  $c$  columns). Thus, if a number of randomly selected forest stands were classified as to soil group and forest type the results might be as follows:

Soil group	I	II	III	Subtotal
1	27	48	62	137
2	32	46	67	145
3	26	51	61	138
Subtotal	85	145	190	420

If the  $r$  by  $c$  table is represented in symbols:

Soil group	I	II	III	Subtotal
1	$a_{11}$	$a_{12}$	$a_{13}$	$S_{1.}$
2	$a_{21}$	$a_{22}$	$a_{23}$	$S_{2.}$
3	$a_{31}$	$a_{32}$	$a_{33}$	$S_{3.}$
Subtotal	$T_{.1}$	$T_{.2}$	$T_{.3}$	G

then the test of independence is

$$\chi^2 = \frac{1}{G} \sum_{i,j} \left( \frac{(Ga_{ij} - S_i T_j)^2}{S_i T_j} \right), \text{ with } (r-1)(c-1) \text{ degrees of freedom}$$

In this example

$$\chi^2_{df=2} = \frac{1}{420} \left[ \frac{((420)(27) - (137)(85))^2}{(85)(137)} + \dots \right] \\ + \frac{((420)(61) - (138)(190))^2}{(138)(190)} \\ = 1.031$$

which is not significant at the 0.05 level. Thus, the test has failed to demonstrate any real association between forest types and soil groups. The test of independence can be extended to more than two classification systems, but formulating meaningful hypotheses may be difficult.

### Test of a Hypothesized Count

A geneticist hypothesized that, if a certain cross were made, the progeny would be of four types, in the proportions

$$A=0.48, \quad B=0.32, \quad C=0.12, \quad D=0.08$$

The actual segregation of 1,225 progeny is shown below, along with the numbers expected according to the hypothesis.

Type	A	B	C	D	Total
Number ( $X_i$ )	542	401	164	118	1,225
Expected ( $m_i$ )	588	392	147	98	1,225

As the observed counts differ from those expected, we might wonder if the hypothesis is false. Or, can departures as large as this occur strictly by chance?

The chi-square test is

$$\chi^2 = \sum_{i=1}^k \left( \frac{(X_i - m_i)^2}{m_i} \right), \text{ with } (k-1) \text{ degrees of freedom}$$

where:

$k$  = The number of groups recognized.

$X_i$  = The observed count for the  $i^{\text{th}}$  group.

$m_i$  = The count expected in the  $i^{\text{th}}$  group if the hypothesis is true.

For the above data,

$$\chi^2_{\text{df}} = \frac{(542-588)^2}{588} + \frac{(401-392)^2}{392} + \frac{(164-147)^2}{147} + \frac{(118-98)^2}{98} = 9.85$$

This value exceeds the tabular  $\chi^2$  with 3 degrees of freedom at the 0.05 level. Hence the hypothesis would be rejected (if the geneticist believed in testing at the 0.05 level).

### Bartlett's Test of Homogeneity of Variance

Many of the statistical methods described later are valid only if the variance is homogeneous. The  $t$  test of the following section assumes that the variance is the same for each group, and so does the analysis of variance. The fitting of an unweighted regression as described in the last section also assumes that the dependent variable has the same degree of variability (variance) for all levels of the independent variables.

Bartlett's test offers a means of evaluating this assumption. Suppose that we have taken random samples in each of four groups and obtained variances ( $s^2$ ) of 84.2, 63.8, 88.6, and 72.1 based on samples of 9, 21, 6, and 11 units, respectively. We would like to know if these variances could have come from populations all having the same variance. The quantities needed for Bartlett's test are tabulated here:

Group	Variance ( $s^2$ )	( $n-1$ )	Corrected sum of squares	$\frac{1}{n-1}$	$\log s^2$	$(n-1)(\log s^2)$
1	84.2	8	673.6	0.125	1.92531	15.40248
2	63.8	20	1,276.0	0.050	1.80482	36.09640
3	88.6	5	443.0	0.200	1.94743	9.73715
4	72.1	10	721.0	0.100	1.85794	18.57940
$k=4$ groups		Sums 43	3,113.6	0.475		79.81543

where:  $k$  = The number of groups (=4).

$$SS = \text{The corrected sum of squares} = \left( \sum X^2 - \frac{(\sum X)^2}{n} \right) = (n-1)s^2$$

From this we compute the pooled within-group variance

$$\bar{s}^2 = \frac{\sum SS_i}{\sum (n_i - 1)} = \frac{3113.6}{43} = 72.4093$$

and

$$\log \bar{s}^2 = 1.85979$$

Then the test of homogeneity is:

$$\chi^2_{(k-1) \text{ df}} = (2.3026) [(\log \bar{s}^2)(\sum (n_i - 1)) - \sum (n_i - 1)(\log s_i^2)] \textcircled{c},$$

with  $(k-1)$  degrees of freedom

In this case,

$$\chi^2_{\text{df}} = (2.3026) [(1.85979)(43) - 79.81543] = 0.358$$

This value of  $\chi^2$  is now compared with the value of  $\chi^2$  in table 4 for the desired probability level. A value greater than that given in the table would lead us to reject the homogeneity assumption.

The  $\chi^2$  value given by the above equation is biased upward. If  $\chi^2$  is nonsignificant, the bias is not important. However, if the computed  $\chi^2$  is just a little above the threshold value for significance, a correction for bias should be applied. The correction is:

$$C = \frac{3(k-1) + \left[ \sum \left( \frac{1}{n_i - 1} \right) - \frac{1}{\sum (n_i - 1)} \right]}{3(k-1)}$$

© The original form of this equation used natural logarithms in place of the common logarithms shown here. The natural log of any number is approximately 2.3026 times its common logarithm—hence the constant of 2.3026 in the equation. In computations, common logarithms are usually more convenient than natural logarithms.

$$\frac{3(4-1) + \left(0.475 - \frac{1}{43}\right)}{3(4-1)} = 1.0502$$

The corrected value of  $\chi^2$  is then

$$\text{Corrected } \chi^2 = \frac{\text{Uncorrected } \chi^2}{C} = \frac{0.358}{1.0502} = 0.341$$

## COMPARING TWO GROUPS BY THE *t* TEST

### The *t* Test for Unpaired Plots

An individual unit in a population may be characterized in a number of different ways. A single tree, for example, can be described as alive or dead, hardwood or softwood, infected or uninfected and so forth. When dealing with observations of this type we usually want to estimate the proportion of a population having a certain attribute. Or, if there are two or more different groups, we will often be interested in testing whether or not the groups differ in the proportions of individuals having the specified attribute. Some methods of handling these problems have been discussed in previous sections.

Alternatively, we might describe a tree by a measurement of some characteristics such as its diameter, height, or cubic volume. For this measurement type of observation we may wish to estimate the mean for a group as discussed in the section on sampling for measurement variables. If there are two or more groups we will frequently want to test whether or not the group means are different. Often the groups will represent types of treatment which we wish to compare. Under certain conditions, the *t* or *F* tests may be used for this purpose.

Both of these tests have a wide variety of applications. For the present we will confine our attention to tests of the hypothesis that there is no difference between treatment (or group) means. The computational routine depends on how the observations have been selected or arranged. The first illustration of a *t* test of the hypothesis that there is no difference between the means of two treatments assumes that the treatments have been assigned to the experimental units completely at random. Except for the fact that there are usually (but not necessarily) an equal number of units or "plots" for each treatment, there is no restriction on the random assignment of treatments.

In this example the "treatments" were two races of white pine which were to be compared on the basis of their volume production over a specified period of time. Twenty-two square one-acre plots were staked out for the study. Eleven of these were selected entirely at random and planted with seedlings of race A. The remaining eleven were planted with seedlings of race B. After the prescribed time period the pulpwood volume (in cords) was determined for each plot. The results were as follows:

Race A	Race B
11	5
8	10
10	8
8	8
Sum = 99	Sum = 88
Average = 9.0	Average = 8.0

To test the hypothesis that there is no difference between the race means (sometimes referred to as a null hypothesis) we compute

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s^2(n_A + n_B)}{(n_A)(n_B)}}}$$

where:  $\bar{X}_A$  and  $\bar{X}_B$  = The arithmetic means for groups A and B.

$n_A$  and  $n_B$  = The number of observations in groups A and B ( $n_A$  and  $n_B$  do not have to be the same).

$s^2$  = The pooled within-group variance (calculation shown below).

To compute the pooled within-group variance, we first get the corrected sum of squares (SS) within each group.

$$SS_A = \sum X_A^2 - \frac{(\sum X_A)^2}{n_A} = 11^2 + 8^2 + \dots + 11^2 - \frac{(99)^2}{11} = 34$$

$$SS_B = \sum X_B^2 - \frac{(\sum X_B)^2}{n_B} = 9^2 + 9^2 + \dots + 6^2 - \frac{(88)^2}{11} = 54$$

Then the pooled variance is

$$s^2 = \frac{SS_A + SS_B}{(n_A - 1) + (n_B - 1)} = \frac{88}{20} = 4.4$$

Hence,

$$t = \frac{9.0 - 8.0}{\sqrt{4.4 \left( \frac{11+11}{(11)(11)} \right)}} = \frac{1.0}{\sqrt{.800000}} = 1.118$$

This value of *t* has  $(n_A - 1) + (n_B - 1)$  degrees of freedom. If it exceeds the tabular value of *t* (table 2) at a specified probability level, we would reject the hypothesis. The difference between the two means would be considered significant (larger than would be expected by chance if there is actually no difference).

In this case, tabular *t* with 20 degrees of freedom at the 0.05 level is 2.086. Since our sample value is less than this, the difference is not significant at the 0.05 level.

*Requirements.*—One of the unfortunate aspects of the *t* test and other statistical methods is that almost any kind of numbers can be plugged into the equations. But if the numbers and methods of obtaining them do not meet certain requirements, the result may be a fancy statistical facade with nothing behind it. In a handbook of this scope it is not possible to make the reader aware of all of the niceties of statistical usage, but a few words of warning are certainly appropriate.

A fundamental requirement in the use of most statistical methods is that the experimental material be a random sample of the population to which the conclusions are to be applied. In the  $t$  test of white pine races, the plots should be a sample of the sites on which the pines are to be grown, and the planted seedlings should be a random sample representing the particular race. A test conducted in one corner of an experimental forest may yield conclusions that are valid only for that particular area or sites that are about the same. Similarly, if the seedlings of a particular race are the progeny of a small number of parents, their performance may be representative of those parents only, rather than of the race.

In addition to assuming that the observations for a given race are a valid sample of the population of possible observations, the  $t$  test described above assumes that the population of such observations follows the normal distribution. With only a few observations, it is usually impossible to determine whether or not this assumption has been met. Special studies can be made to check on the distribution, but often the question is left to the judgment and knowledge of the research worker.

Finally, the  $t$  test of unpaired plots assumes that each group (or treatment) has the same population variance. Since it is possible to compute a sample variance for each group, this assumption can be checked with Bartlett's test for homogeneity of variance. Most statistical textbooks present variations of the  $t$  test that may be used if the group variances are unequal.

### Sample size

If there is a real difference of  $D$  feet between the two races of white pine, how many replicates (plots) would be needed to show that it is significant? To answer this, we first assume that the number of replicates will be the same for each group ( $n_A = n_B = n$ ). The equation for  $t$  can then be written

$$t = \frac{D}{\sqrt{\frac{2s^2}{n}}} \quad \text{or} \quad n = \frac{2t^2 s^2}{D^2}$$

Next we need an estimate of the within-group variance,  $s^2$ . As usual, this must be determined from previous experiments, or by special study of the populations.

*Example.*—Suppose that we plan to test at the 0.05 level and wish to detect a true difference of  $D = 1$  cord if it exists. From previous tests we estimate  $s^2 = 5.0$ . Thus we have

$$n = \frac{2t^2 s^2}{D^2} = 2t^2 \left( \frac{5.0}{1.0} \right)$$

Here we hit a snag. In order to estimate  $n$  we need a value for  $t$ , but the value of  $t$  depends on the number of degrees of freedom, which depends on  $n$ . The situation calls for an iterative solution—a fancy name for trial and error. We start with a guessed value of  $n$ , say  $n_0 = 20$ . As  $t$  has  $(n_A - 1) + (n_B - 1) = 2(n - 1)$  degrees of freedom, we'll use  $t = 2.025$  ( $= t_{.05}$  with 38 df) and compute

$$n_1 = 2(2.025)^2 \left( \frac{5.0}{1.0} \right) = 41$$

The proper value of  $n$  will be somewhere between  $n_0$  and  $n_1$ —much closer to  $n_1$  than to  $n_0$ . We can now make a second guess at  $n$  and repeat the process. If we try  $n_2 = 38$ ,  $t$  will have  $2(n - 1) = 74$  df and  $t_{.05} = 1.992$ . Hence,

$$n_3 = 2(1.992)^2 \left( \frac{5.0}{1.0} \right) = 39.7$$

Thus,  $n$  appears to be over 39 and we will use  $n = 40$  plots for each group or a total of 80 plots.

### The $t$ Test for Paired Plots

A second test was made of the two races of white pine. It also had 11 replicates of each race, but instead of the two races being assigned completely at random over the 22 plots, the plots were grouped into 11 pairs and a different race was randomly assigned to each member of a pair. The cordwood volumes at the end of the growth period were

Plot pair	1	2	3	4	5	6	7	8	9	10	11	Sum	Mean
Race A	12	8	8	11	10	9	11	11	13	10	7	110	10.0
Race B	10	7	8	9	11	6	10	11	10	8	9	99	9.0

$$d_i = A_i - B_i; \quad 2 \quad 1 \quad 0 \quad 2 \quad -1 \quad 3 \quad 1 \quad 0 \quad 3 \quad 2 \quad -2 \quad 11 \quad 1.0$$

As before, we wish to test the hypothesis that there is no real difference between the race means.

The value of  $t$  when the plots have been paired is

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_d^2}{n}}} = \frac{\bar{d}}{\sqrt{s_d^2}}, \quad \text{with } (n - 1) \text{ degrees of freedom}$$

where:  $n$  = The number of pairs of plots

$s_d^2$  = The variance of the individual differences between A and B

$$s_d^2 = \frac{\sum d_i^2}{n - 1} = \frac{2^2 + 1^2 + \dots + (-2)^2 - \frac{11^2}{11}}{10} = 2.6$$

So, in this example we find

$$t_{10 \text{ df}} = \frac{10.0 - 9.0}{\sqrt{2.6/11}} = 2.057$$

Comparing this to the tabular value of  $t$  ( $t_{.05}$  with 10 df = 2.228), we find that the difference is not significant at the 0.05 level. That is, a sample mean difference of 1 cord or more could have occurred by chance more than one time in twenty even if there is no real difference between the race means. Usually such an outcome is not regarded as sufficiently strong evidence to reject the hypothesis.



The paired test will be more sensitive (capable of detecting smaller real differences) than the unpaired test whenever the experimental units (plots in this case) can be grouped into pairs such that the variation between pairs is appreciably larger than the variation within pairs. The basis for pairing plots may be geographic proximity or similarity in any other characteristic that is expected to affect the performance of the plot. In animal-husbandry studies, litter mates are often paired, and where patches of human skin are the "plots," the left and right arms may constitute the pair. If the experimental units are very homogeneous, there may be no advantage in pairing.

### Number of replicates

The number ( $n$ ) of plot pairs needed to detect a true mean difference of size  $D$  is

$$n = \frac{t^2 s_d^2}{D^2}$$

N.B.: Be sure to use the variance of the difference ( $s_d^2$ ) between paired plots in this equation and not the variance among plots.

## COMPARISON OF TWO OR MORE GROUPS BY

### ANALYSIS OF VARIANCE

#### Complete Randomization

A planter wanted to compare the effects of five site-preparation treatments on the early height growth of planted pine seedlings. He laid out 25 plots, and applied each treatment to 5 randomly selected plots. The plots were then hand-planted and at the end of 5 years the height of all pines was measured and an average height computed for each plot. The plot averages (in feet) were as follows:

	Treatments				
	A	B	C	D	E
	15	16	13	11	14
	14	14	12	13	12
	12	13	11	10	12
	13	15	12	12	10
	13	14	10	11	11
Sums	67	72	58	57	59
Treatment means	13.4	14.4	11.6	11.4	11.8
	$s_A^2 = 1.5$	$s_B^2 = 1.3$	$s_C^2 = 1.3$	$s_D^2 = 1.2$	$s_E^2 = 1.7$
					12.52

Looking at the data we see that there are differences among the treatment means: A and B have higher averages than C, D, and E. Soils and planting stock are seldom completely uniform, however, and so we would expect some differences even if every plot had been given exactly the same site-preparation treatment. The question is, can differences as large as this

occur strictly by chance if there is actually no difference among treatments? If we decide that the observed differences are larger than might be expected to occur strictly by chance, the inference is that the treatment means are not equal. Statistically speaking, we reject the hypothesis of no difference among treatment means.

Problems like this are neatly handled by an analysis of variance. To make this analysis, we need to fill in a table like the following:

Source of variation	Degrees of freedom	Sums of squares	Mean squares
Treatments.....	4		
Error.....	20		
Total.....	24		

*Source of variation.*—There are a number of reasons why the height growth of these 25 plots might vary, but only one can be definitely identified and evaluated—that attributable to treatments. The unidentified variation is assumed to represent the variation inherent in the experimental material and is labeled error. Thus, total variation is being divided into two parts: one part attributable to treatments, and the other unidentified and called error.

*Degrees of freedom.*—Degrees of freedom are hard to explain in non-statistical language. In the simpler analyses of variance, however, they are not difficult to determine. For the total, the degrees of freedom are one less than the number of observations: there are 25 plots, so the total has 24 df's. For the sources, other than error, the df's are one less than the number of classes or groups recognized in the source. Thus, in the source labeled treatments there are five groups (five treatments), so there will be four degrees of freedom for treatments. The remaining degrees of freedom ( $24 - 4 = 20$ ) are associated with the error term.

*Sums of squares.*—There is a sum of squares associated with every source of variation. These SS are easily calculated in the following steps: First we need what is known as a "correction term" or C.T. This is simply

$$C.T. = \frac{\left(\sum X\right)^2}{n} = \frac{313^2}{25} = 3918.76$$

where:  $\sum$  = the sum of  $n$  items

Then the total sum of squares is

$$\text{Total SS} = \sum_{24 \text{ df}} X^2 - C.T. = (15^2 + 14^2 + \dots + 11^2) - C.T. = \boxed{64.24}$$

The sum of squares attributable to treatments is

$$\text{Treatment SS} = \frac{\sum (\text{treatment totals}^2)}{\text{No. of plots per treatment}} - C.T. = \frac{67^2 + 72^2 + \dots + 59^2}{5} - C.T. = \frac{19767}{5} - C.T. = \boxed{34.64}$$

Note that in both SS calculations, the number of items squared and added was one more than the number of degrees of freedom associated with the sum of squares. The number of degrees of freedom just below the SS and the numbers of items to be squared and added just over the  $\Sigma$ , provided a partial check as to whether the proper totals are being used in the calculation—the degrees of freedom must be one less than the number of items.

Note also that the divisor in the treatment SS calculation is equal to the number of individual items that go to make up each of the totals being squared in the numerator. This was also true in the calculation of total SS, but there the divisor was 1 and hence did not have to be shown. Note further that the divisor times the number over the summation sign ( $5 \times 5 = 25$  for treatments) must always be equal to the total number of observations in the test—another check.

The sum of squares for error is obtained by subtracting the treatment SS from total SS. A good habit to get into when obtaining sums of squares by subtraction is to perform the same subtraction using df's. In the more complex designs, doing this provides a partial check on whether the right items are being used.

**Mean squares.**—The mean squares are now calculated by dividing the sums of squares by the associated degrees of freedom. It is not necessary to calculate the mean square for the total.

The items that have been calculated are entered directly into the analysis table, which at the present stage would look like this:

Source	df	SS	MS
Treatment.....	4	34.64	8.66
Error.....	20	29.60	1.48
Total	24	64.24	

An  $F$  test of treatments is now made by dividing the MS for treatments by the MS for error. In this case

$$F = \frac{8.66}{1.48} = 5.851$$

This figure is compared to the appropriate value of  $F$  in table 3 of the appendix. Look across the top to the column headed 4 (corresponding to the degrees of freedom for treatments). Follow down the column to the row labeled 20 (corresponding to the degrees of freedom for error). The tabular  $F$  for significance at the 0.05 level is 2.87, and that for the 0.01 level is 4.43. As the calculated value of  $F$  exceeds 4.43, we conclude that the difference in height growth between treatments is significant at the 0.01 level. (More precisely, we reject the hypothesis that there is no difference in mean height growth between the treatments.) If  $F$  had been smaller than 4.43 but larger than 2.87, we would have said that the difference is significant at the 0.05 level. If  $F$  had been less than 2.87, we would have said that the difference between treatments is not significant at the 0.05 level. The researcher should select his own level of significance (preferably in advance of the study), keeping in mind that significance at the  $\alpha$  (alpha) level (for example) means this: if there is actually no difference among treatments, the probability of getting chance differences as large as those observed is  $\alpha$  or less.

*The  $t$  test versus the analysis of variance.*—If only two treatments are being compared, the analysis of variance of variance of a completely randomized design and the  $t$  test of unpaired plots lead to the same conclusion. The choice of test is strictly one of personal preference, as may be verified by applying the analysis of variance to the data used to illustrate the  $t$  test of unpaired plots. The resulting  $F$  value will be equal to the square of the value of  $t$  that was obtained (i.e.,  $F = t^2$ ).

Like the  $t$  test, the  $F$  test is valid only if the variable observed is normally distributed and if all groups have the same variance.

### Multiple Comparisons

In the example illustrating the completely randomized design, the difference among treatments was found to be significant at the 0.01 probability level. This is interesting as far as it goes, but usually we will want to take a closer look at the data, making comparisons among various combinations of the treatments.

Suppose, for example, that A and B involved some mechanical form of site preparation while C, D, and E were chemical treatments. Then we might want to test whether the average of A and B together differed from the combined average of C, D, and E. Or, we might wish to test whether A and B differ significantly from each other. When the number of replications ( $n$ ) is the same for all treatments, such comparisons are fairly easy to define and test.

The question of whether the average of treatments A and B differs significantly from the average of treatments C, D, and E is equivalent to testing whether the linear contrast

$$Q = (3\bar{A} + 3\bar{B}) - (2\bar{C} + 2\bar{D} + 2\bar{E})$$

differs significantly from zero ( $\bar{A}$  = the mean for treatment A, etc.). Note that the coefficients of this contrast sum to zero ( $3 + 3 - 2 - 2 - 2 = 0$ ) and are selected so as to put the two means in the first group on an equal basis with the three means in the second group.

Similarly, testing whether treatment A differs significantly from treatment B is the same as testing whether the contrast  $Q = \bar{A} - \bar{B}$  differs significantly from zero.

### F Test with Single Degree of Freedom

A comparison specified in advance of the study (on logical grounds and before examination of the data) can be tested by an  $F$  test with single degree of freedom. For the linear contrast

$$Q = a_1\bar{X}_1 + a_2\bar{X}_2 + a_3\bar{X}_3 + \dots$$

among means based on the same number ( $n$ ) of observations, the sum of squares has one degree of freedom and is computed as

$$SS_{1 \text{ df}} = \frac{nQ^2}{\sum a_i^2}$$

This sum of squares divided by the mean square for error provides an  $F$  test of the comparison.

Thus, in testing A and B versus C, D, and E we have

$$Q = 3(13.4) + 3(14.4) - 2(11.6) - 2(11.4) - 2(11.8) = 13.8$$

and

$$SS_{1 \text{ df}} = \frac{5(13.8)^2}{3^2 + 3^2 + (-2)^2 + (-2)^2} = \frac{952.20}{30} = 31.74$$

Then dividing by the error mean square gives the  $F$  value for testing the contrast.

$$F = \frac{31.74}{1.48} = 21.446 \text{ with 1 and 20 degrees of freedom}$$

This exceeds the tabular value of  $F(4.35)$  at the 0.05 probability level. If this is the level at which we decided to test, we would reject the hypothesis that the mean of treatments A and B does not differ from the mean of treatments C, D, and E.

If  $Q$  is expressed in terms of the treatment totals rather than their means so that

$$Q_T = a_1(\Sigma X_1) + a_2(\Sigma X_2) + \dots$$

then the equation for the single degree of freedom sum of squares is

$$SS_{1 \text{ df}} = \frac{Q_T^2}{n \Sigma a_i^2}$$

The results will be the same as those obtained with the means. For the test of A and B versus C, D, and E,

$$Q_T = 3(67) + 3(72) - 2(58) - 2(57) - 2(59) = 69$$

And,

$$SS_{1 \text{ df}} = \frac{69^2}{5[3^2 + 3^2 + (-2)^2 + (-2)^2]} = \frac{4761}{150} = 31.74, \text{ as before.}$$

Working with the totals saves the labor of computing means and avoids possible rounding errors.

### Scheffé's Test

Quite often we will want to test comparisons that were not anticipated before the data were collected. If the test of treatments was significant, such unplanned comparisons can be tested by the method of Scheffé. When there are  $n$  replications of each treatment,  $k$  degrees of freedom for treatment, and  $v$  degrees of freedom for error, any linear contrast among the treatment means

$$Q = a_1 \bar{X}_1 + a_2 \bar{X}_2 + \dots$$

is tested by computing

$$F = \frac{nQ^2}{k(\Sigma a_i^2)(\text{Error mean square})}$$

This value is then compared to the tabular value of  $F$  with  $k$  and  $v$  degrees of freedom.

For example, to test treatment B against the means of treatments C and E we would have

$$Q = [2B - (\bar{C} + \bar{E})] = [2(14.4) - 11.6 - 11.8] = 5.4$$

and

$$F = \frac{5(5.4)^2}{(4)[2^2 + (-1)^2 + (-1)^2](1.48)} = 4.105, \text{ with 4 and 20 degrees of freedom}$$

This figure is larger than the tabular value of  $F (= 2.87)$ , and so in testing at the 0.05 level we would reject the hypothesis that the mean for treatment B did not differ from the combined average of treatments C and E.

For a contrast ( $Q_T$ ) expressed in terms of treatment totals, the equation for  $F$  becomes

$$F = \frac{Q_T^2}{nk(\Sigma a_i^2)(\text{Error mean square})}$$

### Unequal Replication

If the number of replications is not the same for all treatments, then for the linear contrast

$$Q = a_1 \bar{X}_1 + a_2 \bar{X}_2 + \dots$$

the sum of squares in the single degree of freedom  $F$  test is given by

$$SS_{1 \text{ df}} = \frac{Q^2}{\left(\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + \dots\right)}$$

where:  $n_i$  = the number of replications on which  $\bar{X}_i$  is based.

With unequal replication, the  $F$  value in Scheffé's test is computed by the equation

$$F = \frac{Q^2}{(k) \left(\frac{a_1^2}{n_1} + \frac{a_2^2}{n_2} + \dots\right) (\text{Error mean square})}$$

Selecting the coefficients ( $a_i$ ) for such contrasts can be tricky. When testing the hypothesis that there is no difference between the means of two groups of treatments, the positive coefficients are usually

$$\text{positive } a_i = \frac{n_i}{p}$$

where  $p$  = the total number of plots in the group of treatments with positive coefficients.

The negative coefficients are

$$\text{negative } a_j = \frac{n_j}{m}$$

where  $m$  = the total number of plots in the group of treatments with negative coefficients.

To illustrate, if we wish to compare the mean of treatments A, B, and C with the mean of treatments D and E and there are two plots of treatment A, three of B, five of C, three of D, and two of E, then  $p = 2 + 3 + 5 = 10$ ,  $m = 3 + 2 = 5$  and the contrast would be

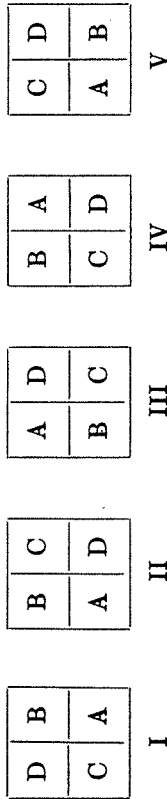
$$Q = \left(\frac{2}{10}A + \frac{3}{10}B + \frac{5}{10}C\right) - \left(\frac{3}{5}D + \frac{2}{5}E\right)$$

## Randomized Block Design

In the completely randomized design the error mean square is a measure of the variation among plots treated alike. It is in fact an average of the within-treatment variances, as may easily be verified by computation. If there is considerable variation among plots treated alike, the error mean square will be large and the  $F$  test for a given set of treatments is less likely to be significant. Only large differences among treatments will be detected as real and the experiment is said to be insensitive.

Often the error can be reduced (thus giving a more sensitive test) by use of a randomized block design in place of complete randomization. In this design, similar plots or plots that are close together are grouped into blocks. Usually the number of plots in each block is the same as the number of treatments to be compared, though there are variations having two or more plots per treatment in each block. The blocks are recognized as a source of variation that is isolated in the analysis.

As an example, a randomized block design with five blocks was used to test the height growth of cottonwood cuttings from four selected parent trees. The field layout looked like this:



Each plot consisted of a planting of 100 cuttings of the clone assigned to that plot. When the trees were 5 years old the heights of all survivors were measured and an average computed for each plot.

The plot averages (in feet) by clones and blocks are summarized below:

Block	Clone				Block totals
	A	B	C	D	
I	18	14	12	16	60
II	15	15	16	13	59
III	16	15	8	15	54
IV	14	12	10	12	48
V	12	14	9	14	49
Clone totals	75	70	55	70	270
Clone means	15	14	11	14	

The hypothesis to be tested is that clones do not differ in mean height. In this design there are two identifiable sources of variation—that attributable to clones and that associated with blocks. The remaining portion of the total variation is used as a measure of experimental error. The outline of the analysis is therefore as follows:

Source of variation	df	Sums of squares	Mean squares
Blocks	4		
Clones	3		
Error	12		
Total	19		

The breakdown in degrees of freedom and computation of the various sums of squares follow the same pattern as in the completely randomized design. Total degrees of freedom (19) are one less than the total number of plots. Degrees of freedom for clones (3) are one less than the number of clones. With five blocks, there will be four degrees of freedom for blocks. The remaining 12 degrees of freedom are associated with the error term.

Sums-of-squares calculations proceed as follows:

1. The correction term

$$C.T. = \frac{(\sum X)^2}{n} = \frac{270^2}{20} = 3,645$$

$$2. \text{ Total SS}_{19 \text{ df}} = \sum X^2 - C.T. = (18^2 + 15^2 + \dots + 14^2) - C.T. = 3,766 - 3,645 = 121$$

$$3. \text{ Clone SS}_{3 \text{ df}} = \frac{\sum (\text{Clone totals}^2)}{\text{No. of plots per clone}} - C.T. = \frac{75^2 + 70^2 + 55^2 + 70^2}{5} - C.T. = 3,690 - 3,645 = 45$$

$$4. \text{ Block SS}_{4 \text{ df}} = \frac{\sum (\text{Block totals}^2)}{\text{No. of plots per block}} - C.T. = \frac{60^2 + 59^2 + \dots + 49^2}{4} - C.T. = 3,675.5 - 3,645 = 30.5$$

$$5. \text{ Error SS}_{12 \text{ df}} = \text{Total SS}_{19 \text{ df}} - \text{Clone SS}_{3 \text{ df}} - \text{Block SS}_{4 \text{ df}} = 45.5$$

Note that in obtaining the error SS by subtraction, we get a partial check on ourselves by subtracting clone and block df's from the total df to see if we come out with the correct number for error df. If these don't check, we have probably used the wrong sums of squares in the subtraction. Mean squares are again calculated by dividing the sums of squares by the associated number of degrees of freedom. Tabulating the results of these computations

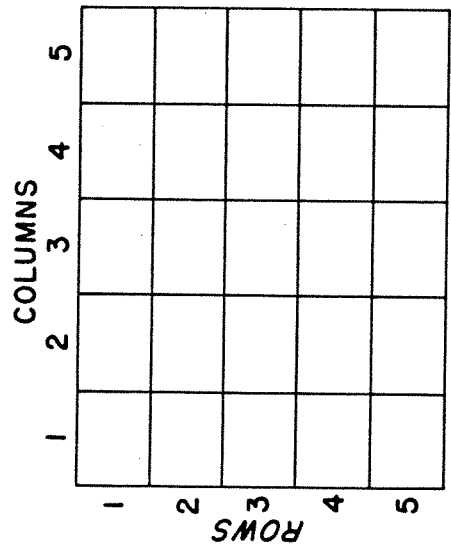
of  $F$  will be equal to the value of  $t^2$  and the inferences derived from the tests will be the same. The choice of tests is a matter of personal preference.

### Latin Square Design

In the randomized block design the purpose of blocking is to isolate a recognizable extraneous source of variation. If successful, blocking reduces the error mean square and hence gives a more sensitive test than could be obtained by complete randomization.

In some situations, however, we have a two-way source of variation that cannot be isolated by blocks alone. In a field, for example, fertility gradients may exist both parallel to and at right angles to plowed rows. Simple blocking isolates only one of these sources of variation, leaving the other to swell the error term and reduce the sensitivity of the test.

When such a two-way source of extraneous variation is recognized or suspected, the Latin square design may be helpful. In this design, the total number of plots or experimental units is made equal to the square of the number of treatments. In forestry and agricultural experiments, the plots are often (but not always) arranged in rows and columns with each row and column having a number of plots equal to the number of treatments being tested. The rows represent different levels of one source of extraneous variation while the columns represent different levels of the other source of extraneous variation. Thus, before the assignment of treatments, the field layout of a Latin square for testing five treatments might look like this:



Treatments are assigned to plots at random, but with the very important restriction that a given treatment cannot appear more than once in any row or any column.

An example of a field layout of a Latin square for testing five treatments is given below. The letters represent the assignment of five treatments (which here are five species of hardwoods). The numbers show the average 5-year height growth by plots. The tabulation shows the totals for rows, columns, and treatments.

$F$  for clones is obtained by dividing clone MS by error MS. In this case  $F = \frac{15.000}{3.792} = 3.956$ . As this is larger than the tabular  $F$  of 3.49 ( $F_{.05}$  with 3 and 12 degrees of freedom) we conclude that the difference between clones is significant at the 0.05 level. The significance appears to be due largely to the low value of C as compared to A, B, and D.

Comparisons among clone means can be made by the methods previously described. For example, to test the hypothesis (i.e., before examining the data) hypothesis that there is no difference between the mean of clone C and the combined average of A, B, and D we would have:

$$SS \text{ for } (A+B+D \text{ vs. } C) = \frac{5(3\bar{C} - \bar{A} - \bar{B} - \bar{D})^2}{3^2 + (-1)^2 + (-1)^2 + (-1)^2}$$

$$= \frac{5(-10)^2}{12} = \frac{41.667}{12}$$

Then,

$$F = \frac{41.667}{3.792} = 10.988$$

*Student's T*  $F = \frac{25.00}{682.6} = 3.66$

Tabular  $F$  at the 0.01 level with 1 and 12 degrees of freedom is 9.33. As calculated  $F$  is greater than this, we conclude that the difference between C and the average of A, B, and D is significant at the 0.01 level.

The sum of squares for this single-degree-of-freedom comparison (41.667) is almost as large as that for clones (45.0) with three degrees of freedom. This result suggests that most of the clonal variation is attributable to the low value of C, and that comparisons between the other three means are not likely to be significant.

There is usually no reason for testing blocks, but the size of the block mean square relative to the mean square for error does give an indication of how much precision was gained by blocking. If the block mean square is large (at least two or three times as large as the error mean square) the test is more sensitive than it would have been with complete randomization. If the block mean square is about equal to or only slightly larger than the error mean square, the use of blocks has not improved the precision of the test. The block mean square should not be appreciably smaller than the error mean square. If it is, the method of conducting the study and the computations should be re-examined.

**Assumptions.**—In addition to the assumptions of homogeneous variance and normality, the randomized block design assumes that there is no interaction between treatments and blocks; i.e., that differences among treatments are about the same in all blocks. Because of this assumption, it is not advisable to have blocks that differ greatly—since they may cause an interaction with treatments.

**N.B.:** With only two treatments, the analysis of variance of a randomized block design is equivalent to the  $t$  test of paired replicates. The value

Source	df	SS	MS
Blocks.....	4	30.5	7.625
Clones.....	3	45.0	15.000
Error.....	12	45.5	3.792
Total.....	19	121.0	

$F_{.01, 1, 12} = 3.49$

		COLUMNS				
		1	2	3	4	5
ROWS	1	C	A	B	E	D
	2	A	B	D	C	E
	3	D	C	E	A	B
	4	E	D	C	B	A
	5	B	E	A	D	C

Row, column, and treatment totals					
Row	Σ	Column	Σ	Treatment	Σ
1	80	1	83	A	95
2	82	2	85	B	80
3	80	3	89	C	75
4	82	4	81	D	85
5	91	5	77	E	80
Σ	415	Σ	415	Σ	415
				Σ	16.6

The partitioning of df's, the calculation of sums of squares, and the subsequent analysis follow much the same pattern illustrated previously for randomized blocks.

$$C.T. = \frac{(\sum X)^2}{n} = \frac{415^2}{25} = \frac{172,225}{25} = 6,889.0$$

$$\text{Total SS} = \sum_{24 \text{ df}} X^2 - C.T. = 7,041 - C.T. = \boxed{152.0}$$

$$\text{Row SS} = \frac{\sum (\text{Row totals}^2)}{\text{No. of plots per row}} - C.T. = \frac{34,529}{5} - C.T. = \boxed{16.8}$$

$$\text{Col. SS} = \frac{\sum (\text{Column totals}^2)}{\text{No. of plots per column}} - C.T. = \frac{34,525}{5} - C.T. = \boxed{16.0}$$

$$\text{Species SS} = \frac{\sum (\text{Species totals}^2)}{\text{No. of plots per species}} - C.T. = \frac{34,675}{5} - C.T. = \boxed{46.0}$$

$$\text{Error SS} = \text{Total SS} - \text{Species SS} - \text{Row SS} - \text{Col. SS} = \boxed{73.2}$$

Analysis of variance

Source	df	SS	MS
Rows	4	16.8	4.2
Columns	4	16.0	4.0
Species	4	46.0	11.5
Error	12	73.2	6.1
Total	24	152.0	

$F$  (for species) =  $\frac{11.5}{6.1} = 1.885$

As the computed value of  $F$  is less than the tabular value of  $F$  at the 0.05 level (with 4/12 df's), the differences among species are considered nonsignificant.

The Latin square design can be used whenever there is a two-way heterogeneity that cannot be controlled simply by blocking. In greenhouse studies, distance from a window could be treated as a row effect while distance from the blower or heater might be regarded as a column effect. Though the plots are often physically arranged in rows or columns, this is not required. In testing the use of materials in a manufacturing process where different machines and machine operators will be involved, the variation between machines could be treated as a row effect and the variation due to operator as a column effect.

The Latin square should not be used if an interaction between rows and treatments or columns and treatments is suspected.

Take care!

### Factorial Experiments

In a comparison of corn yields following three rates or levels of nitrogen fertilization it was found that the yields depended on how much phosphorus was used along with the nitrogen. The differences in yield were smaller when no phosphorus was used than when the nitrogen applications were accompanied by 100 pounds per acre of phosphorus. In statistics this situation is referred to as an interaction between nitrogen and phosphorus. Another example: when leaf litter was removed from the forest floor, the catch of pine seedlings was much greater than when the litter was not removed; but for red oak the reverse was true—the seedling catch was lower where litter was removed. Thus, species and litter treatment were interacting.

Interactions are important in the interpretation of study results. In the presence of an interaction between species and litter treatment it obviously makes no sense to talk about the effects of litter removal without specifying the species. The nitrogen-phosphorus interaction means that it may be misleading to recommend a level of nitrogen without mentioning the associated level of phosphorus.

Factorial experiments are aimed at evaluating known or suspected interactions. In these experiments, each factor to be studied is tested at several levels and each level of a factor is tested at all possible combinations of the levels of the other factors. In a planting test involving three species of trees and four methods of preplanting site preparation, each method will be applied to each species, and the total number of treatment

combinations will be 12. In a factorial test of the effects of two nursery treatments on the survival of four species of pine planted by three different methods, there would be  $2 \times 4 \times 3 = 24$  treatment combinations. The method of analysis can be illustrated by a factorial test of the effects of three levels of nitrogen fertilization (0, 100, and 200 pounds per acre) on the growth of three species (A, B, and C) of planted pine. The nine possible treatment combinations were assigned at random to nine plots in each of three blocks. Treatments were evaluated on the basis of average annual height growth in inches per year over a 3-year period. Field layout and plot data were as follows (with subscripts denoting nitrogen levels: 0=0, 1=100, 2=200):

Block I			Block II		
C <sub>2</sub>	B <sub>2</sub>	C <sub>0</sub>	B <sub>2</sub>	C <sub>2</sub>	A <sub>1</sub>
17	18	37	23	20	18
A <sub>0</sub>	B <sub>0</sub>	C <sub>1</sub>	B <sub>1</sub>	A <sub>2</sub>	C <sub>0</sub>
45	24	20	18	14	43
B <sub>1</sub>	A <sub>2</sub>	A <sub>1</sub>	A <sub>0</sub>	C <sub>1</sub>	B <sub>0</sub>
21	24	17	40	25	35

Block III		
C <sub>1</sub>	C <sub>2</sub>	A <sub>0</sub>
19	21	37
B <sub>1</sub>	A <sub>2</sub>	A <sub>1</sub>
19	17	28
B <sub>0</sub>	C <sub>0</sub>	B <sub>2</sub>
29	39	15

The preliminary analysis of the nine combinations (temporarily ignoring their factorial nature) is made just as though this were a straight randomized block design (which is exactly what it is). (See table, p. 41.)

Sums of squares

$$C.T. = \frac{\left(\sum X\right)^2}{27} = \frac{683^2}{27} = 17,277.3704$$

$$\begin{aligned} \text{Total SS} &= \sum X^2 - C.T. = (45^2 + 17^2 + \dots + 21^2) - C.T. \\ &= \frac{2,275.6296}{26 \text{ df}} \end{aligned}$$

$$\begin{aligned} \text{Block SS} &= \frac{\sum (\text{Block totals}^2)}{\text{No. of plots per block}} - C.T. \\ &= \frac{(223^2 + 236^2 + 224^2)}{9} - C.T. = \frac{11,6296}{2 \text{ df}} \end{aligned}$$

### Summary of plot data

Species	Nitrogen level	I	II	III	Nitrogen subtotals	Species totals
A	0	45	40	37	122	240
	1	17	18	28		
	2	24	14	17		
	Block subtotals	86	72	82		
B	0	24	35	29	88	202
	1	21	18	19		
	2	18	23	15		
	Block subtotals	63	76	63		
C	0	37	43	39	119	241
	1	20	25	19		
	2	17	20	21		
	Block subtotals	74	88	79		
All species	Totals	106	118	105	329	Grand total
		58	61	66	185	
		59	57	53	169	
		223	236	224		683

$$\begin{aligned} \text{Treatment SS} &= \frac{\sum (\text{Treatment totals}^2)}{\text{No. of plots per treatment}} - C.T. \\ &= \frac{(122^2 + 88^2 + \dots + 58^2)}{3} - C.T. = \frac{1,970.2963}{2 \text{ df}} \end{aligned}$$

$$\text{Error SS} = \text{Total SS} - \text{Treatment SS} - \text{Block SS} = \frac{293.7037}{2 \text{ df}}$$

Tabulating these in the usual form:

Source	df	SS	MS
Blocks	2	11,6296	5,8148
Treatments	8	1,970.2963	246.2870
Error	16	293.7037	18.3565
Totals	26	2,275.6296	

Testing treatments:  $F_{8/16 \text{ df}} = \frac{246.2870}{18.3565} = 13.417$ , significant at the 0.01 level.

The next step is to analyze the components of the treatment variability. How do the species compare? What is the effect of fertilization? And does fertilization affect all species the same way (i.e., is there a species-nitrogen interaction)? To answer these questions we have to partition the degrees of freedom and sums of squares associated with treatments. This is easily done by summarizing the data for the nine combinations in a two-way table.

Species	0	Nitrogen levels	2	Totals
A	122	63	55	240
B	88	58	56	202
C	119	64	58	241
Totals	329	185	169	683

The nine individual values will be recognized as those that entered into the calculation of the treatment SS. Keeping in mind that each entry in the body of the table is the sum of three plot values, and that the species and nitrogen totals are each the sum of 9 plots, the sums of squares for species, nitrogen, and the species-nitrogen interaction can be computed as follows:

Treatment SS = 1,970.2963 (as previously calculated)

$$\text{Species SS} = \frac{\sum (\text{species totals})^2}{\text{No. of plots per species}} - C.T.$$

$$= \frac{(240^2 + 202^2 + 241^2)}{9} - C.T.$$

$$= \frac{156,485}{9} - C.T. = \boxed{109.8518}$$

$$\text{Nitrogen SS} = \frac{\sum (\text{Nitrogen totals})^2}{\text{No. of plots per level of nitrogen}} - C.T.$$

$$= \frac{(329^2 + 185^2 + 169^2)}{9} - C.T. = \frac{171,027}{9} - C.T.$$

$$= \boxed{1,725.6296}$$

$$\text{Species-nitrogen interaction SS} = \text{Treatment SS} - \text{Species SS} - \text{Nitrogen SS} = \boxed{134.8149}$$

The analysis now becomes:

Source	df	SS	MS	F
Blocks	2	11.6296	5.8148	—
Treatments	8	1,970.2963	246.2870	13.417**
Species	12	109.8518	54.9259	2.992NS
Nitrogen	12	1,725.6296	862.8148	47.003**
Species-nitrogen	14	134.8149	33.7037	1.836NS
Error	16	293.7037	18.3565	—
Total	26	2,275.6296	—	—

<sup>1</sup> Offset figures are a partitioning of the df's and sum of squares for Treatments, and are therefore not included in the Total at the bottom of the table.

The degrees of freedom for simple interactions can be obtained in two ways. The first way is by subtracting the df's associated with the component factors (in this case two for species and two for nitrogen levels) from the df's associated with all possible treatment combinations (eight in this case). The second way is to calculate the interaction df's as the product of the component factor df's (in this case 2X2=4). Do it both ways as a check.

The *F* values for species, nitrogen, and the species-nitrogen interaction are calculated by dividing their mean squares by the mean square for error. In the above tabulation, last column, NS indicates nonsignificant and \*\* means significant at the 0.01 level.

The analysis indicates a significant difference among levels of nitrogen, but no difference between species and no species-nitrogen interaction.

As before, a prespecified comparison among treatment means can be tested by breaking out the sum of squares associated with that comparison. To illustrate the computations, we will test nitrogen versus no nitrogen and also 100 pounds versus 200 pounds of nitrogen.

$$\text{Nitrogen vs. no nitrogen SS} = \frac{9 \left[ 2 \left( \frac{329}{9} \right) - 1 \left( \frac{185}{9} \right) - 1 \left( \frac{169}{9} \right) \right]^2}{(2^2 + 1^2 + 1^2)}$$

$$= \frac{[2(329) - 185 - 169]^2}{9(6)} = \boxed{1,711.4074}$$

In the numerator the mean for the zero level of nitrogen is multiplied by 2 to give it equal weight with the mean of levels 1 and 2 with which it is compared. The 9 is the number of plots on which each mean is based. The  $(2^2 + 1^2 + 1^2)$  in the denominator is the sum of squares of the coefficients used in the numerator.

$$100 \text{ vs. } 200 \text{ pounds SS} = \frac{9 \left[ 1 \left( \frac{185}{9} \right) - 1 \left( \frac{169}{9} \right) \right]^2}{(1^2 + 1^2)} = \frac{[185 - 169]^2}{9(2)} = \boxed{14.2222}$$

Note that these two sums of squares (1,711.4074 and 14.2222), each with 1 df, add up to the sum of squares for nitrogen (1,725.6296) with 2 df's. This additive characteristic holds true only if the individual comparisons selected are orthogonal (i.e., independent). When the number of observations is the same for all treatments, the orthogonality of any two comparisons can be checked in the following manner: First, tabulate the coefficients and check to see that for each comparison the coefficients sum to zero.

Comparison	Nitrogen level		Sum
	0	1	
2N <sub>0</sub> vs. N <sub>1</sub> +N <sub>2</sub>	2	-	0
N <sub>1</sub> vs. N <sub>2</sub>	0	+	0
Product of coefficients	0	-	+



Then for two comparisons to be orthogonal the sum of the products of corresponding coefficients must be zero. Any sum of squares can be partitioned in a similar manner, with the number of possible orthogonal individual df comparisons being equal to the total number of degrees of freedom with which the sum of squares is associated.

The sum of squares for species can also be partitioned into two orthogonal single df comparisons. If the comparisons were specified before the data were examined, we might make single df tests of the difference between B and the average of A and C and also of the difference between A and C. The method is the same as that illustrated in the comparison of nitrogen treatments. The calculations are as follows:

$$2B \text{ vs. } (A+C) \text{ SS} = \frac{9 \left[ 1 \left( \frac{240}{9} \right) + 1 \left( \frac{241}{9} \right) - 2 \left( \frac{202}{9} \right) \right]^2}{1 \text{ df}} = \frac{[240 + 241 - 2(202)]^2}{9(6)} = \frac{109.7963}{1}$$

$$A \text{ vs. } C \text{ SS} = \frac{9 \left[ 1 \left( \frac{241}{9} \right) - 1 \left( \frac{240}{9} \right) \right]^2}{1 \text{ df}} = \frac{[241 - 240]^2}{9(2)} = \frac{0.0555}{1}$$

These comparisons are orthogonal, so that the sums of squares each with one df add up to the species SS with two df's.

Note that in computing the sums of squares for the single-degree-of-freedom comparisons, the equations have been restated in terms of treatment totals rather than means. This often simplifies the computations and reduces the errors due to rounding.

With the partitioning the analysis has become:

Source	df	SS	MS	F
Blocks.....	2	11.6296	5.8148	2.992 <sup>NS</sup>
Species.....	2	109.8518	54.9259	5.981*
2B vs. (A+C).....	1	109.7963	109.7963	
A vs. C.....	1	.0555	.0555	
Nitrogen.....	2	1,725.6296	862.8148	47.003**
2N <sub>0</sub> vs. (N <sub>1</sub> +N <sub>2</sub> ).....	1	1,711.4074	1,711.4074	93.232**
N <sub>1</sub> vs. N <sub>2</sub> .....	1	14.2222	14.2222	
Species X nitrogen interaction.....	4	134.8149	33.7037	1.836 <sup>NS</sup>
Error.....	16	293.7037	18.3565	
Total.....	26	2,275.6296		

We conclude that species B is poorer than A or C and that there is no difference in growth between A and C. We also conclude that nitrogen adversely affected growth and that 100 pounds was about as bad as 200 pounds. The nitrogen effect was about the same for all species (i.e., no interaction).

It is worth repeating that the comparisons to be made in an analysis should, whenever possible, be planned and specified prior to an examination of the data. A good procedure is to outline the analysis, putting in all the items that are to appear in the first two columns (source and df) of the table. In the above tabulation, last column, \* means significant

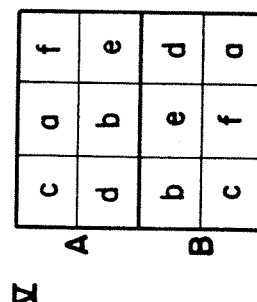
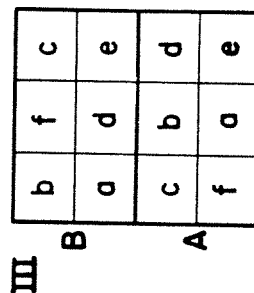
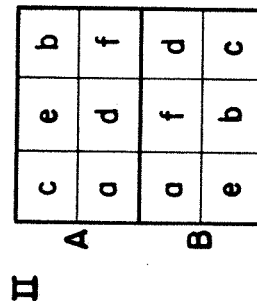
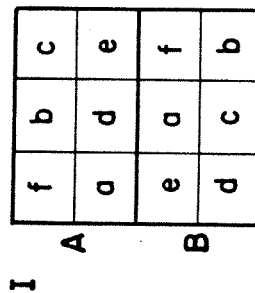
at the 0.05 level. As in the previous table, \*\* means significant at the 0.01 level, and NS means nonsignificant.

The factorial experiment, it will be noted, is not an experimental design. It is, instead, a way of selecting treatments; given two or more factors each at two or more levels, the treatments are all possible combinations of the levels of each factor. If we have three factors with the first at four levels, the second at two levels, and the third at three levels, we will have  $4 \times 2 \times 3 = 24$  factorial combinations or treatments. Factorial experiments may be conducted in any of the standard designs. The randomized block and split plot design are the most common for factorial experiments in forest research.

### The Split Plot Design

When two or more types of treatment are applied in factorial combinations, it may be that one type can be applied on relatively small plots while the other type is best applied to larger plots. Rather than make all plots of the size needed for the second type, a split-plot design can be employed. In this design, the major (large-plot) treatments are applied to a number of plots with replication accomplished through any of the common designs (such as complete randomization, randomized blocks, Latin square). Each major plot is then split into a number of subplots, equal to the number of minor (small-plot) treatments. Minor treatments are assigned at random to subplots within each major plot.

As an example, a test was to be made of direct seeding of loblolly pine at six different dates, on burned and unburned seedbeds. To get typical burn effects, major plots 6 acres in size were selected. There were to be four replications of major treatments in randomized blocks. Each major plot was divided into six 1-acre subplots for seeding at six dates. The field layout was somewhat as follows (blocks denoted by Roman numerals, burning treatment by capital letters, date of seeding by small letters):



One pound of seed was sowed on each 1-acre subplot. Seeding counts were made at the end of the first growing season. Results were as follows:

Summary of seedlings per acre

Date	I		II		III		IV		Date subtotals		Date totals
	A	B	A	B	A	B	A	B	A	B	
a	900	880	810	1,100	760	960	1,040	1,040	3,510	3,980	7,490
b	880	1,050	1,170	1,240	1,060	1,110	910	1,120	4,020	4,520	8,540
c	1,530	1,140	1,160	1,270	1,390	1,320	1,540	1,080	5,620	4,810	10,430
d	1,970	1,360	1,890	1,510	1,820	1,490	2,140	1,270	7,820	5,630	13,450
e	1,960	1,270	1,570	1,350	1,310	1,500	1,480	1,450	6,420	5,600	12,020
f	830	150	420	390	570	420	760	270	2,580	1,220	3,800
Major plot totals	8,070	5,850	7,120	6,880	6,910	6,800	7,870	6,230	29,970	25,760	
Block totals	13,920		14,000		13,710		14,100				55,730

Calculations.—The correction term and total sum of squares are calculated using the 48 subplot values.

$$C.T. = \frac{(\text{Grand total of all subplots})^2}{48} = \frac{55,730^2}{48} = 64,704,852$$

$$\text{Total SS} = \sum_{48 \text{ df}} (\text{Subplot values})^2 - C.T. = 9,339,648$$

$$= (900^2 + 880^2 + \dots + 270^2) - C.T. = 9,339,648$$

Before partitioning the total sum of squares into its components, it may be instructive to ignore subplots for the moment, and examine the major plot phase of the study. The major phase can be viewed as a straight randomized block design with two burning treatments in each of four blocks. The analysis would be:

Source	df
Blocks	3
Burning	1
Error (major plots)	3
Major plots	7

Now, looking at the subplots, we can think of the major plots as blocks. From this standpoint we would have a randomized block design with six dates of treatment in each of eight blocks (major plots) for which the analysis is:

Source	df
Major-plots	7
Dates	5
Remainder	35
Subplots (= Total)	47

In this analysis, the remainder is made up of two components. One of these is the burning-date interaction, with five df's. The rest, with 30 df's, is called the subplot error. Thus, the complete breakdown of the split-plot design is:

Source	df
Blocks	3
Burning	1
Major plot error	3
Date	5
Burning X date	5
Subplot error	30
Total	47

The various sums of squares are obtained in an analogous manner. We first compute

$$\text{Major plot SS} = \frac{\sum (\text{Major plot totals})^2}{\text{Number of subplots per major plot}} - C.T.$$

$$= \frac{8,070^2 + \dots + 6,230^2}{6} - C.T. = 647,498$$

$$\text{Block SS} = \frac{\sum (\text{Block totals})^2}{\text{Subplots per block}} - C.T.$$

$$= \frac{13,920^2 + \dots + 14,100^2}{12} - C.T. = 6,856$$

$$\text{Burning SS} = \frac{\sum (\text{Burning treatment totals})^2}{\text{Subplots per burning treatment}} - C.T.$$

$$= \frac{29,970^2 + 25,760^2}{24} - C.T. = 369,252$$

$$\text{Major-plot error SS} = \text{Major plot SS} - \text{Block SS}$$

$$= \text{Burning SS} = 271,390$$

$$\text{Subplot SS} = \text{Total SS} - \text{Major plot SS} = 8,692,150$$

$$\text{Date SS} = \frac{\sum (\text{Date totals})^2}{\text{Subplots per date}} - C.T.$$

$$= \frac{7,490^2 + \dots + 3,800^2}{8} - C.T. = 7,500,086$$

Date-burning SS:

To get the sum of squares for the interaction between date and burning we resort to a factorial experiment device—the two-way table of the treatment combination totals.

Burning	a	b	c	Date	d	e	f	Burning subtotals
A	3,510	4,020	5,620	7,820	6,420	2,580		29,970
B	3,980	4,520	4,810	5,630	5,600	1,220		25,760
Date subtotals	7,490	8,540	10,430	13,450	12,020	3,800		55,730

$$\text{Date-burning subclass SS} = \frac{1}{11} \sum (\text{Date-burning combination totals}^2) - \text{C.T.} = \frac{3,510^2 + \dots + 1,220^2}{4} - \text{C.T.} = 8,555,723$$

$$\text{Date-burning interaction SS} = \frac{\text{Date-burning subclass SS}}{5 \text{ df}} - \frac{\text{Date SS} - \text{Burning SS}}{5 \text{ df}} = \frac{686,385}{5 \text{ df}}$$

$$\text{Subplot error SS} = \frac{\text{Subplot SS} - \text{Date SS}}{30 \text{ df}} - \frac{\text{Date-burning interaction SS}}{5 \text{ df}} = \frac{505,679}{30 \text{ df}}$$

Thus the completed analysis table is

Source	df	SS	MS
Blocks	3	6,856	
Burning	1	369,252	369,252
Major-plot error	3	271,390	90,463
Date	5	7,500,086	1,500,017
Date-burning	5	686,385	137,277
Subplot error	30	505,679	16,856
Total	47	9,339,648	

The  $F$  test for burning is

$$F_{1/3 \text{ df}} = \frac{\text{Burning MS}}{\text{Major-plot error MS}} = \frac{369,252}{90,463} = 4.082, \text{ not significant at the 0.05 level.}$$

For dates,

$$F_{5/30 \text{ df}} = \frac{\text{Date MS}}{\text{Subplot error MS}} = \frac{1,500,017}{16,856} = 88.99, \text{ significant at the 0.01 level.}$$

And for the date-burning interaction,

$$F_{5/30 \text{ df}} = \frac{\text{Date-burning MS}}{\text{Subplot error MS}} = \frac{137,277}{16,856} = 8.14, \text{ significant at the 0.01 level.}$$

Note that the major-plot error is used to test the sources above the dashed line while the subplot error is used for the sources below the line. Because the subplot error is a measure of random variation within major

plots it will usually be smaller than the major-plot error, which is a measure of the random variation between major plots. In addition to being smaller, the subplot error will generally have more degrees of freedom than the major-plot error, and for these reasons the sources below the dashed line will usually be tested with greater sensitivity than the sources above the line. This fact is important in planning a split-plot experiment; the designer should try to get the items of greatest interest below the line rather than above.

Rarely will the major-plot error be appreciably smaller than the subplot error. If it is, the conduct of the study and the computations should be carefully examined.

Subplots can also be split.—If desired, the subplots can also be split for a third level of treatment, producing a split-split-plot design. The calculations follow the same general pattern but are more involved. A split-split-plot design has three separate error terms.

Comparisons among means in a split-plot design.—For comparisons among major- or subplot treatments,  $F$  tests with a single degree of freedom may be made in the usual manner. Comparisons among major-plot treatments should be tested against the major-plot error mean square, while subplot treatment comparisons are tested against the subplot error. In addition, it is sometimes desirable to compare the means of two treatment combinations. This can get tricky, for the variation among such means may contain more than one source of error. A few of the more common cases are discussed below.

In general, the  $t$  test for comparing two equally replicated treatment means is

$$t = \frac{\text{Mean difference}}{\text{Standard error of the mean difference}} = \frac{\bar{D}}{s\bar{D}}$$

1. For the difference between two major-treatment means:

$$s\bar{D} = \sqrt{\frac{2(\text{Major-plot error MS})}{(m)(R)}}; t \text{ has df equal to the df for the major-plot error.}$$

where:  $R$  = Number of replications of major treatments.  
 $m$  = Number of subplots per major plot.

2. For the difference between two minor-treatment means:

$$s\bar{D} = \sqrt{\frac{2(\text{Subplot error MS})}{(R)(M)}}; t \text{ has df equal to the df for subplot error.}$$

where:  $M$  = Number of major-plot treatments.

3. For the difference between two minor treatments within a single major treatment:

$$s\bar{D} = \sqrt{\frac{2(\text{Subplot error MS})}{R}}; \text{ df for } t = \text{df for the subplot error}$$

4. For the difference between the means of two major treatments at a single level of a minor treatment, or between the means of two major treatments at different levels of a minor treatment:

$$s\bar{D} = \sqrt{\frac{2 \left[ \frac{(m-1)(\text{Subplot error MS}) + \text{Major-plot error MS}}{(m)(R)} \right]}}{m}}$$

In this case,  $t$  will not follow the  $t$  distribution. A close approximation to the value of  $t$  required for significance at the  $\alpha$  level is given by

$$t = \frac{(m-1)(\text{Subplot error MS})t_m + (\text{Major-plot error MS})t_M}{(m-1)(\text{Subplot error MS}) + (\text{Major-plot error MS})}$$

where:

$t_m$  = Tabular value of  $t$  at the  $\alpha$  level for  $df$  equal to the  $df$  for the subplot error.

$t_M$  = Tabular value of  $t$  at the  $\alpha$  level for  $df$  equal to the  $df$  for the major-plot error.

Other symbols are as previously defined.

### Missing Plots

A mathematician who had developed a complex electronic computer program for analyzing a wide variety of experimental designs was asked how he handled missing plots. His disdainful reply was, "We tell our research workers not to have missing plots."

This is good advice. But it is sometimes hard to follow, and particularly so in forest research, where close control over experimental material is difficult and studies may run for several years.

The likelihood of plots being lost during the course of a study should be considered when selecting an experimental design. Lost plots are least troublesome in the simple designs. For this reason, complete randomization and randomized blocks may be preferable to the more intricate designs when missing data can be expected.

In the complete randomization design, loss of one or more plots causes no computational difficulties. The analysis is made as though the missing plots never existed. Of course, a degree of freedom will be lost from the total and error terms for each missing plot and the sensitivity of the test will be reduced. If missing plots are likely, the number of replications should be increased accordingly.

In the randomized block design, completion of the analysis will usually require an estimate of the values for the missing plots. A single missing value can be estimated by

$$X = \frac{bB + tT - G}{(b-1)(t-1)}$$

where:  $b$  = Number of blocks

$t$  = Number of treatments

$B$  = Total of all other units in the block with a missing plot

$T$  = Total of all other units that received the same treatment as the missing plot

$G$  = Total of all observed units

If more than one plot is missing, the customary procedure is to insert guessed values for all but one of the missing units, which is then estimated by the above formula. This estimate is used in obtaining an estimated value for one of the guessed plots, and so on through each missing unit. Then the process is repeated with the first estimates replacing the guessed values. The cycle should be repeated until the new approximations differ little from the previous estimates.

The estimated values are now applied in the usual analysis-of-variance calculations. For each missing unit one degree of freedom is deducted from the total and from the error term.

A similar procedure is used with the Latin square design, but the formula for a missing plot is

$$X = \frac{r(R+C+T) - 2G}{(r-1)(r-2)}$$

where:  $r$  = Number of rows

$R$  = Total of all observed units in the row with the missing plot

$C$  = Total of all observed units in the column with the missing plot

$T$  = Total of all observed units in the missing plot treatment

$G$  = Grand total of all observed units

With the split-plot design, missing plots can cause trouble. A single missing subplot value can be estimated by the equation

$$X = \frac{rP + m(T_i) - (T_i)}{(r-1)(m-1)}$$

where:  $r$  = Number of replications of major-plot treatments

$P$  = Total of all observed subplots in the major plot having a missing subplot

$m$  = Number of subplot treatments

$T_i$  = Total of all subplots having the same treatment combination as the missing unit

$T_i$  = Total of all subplots having the same major-plot treatment as the missing unit

For more than one missing subplot the iterative process described for randomized blocks must be used. In the analysis, one  $df$  will be deducted from the total and subplot error terms for each missing subplot.

When data for missing plots are estimated, the treatment mean square for all designs is biased upwards. If the proportion of missing plots is small, the bias can usually be ignored. Where the proportion is large, adjustments can be made as described in the standard references on experimental designs.

## REGRESSION

### Simple Linear Regression

A forester had an idea that he could tell how well a loblolly pine was growing from the volume of the crown. Very simple: big crown—good growth, small crown—poor growth. But he couldn't say how big and how good, or how small and how poor. What he needed was regression analysis: it would enable him to express a relationship between tree growth and crown volume in an equation. Given a certain crown volume, he could use the equation to predict what the tree growth was.

To gather data, he ran parallel survey lines across a large tract that was representative of the area in which he was interested. The lines were 5 chains apart. At each 2-chain mark along the lines, he measured the nearest loblolly pine of at least 5.6 inches d.b.h. for crown volume and basal area growth over the past 10 years.

A portion of the data is printed below to illustrate the methods of calculation. Crown volume in hundreds of cubic feet is labeled  $X$  and basal area growth in square feet is labeled  $Y$ . Now, what can we tell the forester about the relationship?

X Crown volume	Y Growth	X Crown volume	Y Growth	X Crown volume	Y Growth
22	.36	53	.47	51	.41
6	.09	70	.55	75	.66
93	.67	5	.07	6	.18
62	.44	90	.69	20	.21
84	.72	46	.42	36	.29
14	.24	36	.39	50	.56
52	.33	14	.09	9	.13
69	.61	60	.54	2	.10
104	.66	103	.74	21	.18
100	.80	43	.64	17	.17
41	.47	22	.50	87	.63
85	.60	75	.39	97	.66
90	.51	29	.30	33	.18
27	.14	76	.61	20	.06
18	.32	20	.29	96	.58
48	.21	29	.38	61	.42
37	.54	50	.53		
67	.70	59	.58		
56	.67	70	.62		
31	.42	81	.66		
17	.39	93	.69		
7	.25	99	.71		
2	.06	14	.14		
Totals.....					26.62
Means (n=62).....					0.42935

Often, the first step is to plot the field data on coordinate paper (fig. 1). This is done to provide some visual evidence of whether the two variables are related. If there is a simple relationship, the plotted points will tend to form a pattern (a straight line or curve). If the relationship is very strong, the pattern will generally be distinct. If the relationship is weak, the points will be more spread out and the pattern less definite. If the points appear to fall pretty much at random, there may be no simple relationship or one that is so very poor as to make it a waste of time to fit any regression.

The type of pattern (straight line, parabolic curve, exponential curve, etc.) will influence the regression model to be fitted. In this particular case, we will assume a simple straight-line relationship.

After selecting the model to be fitted, the next step will be to calculate the corrected sums of squares and products. In the following equations, capital letters indicate uncorrected values of the variables; lower-case letters will be used for the corrected values ( $y = Y - \bar{Y}$ ).

$$\begin{aligned} \text{The corrected sum of squares for } Y: \sum y^2 &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\ &= \frac{(0.36^2 + 0.09^2 + \dots + 0.42^2)}{62} \\ &= \frac{2.7826}{62} \end{aligned}$$

$$\begin{aligned} \text{The corrected sum of squares for } X: \sum x^2 &= \sum X^2 - \frac{(\sum X)^2}{n} \\ &= \frac{(22^2 + 6^2 + \dots + 61^2)}{62} - \frac{3,050^2}{62} \\ &= \frac{59,397.6775}{62} \end{aligned}$$

$$\begin{aligned} \text{The corrected sum of products:} \quad \sum xy &= \sum(XY) - \frac{(\sum X)(\sum Y)}{n} \\ &= \frac{[(22)(.36) + (6)(.09) + \dots + (61)(.42)]}{62} - \frac{(3,050)(26.62)}{62} \\ &= \frac{354.1477}{62} \end{aligned}$$

The general form of equation for a straight line is  $Y = a + bX$

In this equation,  $a$  and  $b$  are constants or regression coefficients that must be estimated. According to the principle of least squares, the best estimates of these coefficients are:

$$\begin{aligned} b &= \frac{\sum xy}{\sum x^2} = \frac{354.1477}{59,397.6775} = 0.005962 \\ a &= \bar{Y} - b\bar{X} = 0.42935 - (0.005962)(49.1935) = 0.13606 \end{aligned}$$

Substituting these estimates in the general equation gives

$$\hat{Y} = 0.13606 + 0.005962X$$

where  $\hat{Y}$  is used to indicate that we are dealing with an estimated value of  $Y$ .

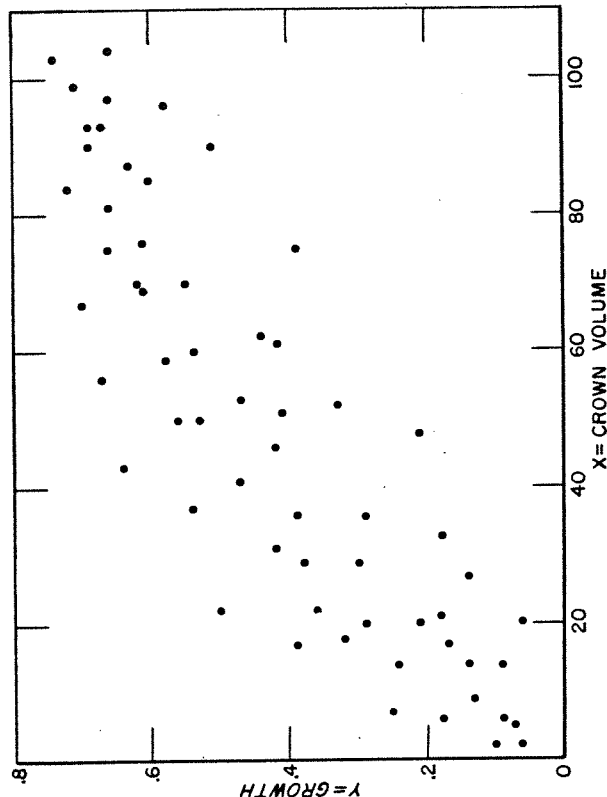


FIGURE 1.—Plotting of growth (Y) over crown volume (X).

With this equation we can estimate the basal area growth for the past 10 years ( $\bar{Y}$ ) from the measurements of the crown volume  $X$ .

Because  $Y$  is estimated from a known value of  $X$ , it is called the dependent variable and  $X$  the independent variable. In plotting on graph paper, the values of  $Y$  are usually (purely by convention) plotted along the vertical axis (ordinate) and the values of  $X$  along the horizontal axis (abscissa).

### How Well Does the Regression Line Fit the Data?

A regression line can be thought of as a moving average. It gives an average value of  $Y$  associated with a particular value of  $X$ . Of course, some values of  $Y$  will be above the regression line (or moving average) and some below, just as some values of  $Y$  are above or below the general average of  $Y$ .

The corrected sum of squares for  $Y$  (i.e.,  $\sum y^2$ ) estimates the amount of variation of individual values of  $Y$  about the mean value of  $Y$ . A regression equation is a statement that part of the observed variation in  $Y$  (estimated by  $\sum y^2$ ) is associated with the relationship of  $Y$  to  $X$ . The amount of variation in  $Y$  that is associated with the regression on  $X$  is called the reduction or regression sum of squares.

$$\text{Reduction SS} = \frac{(\sum xy)^2}{\sum x^2} = \frac{(354.1477)^2}{(59,397.6775)} = 2.1115$$

As noted above, the total variation in  $Y$  is estimated by  $\sum y^2 = 2.7826$  (as previously calculated).

The part of the total variation in  $Y$  that is not associated with the regression is called the residual sum of squares. It is calculated by

$$\text{Residual SS} = \sum y^2 - \text{Reduction SS} = 2.7826 - 2.1115 = 0.6711$$

In analysis of variance we used the unexplained variation as a standard for testing the amount of variation attributable to treatments. We can do the same in regression. What's more, the familiar  $F$  test will serve.

Source of variation	df <sup>1</sup>	SS	MS <sup>2</sup>
Due to regression $\left[ = \frac{(\sum xy)^2}{\sum x^2} \right]$	1	2.1115	2.1115
Residual (i.e., unexplained)	60	0.6711	0.01118
Total ( $= \sum y^2$ )	61	2.7826	

<sup>1</sup> As there are 62 values of  $Y$ , the total sum of squares has 61 df. The regression of  $Y$  on  $X$  has one df. The residual df are obtained by subtraction.

<sup>2</sup> MS is, as always  $= \frac{SS}{df}$

The regression is tested by

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} = \frac{2.1115}{0.01118} = 188.86$$

As the calculated  $F$  is much greater than tabular  $F_{.01}$  with 1/60 df, the regression is deemed significant at the 0.01 level.

Before we fitted a regression line to the data,  $Y$  had a certain amount of variation about its mean ( $\bar{Y}$ ). Fitting the regression was, in effect, an attempt to explain part of this variation by the linear association of  $Y$  with  $X$ . But even after the line had been fitted, some variation was unexplained—that of  $Y$  about the regression line. When we tested the regression line above, we merely showed that the part of the variation in  $Y$  that is explained by the fitted line is significantly greater than the part that the line left unexplained. The test did not show that the line we fitted gives the best possible description of the data (a curved line might be even better). Nor does it mean that we have found the true mathematical relationship between the two variables. There is a dangerous tendency to ascribe more meaning to a fitted regression than is warranted.

It might be noted that the residual sum of squares is equal to the sum of the squared deviations of the observed values of  $Y$  from the regression line. That is,

$$\text{Residual SS} = \sum (Y - \hat{Y})^2 = \sum (Y - a - bX)^2$$

The principle of least squares says that the best estimates of the regression coefficients ( $a$  and  $b$ ) are those that make this sum of squares a minimum.

### Coefficient of Determination

As a measure of how well a regression fits the sample data, we can compute the proportion of the total variation in  $Y$  that is associated with the regression. This ratio is sometimes called the coefficient of determination.

$$\begin{aligned} \text{Coefficient of determination} &= \frac{\text{Reduction SS}}{\text{Total SS}} \\ &= \frac{2.1115}{2.7826} = 0.758823 \end{aligned}$$

When someone says, "76 percent of the variation in  $Y$  was associated with  $X$ ," he means that the coefficient of determination was 0.76.

The coefficient of determination is equal to the square of the correlation coefficient.

$$\frac{\text{Reduction SS}}{\text{Total SS}} = \frac{(\sum xy)^2 / \sum x^2}{\sum y^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)} = r^2$$

In fact, most present-day users of regression refer to  $r^2$  values rather than to coefficients of determination.

In the older literature,  $1 - r^2$  is sometimes called the coefficient of non-determination, and  $\sqrt{1 - r^2}$  has been called the alienation index.

### Confidence Intervals

Since it is based on sample data, a regression equation is subject to sample variation. Confidence limits on the regression line can be obtained by specifying several values over the range of  $X$  and computing

$$\text{Confidence limits} = \hat{Y} \pm \sqrt{(\text{Residual MS}) \left( \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right)}$$

Where  $X_0$  = a selected value of  $X$ , and  
 Degrees of freedom for  $t$  =  $df$  for residual MS.  
 In the example we had:

$$\begin{aligned} \bar{Y} &= 0.13606 + 0.005962X \\ \text{Residual MS} &= 0.01118 \text{ with } 60 \text{ df's.} \\ n &= 62 \\ \bar{X} &= 49.1935 \\ \Sigma x^2 &= 59,397.6775. \end{aligned}$$

So, if we pick  $X_0 = 28$  we have  $\bar{Y} = 0.303$ , and  
 95-percent confidence limits

$$\begin{aligned} &= 0.303 \pm 2.000 \sqrt{(0.01118) \left( \frac{1}{62} + \frac{(28 - 49.1935)^2}{59,397.6775} \right)} \\ &= 0.270 \text{ to } 0.336 \end{aligned}$$

For other values of  $X_0$  we would get:

$X_0$	$\bar{Y}$	95 percent limits	
		Lower	Upper
8	0.184	0.139	0.229
49.1935	.429	.402	.456
70	.553	.521	.585
90	.673	.629	.717

In figure 2 these points have been plotted and connected by smooth curves.

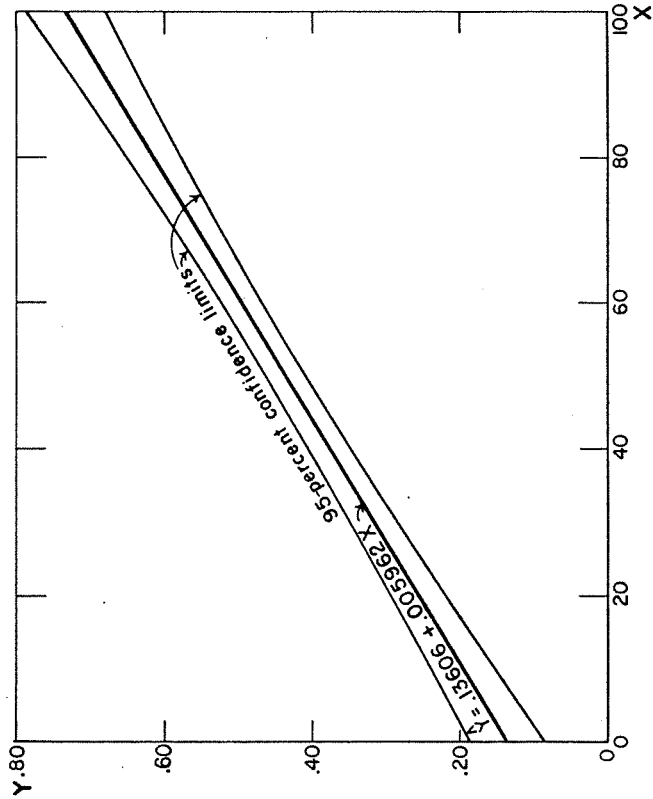


FIGURE 2.—Confidence limits for the regression of  $Y$  on  $X$ .

It should be especially noted that these are confidence limits on the regression of  $Y$  on  $X$ . They indicate the limits within which the true mean of  $Y$  for a given  $X$  will lie unless a one-in-twenty chance has occurred. The limits do not apply to a single predicted value of  $Y$ . The limits within which a single  $Y$  might lie are given by

$$\bar{Y} \pm t \sqrt{(\text{Residual MS}) \left( 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x^2} \right)}$$

*Assumptions.*—In addition to assuming that the relationship of  $Y$  to  $X$  is linear, the above method of fitting assumes that the variance of  $Y$  about the regression line is the same at all levels of  $X$  (the assumption of homogeneous variance or homoscedasticity—if you want to impress your friends). The fitting does not assume nor does it require that the variation of  $Y$  about the regression line follows the normal distribution. However, the  $F$  test does assume normality, and so does the use of  $t$  for the computation of confidence limits.

There is also an assumption of independence of the errors (departures from regression) of the sample observations. The validity of this assumption is best insured by selecting the sample units at random. The requirement of independence may not be met if successive observations are made on a single unit or if the units are observed in clusters. For example, a series of observations of tree diameter made by means of a growth band would probably lack independence.

Selecting the sample units so as to get a particular distribution of the  $X$  values does not violate any of the regression assumptions, provided the  $Y$  values are a random sample of all  $Y$ 's associated with the selected values of  $X$ . Spreading the sample over a wide range of  $X$  values will usually increase the precision with which the regression coefficients are estimated. This device must be used with caution however, for if the  $Y$  values are not random, the regression coefficients and residual mean squares may be improperly estimated.

## Multiple Regression

It frequently happens that a variable ( $Y$ ) in which we are interested is related to more than one independent variable. If this relationship can be estimated, it may enable us to make more precise predictions of the dependent variable than would be possible by a simple linear regression. This brings us up against multiple regression, which is a little more work but no more complicated than a simple linear regression.

The calculation methods can be illustrated with the following set of hypothetical data from a study relating the growth of even-aged loblolly-shortleaf pine stands to the total basal area ( $X_1$ ), the percentage of the basal area in loblolly pine ( $X_2$ ), and loblolly pine site index ( $X_3$ ).

Y	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>
65	41	79	75
78	90	48	83
85	53	67	74
50	42	52	61
55	57	52	59
59	32	82	73
82	71	80	72
66	60	65	66
113	98	96	99
86	80	81	90
104	101	78	86
92	100	59	88
96	84	84	93
65	72	48	70
81	55	93	85
77	77	68	71
83	98	51	84
97	95	82	81
90	90	70	78
87	93	61	89
74	45	96	81
70	50	80	77
75	60	76	70
75	68	74	76
93	75	96	85
76	82	58	80
71	72	58	68
61	46	69	65
Sums	2,206	1,987	2,179
Means (n=28)	78.7857	70.9643	77.8214

With this data we would like to fit an equation of the form

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3$$

According to the principle of least squares, the best estimates of the X coefficients can be obtained by solving the set of least squares normal equations.

$$\begin{aligned} b_1 \text{ equation: } & (\sum x_1^2)b_1 + (\sum x_1x_2)b_2 + (\sum x_1x_3)b_3 = \sum x_1y \\ b_2 \text{ equation: } & (\sum x_1x_2)b_1 + (\sum x_2^2)b_2 + (\sum x_2x_3)b_3 = \sum x_2y \\ b_3 \text{ equation: } & (\sum x_1x_3)b_1 + (\sum x_2x_3)b_2 + (\sum x_3^2)b_3 = \sum x_3y \end{aligned}$$

$$\text{where: } \sum x_ix_j = \sum X_iX_j - \frac{(\sum X_i)(\sum X_j)}{n}$$

Having solved for the X coefficients ( $b_1$ ,  $b_2$ , and  $b_3$ ), we obtain the constant term by solving

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - b_3\bar{X}_3$$

Derivation of the least squares normal equations requires a knowledge of differential calculus. However, for the general linear model with a

constant term

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

the normal equations can be written quite mechanically once their pattern has been recognized. Every term in the first row contains an  $x_1$ , every term in the second row an  $x_2$ , and so forth down to the  $k^{\text{th}}$  row, every term of which will have an  $x_k$ . Similarly, every term in the first column has an  $x_1$  and a  $b_1$ , every term in the second column has an  $x_2$  and a  $b_2$ , and so through the  $k^{\text{th}}$  column, every term of which has an  $x_k$  and a  $b_k$ . On the right side of the equations, each term has a  $y$  times the  $x$  that is appropriate for a particular row. So, for the general linear model given above, the normal equations are:

$$b_1 \text{ equation: } (\sum x_1^2)b_1 + (\sum x_1x_2)b_2 + (\sum x_1x_3)b_3 + \dots + (\sum x_1x_k)b_k = \sum x_1y$$

$$b_2 \text{ equation: } (\sum x_1x_2)b_1 + (\sum x_2^2)b_2 + (\sum x_2x_3)b_3 + \dots + (\sum x_2x_k)b_k = \sum x_2y$$

$$b_3 \text{ equation: } (\sum x_1x_3)b_1 + (\sum x_2x_3)b_2 + (\sum x_3^2)b_3 + \dots + (\sum x_3x_k)b_k = \sum x_3y$$

$$b_k \text{ equation: } (\sum x_1x_k)b_1 + (\sum x_2x_k)b_2 + (\sum x_3x_k)b_3 + \dots + (\sum x_k^2)b_k = \sum x_ky$$

Given the X coefficients, the constant term can be computed as

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - \dots - b_k\bar{X}_k$$

Note that the normal equations for the general linear model include the solution for the simple linear regression

$$(\sum x_1^2)b_1 = \sum x_1y$$

Hence,

$$b_1 = (\sum x_1y) / \sum x_1^2 \text{ as previously given.}$$

In fact, all of this section on multiple regression can be applied to the simple linear regression as a special case.

The corrected sums of squares and products are computed in the familiar (by now) manner:

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} = (65^2 + \dots + 61^2) - \frac{(2206)^2}{28} = 5,974.7143$$

$$\sum x_1^2 = \sum X_1^2 - \frac{(\sum X_1)^2}{n} = (41^2 + \dots + 46^2) - \frac{(1987)^2}{28} = 11,436.9643$$

$$\sum x_1y = \sum X_1Y - \frac{(\sum X_1)(\sum Y)}{n} = (41)(65) + \dots + (46)(61) - \frac{(1,987)(2,206)}{28} = 6,428.7858$$

Similarly,

$$\sum x_2x_2 = -1,171.4642$$

$$\sum x_2y = 2,632.2143$$

$$\sum x_1x_3 = 3,458.8215$$

$$\sum x_3^2 = 2,606.1072$$

$$\sum x_2^2 = 5,998.9643$$

$$\sum x_3y = 3,327.9286$$

$$\sum x_3x_3 = 1,789.6786$$

Putting these values in the normal equations gives:

$$\begin{aligned} 11,436.9643b_1 - 1,171.4642b_2 + 3,458.8215b_3 &= 6,428.7858 \\ -1,171.4642b_1 + 5,998.9643b_2 + 1,789.6786b_3 &= 2,632.2143 \\ 3,458.8215b_1 + 1,789.6786b_2 + 2,606.1072b_3 &= 3,327.9286 \end{aligned}$$



These equations can be solved by any of the standard procedures for simultaneous equations. One approach (applied to the above equations) is as follows:

1. Divide through each equation by the numerical coefficient of  $b_1$ .  
 $b_1 - 0.102,427,897b_2 + 0.302,424,788b_3 = 0.562,105,960$   
 $b_1 - 5.120,911,334b_2 - 1.527,727,949b_3 = -2.246,943,867$   
 $b_1 + 0.517,424,389b_2 + 0.753,466,809b_3 = 0.962,156,792$
2. Subtract the second equation from the first and the third from the first so as to leave two equations in  $b_2$  and  $b_3$ .  
 $5.018,483,437b_2 + 1.830,152,737b_3 = 2.809,049,827$   
 $-0.619,852,286b_2 - 0.451,042,021b_3 = -0.400,050,832$
3. Divide through each equation by the numerical coefficient of  $b_2$ .  
 $b_2 + 0.364,682,430b_3 = 0.559,740,779$   
 $b_2 + 0.727,660,494b_3 = 0.645,397,042$
4. Subtract the second of these equations from the first, leaving one equation in  $b_3$ .  
 $-0.362,978,064b_3 = -0.085,656,263$
5. Solve for  $b_3$ .  
 $b_3 = \frac{-0.085,656,263}{-0.362,978,064} = 0.235,981,927$
6. Substitute this value of  $b_3$  in one of the equations (say the first) of step 3 and solve for  $b_2$ .  
 $b_2 + (0.364,682,430)(0.235,981,927) = 0.559,740,779$   
 $b_2 = 0.473,682,316$
7. Substitute the solutions for  $b_2$  and  $b_3$  in one of the equations (say the first) of step 1, and solve for  $b_1$ .  
 $b_1 - (0.102,427,897)(0.473,682,316) + (0.302,424,788)(0.235,981,927) = 0.562,105,960$   
 $b_1 = 0.539,257,459$
8. As a check, add up the original normal equations and substitute the solutions for  $b_1, b_2,$  and  $b_3$ .  
 $13,724,3216b_1 + 6,617,1787b_2 + 7,854,6073b_3 = 12,388,9287$   
 $12,388,92869 = 12,388,9287, \text{ check}$

Given the values of  $b_1, b_2,$  and  $b_3$  we can now compute

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - b_3\bar{X}_3 = -11.7320$$

Thus, after rounding of the coefficients, the regression equation is

$$\hat{Y} = -11.732 + 0.539X_1 + 0.474X_2 + 0.236X_3$$

It should be noted that in solving the normal equations more digits have been carried than would be justified by the rules for number of significant digits. Unless this is done, the rounding errors may make it difficult to check the computations.

## Tests of Significance

To test the significance of the fitted regression, the outline for the analysis of variance is

Source	df
Reduction due to regression on $X_1, X_2,$ and $X_3$ .....	3
Residuals.....	24
Totals.....	27

The degrees of freedom for the total are equal to the number of observations minus 1. The total sum of squares is

$$\text{Total SS} = \sum y^2 = 5,974.7143$$

The degrees of freedom for the reduction are equal to the number of independent variables fitted, in this case 3. The reduction sum of squares for any least squares regression is

$$\text{Reduction SS} = \sum (\text{estimated coefficients})(\text{right side of their normal equations})$$

In this example there are three coefficients estimated by the normal equations, and so

$$\begin{aligned} \text{Reduction SS} &= b_1(\sum x_1y) + b_2(\sum x_2y) + b_3(\sum x_3y) \\ &= (0.53926)(6,428.7858) + (0.47368)(2,632.2143) \\ &\quad + (0.23598)(3,327.9286) \\ &= 5,498.9389 \end{aligned}$$

The residual df and sum of squares are obtained by subtraction. Thus the analysis becomes

Source	df	SS	MS
Reduction due to $X_1, X_2,$ and $X_3$ .....	3	5,498.9389	1,832.9796
Residuals.....	24	475.7754	19.8240
Total.....	27	5,974.7143	

To test the regression we compute

$$F_{3,24,df} = \frac{1,832.9796}{19.8240} = 92.46$$

which is significant at the 0.01 level.

Often we will want to test individual terms of the regression. In the previous example we might want to test the hypothesis that the true value of  $b_3$  is zero. This would be equivalent to testing whether the variable  $X_3$  makes any contribution to the prediction of  $Y$ . If we decide that  $b_3$  may be equal to zero, we might rewrite the equation in terms of  $X_1$  and  $X_2$ . Similarly, we could test the hypothesis that  $b_1$  and  $b_2$  are both equal to zero.

To test the contribution of any set of the independent variables in the presence of the remaining variables:

1. Fit all independent variables and compute the reduction and residual sums of squares.
2. Fit a new regression that includes only the variables not being tested. Compute the reduction due to this regression.
3. The reduction obtained in the first step minus the reduction in the second step is the gain due to the variables being tested.
4. The mean square for the gain (step 3) is tested against the residual mean square from the first step.

Two examples will illustrate the procedure:

I. Test  $X_1$  and  $X_2$  in the presence of  $X_3$ .

1. The reduction due to  $X_1$ ,  $X_2$ , and  $X_3$  is 5,498.9389 with 3 df. The residual is 475.7754 with 24 degrees of freedom (from previous example).

2. For fitting  $X_3$  alone, the normal equation is

$$(\sum x_3^2)b_3 = (\sum x_3y)$$

$$\text{or } 2,606.1072b_3 = 3,327.9286$$

$$b_3 = 1.27697$$

The reduction due to  $X_3$  alone is

$$\begin{aligned} \text{Red. SS} &= b_3(\sum x_3y) = 1.27697(3,327.9286) \\ &= 4,249.6650 \text{ with 1 df.} \end{aligned}$$

3. The gain due to  $X_1$  and  $X_2$  after  $X_3$  is

$$\begin{aligned} \text{Gain SS} &= \text{Reduction due to } X_1, X_2, X_3 - \text{reduction} \\ &\quad \text{due to } X_3 \text{ alone.} \\ &= 5,498.9389 - 4,249.6650 \\ &= 1,249.2739 \text{ with } (3 - 1) = 2 \text{ df.} \end{aligned}$$

4. Then

$$\begin{aligned} F_{2/24 \text{ df}} &= \frac{\text{Gain MS}}{\text{Residual MS}} = \frac{624.63695}{19.8240} \\ &= 31.51, \text{ significant at the 0.01 level.} \end{aligned}$$

This test is usually presented in the analysis of variance form:

Source	df	SS	MS
Reduction due to $X_1, X_2$ , and $X_3$ .....	3	5,498.9389	
Reduction due to $X_3$ alone.....	1	4,249.6650	
Gain due to $X_1$ and $X_2$ after $X_3$ .....	2	1,249.2739	624.63695
Residuals.....	24	475.7754	19.8240
Total.....	27	5,974.7143	

$$F_{2/24 \text{ df}} = \frac{624.63695}{19.8240} = 31.51, \text{ significant at the 0.01 level.}$$

II. Test  $X_2$  in the presence of  $X_1$  and  $X_3$

The normal equations for fitting  $X_1$  and  $X_3$  are

$$(\sum x_1^2)b_1 + (\sum x_1x_3)b_3 = \sum x_1y$$

$$(\sum x_1x_3)b_1 + (\sum x_3^2)b_3 = \sum x_3y$$

or

$$11,436.9643b_1 + 3,458.8215b_3 = 6,428.7858$$

$$3,458.8215b_1 + 2,606.1072b_3 = 3,327.9286$$

The solutions are

$$b_1 = 0.29387 \quad b_3 = 0.88695$$

The reduction sum of squares is

$$\begin{aligned} \text{Reduction SS} &= b_1(\sum x_1y) + b_3(\sum x_3y) \\ &= (0.29387)(6,428.7858) + (0.88695)(3,327.9286) \\ &= 4,840.9336, \text{ with 2 df.} \end{aligned}$$

The analysis is:

Source	df	SS	MS
Reduction due to $X_1, X_2$ , and $X_3$ .....	3	5,498.9389	
Reduction due to $X_1$ and $X_3$ .....	2	4,840.9336	
Gain due to $X_2$ after $X_1$ and $X_3$ .....	1	658.0053	658.0053
Residuals.....	24	475.7754	19.8240
Total.....	27	5,974.7143	

$$F_{1/24 \text{ df}} = \frac{658.0053}{19.8240} = 33.19, \text{ significant at the 0.01 level.}$$

### Coefficient of Multiple Determination

As a measure of how well the regression fits the data, it is customary to compute the ratio of the reduction sum of squares to the total sum of squares. This ratio is symbolized by  $R^2$  and is sometimes called the coefficient of determination:

$$R^2 = \frac{\text{Reduction SS}}{\text{Total SS}}$$

For the regression of  $Y$  on  $X_1, X_2$ , and  $X_3$ ,

$$R^2 = \frac{5,498.9389}{5,974.7143} = 0.92$$

The  $R^2$  value is usually referred to by saying that a certain percentage (92 in this case) of the variation in  $Y$  is associated with the regression.

The square root ( $R$ ) of the ratio is called the multiple correlation coefficient.

### The c-multipliers

Putting confidence limits on a multiple regression requires computation of the Gauss or c-multipliers. The c-multipliers are the elements of the

inverse of the matrix of corrected sums of squares and products as they appear in the normal equations. Thus, in fitting the regression of  $Y$  on  $X_1$ ,  $X_2$ , and  $X_3$  the matrix of corrected sums of squares and products is:

$$\begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 & \sum x_1 x_3 \\ \sum x_1 x_2 & \sum x_2^2 & \sum x_2 x_3 \\ \sum x_1 x_3 & \sum x_2 x_3 & \sum x_3^2 \end{bmatrix} = \begin{bmatrix} 11,436.9643 & -1,171.4642 & 3,458.8215 \\ -1,171.4642 & 5,998.9643 & 1,789.6786 \\ 3,458.8215 & 1,789.6786 & 2,606.1072 \end{bmatrix}$$

The inverse of this matrix is:

$$\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \begin{bmatrix} 0.000,237,573 & 0.000,176,649 & -0.000,436,615 \\ 0.000,176,649 & 0.000,340,994 & -0.000,468,616 \\ -0.000,436,615 & -0.000,468,616 & 0.001,285,000 \end{bmatrix}$$

The matrix of  $c$ -multipliers is symmetric, and therefore  $c_{12} = c_{21}$ ,  $c_{13} = c_{31}$ , etc.

The procedure for calculating the  $c$ -multipliers will not be given here. Those who are interested can refer to one of the standard statistical textbooks such as Goulden or Snedecor. However, because the  $c$ -multipliers are the output of many electronic computer programs, some of their applications will be described.

One of the important uses is in the calculation of confidence limits on the mean value of  $Y$  (i.e., regression  $Y$ ) associated with a specified set of  $X$  values. The general equation for  $k$  independent variables is:

$$\text{confidence limits} = \hat{Y} \pm t \sqrt{\text{Residual MS} \left( \frac{1}{n} + \sum_{i=1}^k c_{ij} (X_i - \bar{X}_i)(X_j - \bar{X}_j) \right)}$$

where:  $i$  and  $j$  each go from 1 to  $k$

$t$  has  $df$  equal to the degrees of freedom for the residual mean square.

In the example, if we specify  $X_1 = 80.9643$ ,  $X_2 = 66.5357$ , and  $X_3 = 76.8214$ , then  $\hat{Y} = 81.576$  and the 95-percent confidence limits are:

$$\begin{aligned} & 81.576 \pm 2.064 \sqrt{19.8240 \left[ \frac{1}{28} + c_{11}(80.9643 - \bar{X}_1)^2 + c_{22}(66.5357 - \bar{X}_2)^2 \right.} \\ & \quad \left. + c_{33}(76.8214 - \bar{X}_3)^2 + 2c_{12}(80.9643 - \bar{X}_1)(66.5357 - \bar{X}_2) \right.} \\ & \quad \left. + 2c_{13}(80.9643 - \bar{X}_1)(76.8214 - \bar{X}_3) \right.} \\ & \quad \left. + 2c_{23}(66.5357 - \bar{X}_2)(76.8214 - \bar{X}_3) \right] } \\ & = 81.576 \pm 2.064 \sqrt{19.8240 \left[ \frac{1}{28} + 0.019,948,390 \right]} \\ & = 79.41 \text{ to } 83.74 \end{aligned}$$

Note that each cross-product term such as  $c_{13}(X_1 - \bar{X}_1)(X_3 - \bar{X}_3)$  is multiplied by 2. This results because in summing over both  $i$  and  $j$  we get the terms  $c_{13}(X_1 - \bar{X}_1)(X_3 - \bar{X}_3)$  and  $c_{31}(X_3 - \bar{X}_3)(X_1 - \bar{X}_1)$ . As previously noted, the matrix of  $c$ -multipliers is symmetric, so that  $c_{13} = c_{31}$ ; hence we can combine these two terms to get  $2c_{13}(X_1 - \bar{X}_1)(X_3 - \bar{X}_3)$ .

For the confidence limits on a single predicted value of  $Y$  (as opposed to mean  $Y$ ) associated with a specified set of  $X$  values the equation is

$$\hat{Y} \pm t \sqrt{\text{Residual MS} \left[ 1 + \frac{1}{n} + \sum_{i=1}^k c_{ij} (X_i - \bar{X}_i)(X_j - \bar{X}_j) \right]}$$

With the above set of  $X$  values these limits would be

$$72.13 \text{ to } 91.02$$

The  $c$ -multipliers may also be used to calculate the estimated regression coefficients. The general equation is

$$b_j = \sum_{i=1}^k c_{ij} (\sum x_i y)$$

where:  $\sum x_i y$  = The right hand side of the  $i^{\text{th}}$  normal equation ( $i = 1, \dots, k$ ).

To illustrate,  $b_2$  in the previous example would be calculated as:

$$\begin{aligned} b_2 &= c_{21}(\sum x_1 y) + c_{22}(\sum x_2 y) + c_{23}(\sum x_3 y) \\ &= 0.000,176,649(6,428.7858) + 0.000,340,994(2,632.2143) \\ &\quad - 0.000,468,616(3,327.9286) \\ &= -0.47369 \end{aligned}$$

as before (except for rounding errors)

The regression coefficients are sample estimates and are, of course, subject to sampling variation. This means that any regression coefficient has a standard error and any pair of coefficients will have a covariance. The standard error of a regression coefficient is estimated by

$$\text{Standard error of } b_j = \sqrt{c_{jj}(\text{Residual MS})}$$

Hence,

$$\text{Variance of } b_j = c_{jj}(\text{Residual MS})$$

The covariance of any two coefficients is

$$\text{Covariance of } b_i \text{ and } b_j = c_{ij}(\text{Residual MS}).$$

The variance and covariance equations permit testing various hypotheses about the regression coefficients by means of a  $t$  test in the general form

$$t = \frac{\theta - \theta}{\sqrt{\text{Variance of } \theta}}$$

where:  $\theta$  = Any linear function of the estimated coefficients  
 $\theta$  = Hypothesized value of the function.

In the discussion of the analysis of variance we tested the contribution of  $X_2$  in the presence of  $X_1$  and  $X_3$  (obtaining  $F = 33.19^{**}$ ). This is actually equivalent to testing the hypothesis that in the model  $Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3$  the true value of  $b_2$  is zero. The  $t$  test of this hypothesis is as follows:

$$\begin{aligned} t &= \frac{b_2 - 0}{\sqrt{\text{Variance of } b_2}} = \frac{b_2}{\sqrt{c_{22}(\text{Residual MS})}} \\ &= \frac{-0.47368}{\sqrt{0.000,340,994(19.8240)}} \\ &= -5.761, \text{ with } 24 \text{ df} = \text{df for Residual MS} \end{aligned}$$

Note that  $t^2 = 33.19$  = the  $F$  value obtained in testing this same hypothesis.

It is also possible to test the hypothesis that a coefficient has some value other than zero or that a linear function of the coefficients has some specified value. For example, if there were some reason for believing that  $b_1 = 2b_3$  or  $b_1 - 2b_3 = 0$ , this hypothesis could be tested by

$$t = \frac{b_1 - 2b_3}{\sqrt{\text{Variance of } (b_1 - 2b_3)}}$$

Referring to the section on the variance of a linear function, we find that:

$$\begin{aligned} \text{Variance of } (b_1 - 2b_3) &= \text{Variance of } b_1 + 4(\text{variance of } b_3) \\ &\quad - 4(\text{covariance of } b_1 \text{ and } b_3) \\ &= (c_{11} + 4c_{33} - 4c_{13})(\text{Residual MS}) \\ &= 19.8240[(0.000,237,573) \\ &\quad + 4(0.001,285,000) - 4(-0.000,436,615)] \\ &= 0.141,226,830 \end{aligned}$$

Then,

$$t_{24 \text{ df}} = \frac{0.53926 - 2(0.23598)}{\sqrt{0.141,226,830}} = \frac{0.06730}{0.37580} = 0.179$$

The hypothesis would not be rejected at the 0.05 level.

*Assumptions.*—The assumptions underlying these methods of fitting a multiple regression are the same as those for a simple linear regression: equal variance of  $Y$  at all combinations of  $X$  values and independence of the errors (i.e., departures from regression) of the sample observations. Application of the  $F$  or  $t$  distributions (for testing or setting confidence limits) requires the further assumption of normality.

The reader is again warned against inferring more than is actually implied by a regression equation and analysis. For one thing, no matter how well a particular equation may fit a set of data, it is only a mathematical approximation of the relationship between a dependent and a set of independent variables. It should not be construed as representing a biological or physical law. Nor does it prove the existence of a cause and effect relationship. It is merely a convenient way of describing an observed association.

Tests of significance must also be interpreted with caution. A significant  $F$  or  $t$  test means that the estimated values of the regression coefficients differ from the hypothesized values (usually zero) by more than would be expected by chance. Even though a regression is highly significant, the predicted values may not be very close to the actual (look at the standard errors). Conversely, the fact that a particular variable (say  $X_i$ ) is not significantly related to  $Y$  does not necessarily mean that a relationship is lacking. Perhaps the test was insensitive or we did not select the proper model to represent the relationship.

Regression analysis is a very useful technique, but it does not relieve the research worker of the responsibility for thinking.

### Curvilinear Regressions and Interactions

*Curves.*—Many forms of curvilinear relationships can be fitted by the regression methods that have been described in the previous sections.

If the relationship between height and age is assumed to be hyperbolic so that

$$\text{Height} = a + \frac{b}{\text{Age}}$$

then we could let  $Y = \text{Height}$  and  $X_1 = 1/\text{Age}$  and fit

$$Y = a + b_1 X_1$$

Similarly, if the relationship between  $Y$  and  $X$  is quadratic

$$Y = a + bX + cX^2$$

we can let  $X = X_1$  and  $X^2 = X_2$  and fit

$$Y = a + b_1 X_1 + b_2 X_2$$

Functions such as

$$Y = aX^b$$

$$Y = a(b^X)$$

$$10^Y = aX^b$$

which are nonlinear in the coefficients can sometimes be made linear by a logarithmic transformation. The equation

$$Y = aX^b$$

would become

$$\log Y = \log a + b(\log X)$$

which could be fitted by

$$Y' = a' + b_1 X_1$$

where  $Y' = \log Y$ , and

$$X_1 = \log X.$$

The second equation transforms to

$$\log Y = \log a + (\log b)X$$

The third becomes

$$Y = \log a + b(\log X)$$

Both can be fitted by the linear model.

In making these transformations the effect on the assumption of homogeneous variance must be considered. If  $Y$  has homogeneous variance,  $\log Y$  probably will not have—and vice versa.

Some curvilinear models cannot be fitted by the methods that have been described. Some examples are

$$Y = a + b^x$$

$$Y = a(X - b)^2$$

$$Y = a(X_1 - b)(X_2 - c)$$

Fitting these models requires more cumbersome procedures.

*Interactions.*—Suppose that there is a simple linear relationship between  $Y$  and  $X_1$ . If the slope ( $b$ ) of this relationship varies, depending on the level of some other independent variable ( $X_2$ ), then  $X_1$  and  $X_2$  are said to interact. Such interactions can sometimes be handled by introducing interaction variables.

To illustrate, suppose that we know that there is a linear relationship between  $Y$  and  $X_1$ .

$$Y = a + bX_1$$

Suppose further that we know or suspect that the slope ( $b$ ) varies linearly with  $Z$

$$b = a' + b'Z$$

This implies the relationship

$$Y = a + (a' + b'Z)X_1$$

or

$$Y = a + a'X_1 + b'X_1Z$$

which can be fitted by

$$Y = a + b_1X_1 + b_2X_2$$

where  $X_2 = X_1Z$ , an interaction variable.

If the  $Y$ -intercept is also a linear function of  $Z$ , then

$$a = a'' + b''Z$$

and the form of relationship is

$$Y = a'' + b''Z + a'X_1 + b'X_1Z$$

which could be fitted by

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3$$

where  $X_3 = Z$ , and

$$X_3 = X_1Z.$$

### Group Regressions

Linear regressions of  $Y$  on  $X$  were fitted for each of two groups. The basic data and fitted regressions were:

Group A		Sum	Mean
Y	3 7 9 6 8 13 10 12 14	82	9.111
X	1 4 7 7 2 9 10 6 12	58	6.444

$$n = 9, \Sigma Y^2 = 848, \Sigma XY = 609, \Sigma X^2 = 480$$

$$\Sigma y^2 = 100.8889, \Sigma xy = 80.5556, \Sigma x^2 = 106.2222,$$

$$\hat{Y} = 4.224 + 0.7584X$$

Residual SS = 39.7980, with 7 df.

Group B		Sum	Mean
Y	4 6 12 2 8 7 0 5 9 2 11 3 10	79	6.077
X	4 9 14 6 9 12 2 7 5 5 11 2 13	99	7.615

$$n = 13, \Sigma Y^2 = 653, \Sigma XY = 753, \Sigma X^2 = 951$$

$$\Sigma y^2 = 172.9231, \Sigma xy = 151.3846, \Sigma x^2 = 197.0769$$

$$\hat{Y} = 0.228 + 0.7681X$$

Residual SS = 56.6370, with 11 df.

Now we might ask, are these really different regressions? Or could the data be combined to produce a single regression that would be applicable to both groups? If there is no significant difference between the residual mean squares for the two groups (this matter may be determined by Bartlett's test, page 22), the test described below helps to answer the question.

*Testing for the common regressions.*—Simple linear regressions may differ either in their slope or in their level. In testing for common regressions the procedure is to test first for common slopes. If the slopes differ significantly, the regressions are different and no further testing is needed. If the slopes are not significantly different, the difference in level is tested. The analysis table is:

Line	Group	df	$\Sigma y^2$	$\Sigma xy$	$\Sigma x^2$	Residuals	
						df	MS
1	A	8	100.8889	80.5556	106.2222	7	39.7980
2	B	12	172.9231	151.3846	197.0769	11	56.6370
3						18	96.4350
4						1	0.0067
5	Common slope	20	273.8120	231.9402	303.2991	19	96.4417
6						1	80.1954
7	Single regression	21	322.7727	213.0455	310.5909	20	176.6371

The first two lines in this table contain the basic data for the two groups. To the left are the total df for the groups (8 for A and 12 for B). In the center are the corrected sums of squares and products. The right side of the table gives the residual sum of squares and df. Since only simple linear regressions have been fitted, the residual df for each group are one less than the total df. The residual sum of squares is obtained by first computing the reduction sum of squares for each group.

$$\text{Reduction SS} = \frac{(\Sigma xy)^2}{\Sigma x^2}$$

This reduction is then subtracted from the total sum of squares ( $\Sigma y^2$ ) to give the residuals.

Line 3 is obtained by pooling the residual df and residual sums of squares for the groups. Dividing the pooled sum of squares by the pooled df gives the pooled mean square.

The left side and center of line 5 (we will skip line 4 for the moment) is obtained by pooling the total df and the corrected sums of squares and products for the groups. These are the values that are obtained under the assumption of no difference in the slopes of the group regressions. If the assumption is wrong, the residuals about this common slope regression will be considerably larger than the mean square residual about the separate regressions. The residual df and sum of squares are obtained by fitting a straight line to this pooled data. The residual df are, of course, one less than the total df. The residual sum of squares is, as usual,

$$\text{Residual SS} = 273.8120 - \frac{(231.9402)^2}{303.2991} = 96.4417$$

Now, the difference between these residuals (line 4 = line 5 - line 3) provides a test of the hypothesis of common slopes. The error term for this test is the pooled mean square from line 3.

$$\text{Test of common slopes: } F_{1/18} \text{ at } = \frac{0.0067}{5.3575}$$

The difference is not significant.

If the slopes differed significantly, the groups would have different regressions, and we would stop here. Since the slopes did not differ, we now go on to test for a difference in the levels of the regression.

Line 7 is what we would have if we ignored the groups entirely, lumped all the original observations together, and fitted a single linear regression. The combined data are as follows:

$$n = (9 + 13) = 22 \text{ (so the df for total} = 21)$$

$$\Sigma Y = (82 + 79) = 161, \Sigma Y^2 = (848 + 653) = 1,501$$

$$\Sigma y^2 = 1,501 - \frac{(161)^2}{22} = 322.7727$$

$$\Sigma X = (58 + 99) = 157, \Sigma X^2 = (480 + 951) = 1,431$$

$$\Sigma x^2 = 1,431 - \frac{(157)^2}{22} = 310.5909$$

$$\Sigma XY = (609 + 753) = 1,362, \Sigma xy = 1,362 - \frac{(157)(161)}{22} = 213.0455$$

From this we obtain the residual values on the right side of line 7.

$$\text{Residual SS} = 322.7727 - \frac{(213.0455)^2}{310.5909} = 176.6371$$

If there is a real difference among the levels of the groups, the residuals about this single regression will be considerably larger than the mean square residual about the regression that assumed the same slopes but different levels. This difference (line 6 = line 7 - line 5) is tested against the residual mean square from line 5.

$$\text{Test of levels: } F_{1/19} \text{ at } = \frac{80.1954}{5.0759} = 15.80^{**}$$

As the levels differ significantly, the groups do not have the same regressions.

The test is easily extended to cover several groups, though there may be a problem in finding which groups are likely to have separate regressions and which can be combined. The test can also be extended to multiple regressions.

### Analysis of Covariance in a Randomized Block Design

A test was made of the effect of three soil treatments on the height growth of 2-year-old seedlings. Treatments were assigned at random to the three plots within each of 11 blocks. Each plot was made up of 50 seedlings. Average 5-year height growth was the criterion for evaluating treatments. Initial heights and 5-year growths, all in feet, were:

Block	Treatment A		Treatment B		Treatment C		Block totals	
	Height	Growth	Height	Growth	Height	Growth	Height	Growth
1	3.6	8.9	3.1	10.7	4.7	12.4	11.4	32.0
2	4.7	10.1	4.9	14.2	2.6	9.0	12.2	33.3
3	2.6	6.3	8	5.9	1.5	7.4	4.9	19.6
4	5.3	14.0	4.6	12.6	4.3	10.1	14.2	36.7
5	3.1	9.6	3.9	12.5	3.3	6.8	10.3	28.9
6	1.8	6.4	1.7	9.6	3.6	10.0	7.1	26.0
7	5.8	12.3	5.5	12.8	5.8	11.9	17.1	37.0
8	3.8	10.8	2.6	8.0	2.0	7.5	8.4	26.3
9	2.4	8.0	1.1	7.5	1.6	5.2	5.1	20.7
10	5.3	12.6	4.4	11.4	5.8	13.4	15.5	37.4
11	3.6	7.4	1.4	8.4	4.8	10.7	9.8	26.5
Sums	42.0	106.4	34.0	113.6	40.0	104.4	116.0	324.4
Means	3.82	9.67	3.09	10.33	3.64	9.49	3.52	9.83

The analysis of variance of growth is:

Source	df	SS	MS
Blocks	10	132.83	
Treatment	2	4.26	2.130
Error	20	68.88	3.444
Total	32	205.97	

$$F \text{ (for testing treatments)}_{2/20} \text{ at } = \frac{2.130}{3.444}$$

Not significant at 0.05 level.

There is no evidence of a real difference in growth due to treatments. There is, however, reason to believe that, for young seedlings, growth is affected by initial height. A glance at the block totals seems to suggest that plots with greatest initial height had greatest 5-year growth. The possibility that effects of treatment are being obscured by differences in initial heights raises the question of how the treatments would compare if adjusted for differences in initial heights.

If the relationship between height growth and initial height is linear and if the slope of the regression is the same for all treatments, the test of adjusted treatment means can be made by an analysis of covariance as described below. In this analysis, the growth will be labeled  $Y$  and initial height  $X$ .

Computationally the first step is to obtain total, block, treatment, and error sums of squares of  $X$  ( $SS_x$ ) and sums of products of  $X$  and  $Y$  ( $SP_{xy}$ ), just as has already been done for  $Y$ .

$$\text{For } X: \quad C.T._x = \frac{116.0^2}{33} = 407.76$$

$$\text{Total } SS_x = (3.6^2 + \dots + 4.8^2) - C.T._x = 73.26$$

$$\text{Block } SS_x = \left( \frac{11.4^2 + \dots + 9.8^2}{3} \right) - C.T._x = 54.31$$

$$\text{Treatment } SS_x = \left( \frac{42.0^2 + 34.0^2 + 40.0^2}{11} \right) - C.T._x = 3.15$$

$$\text{Error } SS_x = \text{Total} - \text{Block} - \text{Treatment} = 15.80$$

$$\text{For } XY: \quad C.T._{xy} = \frac{(116.0)(324.4)}{33} = 1,140.32$$

$$\text{Total } SP_{xy} = (3.6)(8.9) + \dots + (4.8)(10.7) - C.T._{xy} = 103.99$$

$$\text{Block } SP_{xy} = \left( \frac{(11.4)(32.0) + \dots + (9.8)(26.5)}{3} \right) - C.T._{xy} = 82.71$$

$$\text{Treatment } SP_{xy} = \left( \frac{(42.0)(106.4) + (34.0)(113.6) + (40.0)(104.4)}{11} \right) - C.T._{xy} = -3.30$$

$$\text{Error } SP_{xy} = \text{Total} - \text{Block} - \text{Treatment} = 24.58$$

These computed terms are arranged in a manner similar to that for the test of group regressions (which is exactly what the covariance analysis is). One departure is that the total line is put at the top.

Source	Residuals						
	df	SS <sub>x</sub>	SP <sub>xy</sub>	SS <sub>y</sub>	df	SS	MS
Total.....	32	205.97	103.99	73.26	19	30.641	1.613
Blocks.....	10	132.83	82.71	54.31			
Treatment.....	2	4.26	-3.30	3.15			
Error.....	20	68.88	24.58	15.80			

On the error line, the residual sum of squares after adjusting for a linear regression is

$$\text{Residual } SS = SS_y - \frac{(SP_{xy})^2}{SS_x} = 68.88 - \frac{24.58^2}{15.80} = 30.641$$

This sum of squares has 1 df less than the unadjusted sum of squares. To test treatments we first pool the unadjusted df and sums of squares and products for treatment and error. The residual terms for this pooled line are then computed just as they were for the error line

Treatment + error..	Residuals					
	df	SS <sub>x</sub>	SP <sub>xy</sub>	SS <sub>y</sub>	df	SS
	22	73.14	21.28	18.95	21	49.244

Then to test for a difference among treatments after adjustment for the regression of growth on initial height, we compute the difference in residuals between the error and the treatment + error lines

$$\text{Difference for testing adjusted treatments} \quad \begin{array}{|l} \text{df} \quad \text{SS} \quad \text{MS} \\ \hline 2 \quad 18.603 \quad 9.302 \end{array}$$

The mean square for the difference in residuals is now tested against the residual mean square for error.

$$F_{2/19 \text{ df}} = \frac{9.302}{1.613} = 5.77^*$$

Thus, after adjustment, the difference in treatment means is found to be significant at the 0.05 level. It may also happen that differences that were significant before adjustment are not significant afterwards.

If the independent variable has been affected by treatment, interpretation of a covariance analysis requires careful thinking. The covariance adjustment may have the effect of removing the treatment differences that are being tested. On the other hand, it may be informative to know that treatments are or are not significantly different in spite of the covariance adjustment. The beginner who is uncertain of the interpretations would do well to select as covariates only those that have not been affected by treatments.

The covariance test may be made in a similar manner for any experimental design and, if desired (and justified), adjustment may be made for multiple or curvilinear regressions.

The entire analysis is usually presented in the following form:

Source	df	SS <sub>x</sub>	SP <sub>xy</sub>	SS <sub>y</sub>	Adjusted		
					df	SS	MS
Total.....	32	205.97	103.99	73.26			
Blocks.....	10	132.83	82.71	54.31			
Treatment.....	2	4.26	-3.30	3.15			
Error.....	20	68.88	24.58	15.80	19	30.641	
Treatment + Error.....	22	73.14	21.28	18.95	21	49.244	
Difference for testing adjusted treatment means					2	18.603	9.302

Unadjusted treatments:  $F_{2/20 \text{ df}} = \frac{2.130}{3.444}$ . Not significant.

Adjusted treatments:  $F_{2/19 \text{ df}} = \frac{9.302}{1.613} = 5.77$ . Significant at the 0.05 level.

*Adjusted means.*—If we wish to know what the treatment means are after adjustment for regression, the equation is

$$\text{Adjusted } \bar{Y}_i = \bar{Y}_i - b(\bar{X}_i - \bar{X})$$

where:  $\bar{Y}_i$  = Unadjusted mean for treatment  $i$ .

$$b = \text{Coefficient of the linear regression} = \frac{\text{Error } SP_{xy}}{\text{Error } SS_x}$$

$\bar{X}_i$  = Mean of the independent variable for treatment  $i$ .

$\bar{X}$  = Mean  $X$  for all treatments.

In the example we had  $\bar{X}_A = 3.82$ ,  $\bar{X}_B = 3.09$ ,  $\bar{X}_C = 3.64$ ,  $\bar{X} = 3.52$ , and

$$b = \frac{24.58}{15.80} = 1.56$$

So, the unadjusted and adjusted mean growths are

Treatment	Mean growths	
	Unadjusted	Adjusted
A	9.67	9.20
B	10.33	11.00
C	9.49	9.30

*Tests among adjusted means.*—In an earlier section we encountered methods of making further tests among the means. Ignoring the covariance adjustment, we could for example make an  $F$  test for pre-specified comparisons such as  $A+C$  vs.  $B$ , or  $A$  vs.  $C$ . Similar tests can also be made after adjustment for covariance, though they involve more labor. The  $F$  test will be illustrated for the comparison  $B$  vs.  $A+C$  after adjustment.

As might be suspected, to make the  $F$  test we must first compute sums of squares and products of  $X$  and  $Y$  for the specified comparison:

$$SS_y = \frac{[2(\sum Y_B) - (\sum Y_A + \sum Y_C)]^2}{[2^2 + 1^2 + 1^2][11]} = \frac{[2(113.6) - (106.4 + 104.4)]^2}{66} = 4.08$$

$$SS_x = \frac{[2(\sum X_B) - (\sum X_A + \sum X_C)]^2}{[2^2 + 1^2 + 1^2][11]} = \frac{[2(34.0) - (42.0 + 40.0)]^2}{66} = 2.97$$

$$SP_{xy} = \frac{[2(\sum Y_B) - (\sum Y_A + \sum Y_C)][2(\sum X_B) - (\sum X_A + \sum X_C)]}{[2^2 + 1^2 + 1^2][11]} = -3.48$$

From this point on, the  $F$  test of  $A+B$  vs.  $C$  is made in exactly the same manner as the test of treatments in the covariance analysis.

Source	df	$SS_y$	$SP_{xy}$	$SS_x$	Residuals SS	MS
2B-(A+C).....	1	4.08	-3.48	2.97	—	—
Error.....	20	68.88	24.58	15.80	30.641	1.613
Sum.....	21	72.96	21.10	18.77	49.241	—
Difference for testing adjusted comparison					1	18.600

$$F_{1/19, df} = \frac{18.600}{1.613} = 11.531. \quad \text{Significant at the 0.01 level.}$$

## REFERENCES FOR FURTHER READING

- Arkin, H., and Colton, R. R.  
1963. Tables for statisticians. Ed. 2, 168 pp., illus. New York: Barnes and Noble.
- Cochran, W. G.  
1963. Sampling techniques. Ed. 2, 413 pp., illus. New York: Wiley.
- \_\_\_\_\_ and Cox, G. M.  
1957. Experimental designs. Ed. 2, 611 pp., illus. New York: Wiley.
- Fisher, R. A.  
1954. Statistical methods for research workers. Ed. 12, 356 pp., illus. New York: Hafner.
- Freese, F.  
1962. Elementary forest sampling. U.S. Dept. Agr., Agr. Handb. 232, 91 pp.
- \_\_\_\_\_ 1964. Linear regression methods for forest research. U.S. Forest Serv. Res. Paper FPL-17, 136 pp., illus. Forest Products Laboratory, Madison, Wis.
- Goulden, C. H.  
1952. Methods of statistical analysis. Ed. 2, 467 pp., illus. New York: Wiley.
- Natrella, M. G.  
1963. Experimental statistics. U.S. Natl. Bureau Standards Handb. 91, 522 pp., illus.
- Schumacher, F. X., and Chapman, R. A.  
1954. Sampling methods in forestry and range management. Duke Univ. School Forestry Bul. 7. Ed. 3, 222 pp., illus. Durham, N. C.
- Snedecor, G. W.  
1956. Statistical methods. Ed. 5, 534 pp., illus. Ames, Iowa: Iowa State Univ. Press.
- Steel, R. G. D., and Torrie, J. H.  
1960. Principles and procedures of statistics. 481 pp., illus. New York: McGraw-Hill.
- Walker, H. M.  
1951. Mathematics essential for elementary statistics. Rev. ed., 382 pp., illus. New York: Holt.
- Wallis, W. A., and Roberts, H. V.  
1956. Statistics: a new approach. 646 pp., illus. Glencoe, Ill.: Free Press.
- Wilcoxon, F.  
1949. Some rapid approximate statistical procedures. Rev. ed., 16 pp., illus. New York: American Cyanamid Co.



**APPENDIX TABLES**

**Table 1.—Ratio of standard deviation to range for simple random samples of size n from normal populations**

n	$\sigma$		n	$\sigma$	
	Range	Range		Range	Range
2	0.886		12	0.307	
3	.591		14	.294	
4	.486		16	.283	
5	.430		18	.275	
6	.395		20	.268	
7	.370		30	.245	
8	.351		40	.231	
9	.337		50	.222	
10	.325				

Abridged by permission of the author and publishers from table 2.2.2 of Snedecor's *Statistical Methods* (ed. 5), Iowa State University Press.

**Table 2.—Distribution of t**

df	Probability									
	.5	.4	.3	.2	.1	.05	.02	.01	.001	
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619	
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598	
3	.765	.978	1.250	1.638	2.353	3.182	4.341	5.841	12.941	
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610	
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859	
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959	
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405	
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041	
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781	
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587	
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437	
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318	
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221	
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140	
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073	
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015	
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965	
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922	
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883	
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850	
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819	
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792	
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767	
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745	
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725	
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707	
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690	
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674	
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659	
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646	
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551	
60	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460	
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373	
$\infty$	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291	

Abridged from table III of Fisher and Yates' *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, Ltd., Edinburgh. Permission has been given by Dr. F. Yates, by the literary executor of the late Professor Sir Ronald A. Fisher, and by the publishers.

1 First line of figures in each pair is for the 5% level; second line in each pair is for the 1% level.

30	4.17	5.32	4.51	4.02	2.69	2.92	2.53	2.42	3.70	3.47	3.30	3.17	3.06	3.01	2.27	2.24	2.21	2.16	2.90	2.12	2.09	2.08	2.04	1.99	1.98	1.93	1.89	1.84	2.29	2.24	2.16	1.72	1.72	1.69	1.66	1.64	1.62
29	4.18	5.33	4.54	4.04	2.70	2.93	2.54	2.43	3.73	3.50	3.33	3.20	3.08	3.03	2.22	2.22	2.18	2.12	2.92	2.10	2.07	2.05	2.02	1.97	1.94	1.89	1.85	2.29	2.27	2.18	1.73	1.71	1.68	1.65	1.64	1.64	2.03
28	4.20	5.34	4.55	4.07	2.71	2.94	2.55	2.44	3.76	3.53	3.36	3.23	3.11	3.06	2.23	2.23	2.19	2.12	2.93	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.30	2.22	2.12	1.75	1.72	1.69	1.67	1.65	1.65	2.06
27	4.21	5.35	4.57	4.11	2.73	2.96	2.57	2.46	3.79	3.56	3.39	3.26	3.14	3.09	2.24	2.24	2.16	2.12	2.94	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.31	2.22	2.12	1.76	1.74	1.71	1.68	1.67	1.67	2.10
26	4.22	5.37	4.64	4.14	2.74	2.98	2.59	2.47	3.82	3.59	3.42	3.29	3.17	3.09	2.25	2.25	2.18	2.15	2.96	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.32	2.23	2.13	1.78	1.76	1.73	1.70	1.69	1.69	2.13
25	4.24	5.38	4.68	4.18	2.76	2.99	2.60	2.49	3.86	3.63	3.46	3.32	3.21	3.13	2.26	2.26	2.19	2.16	2.98	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.33	2.24	2.14	1.79	1.77	1.74	1.72	1.71	1.71	2.17
24	4.26	5.40	4.71	4.21	2.78	3.01	2.62	2.51	3.90	3.67	3.50	3.36	3.25	3.17	2.27	2.27	2.20	2.17	3.00	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.34	2.25	2.15	1.80	1.78	1.75	1.72	1.72	1.72	2.21
23	4.28	5.42	4.74	4.24	2.80	3.03	2.64	2.53	3.94	3.71	3.54	3.41	3.30	3.21	2.28	2.28	2.21	2.18	3.02	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.35	2.26	2.16	1.81	1.79	1.76	1.73	1.73	1.73	2.26
22	4.30	5.44	4.77	4.27	2.82	3.05	2.66	2.55	3.99	3.76	3.59	3.45	3.35	3.26	2.29	2.29	2.22	2.19	3.04	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.36	2.27	2.17	1.82	1.80	1.77	1.74	1.74	1.74	2.28
21	4.32	5.47	4.80	4.30	2.84	3.07	2.68	2.57	4.04	3.81	3.65	3.51	3.40	3.31	2.30	2.30	2.23	2.20	3.06	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.37	2.28	2.18	1.83	1.81	1.78	1.75	1.75	1.75	2.31
20	4.35	5.49	4.83	4.33	2.87	3.10	2.69	2.58	4.10	3.87	3.71	3.56	3.45	3.37	2.31	2.31	2.24	2.21	3.08	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.38	2.29	2.19	1.84	1.82	1.79	1.76	1.76	1.76	2.34
19	4.38	5.52	4.86	4.36	2.90	3.13	2.68	2.55	4.17	3.94	3.77	3.63	3.52	3.43	2.32	2.32	2.25	2.22	3.10	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.39	2.30	2.20	1.85	1.83	1.80	1.77	1.77	1.77	2.36
18	4.41	5.55	4.89	4.39	2.91	3.16	2.69	2.56	4.25	4.01	3.85	3.71	3.60	3.51	2.33	2.33	2.26	2.23	3.12	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.40	2.31	2.21	1.86	1.84	1.81	1.78	1.78	1.78	2.39
17	4.45	5.58	4.92	4.42	2.92	3.20	2.70	2.57	4.34	4.10	3.93	3.79	3.68	3.59	2.34	2.34	2.27	2.24	3.14	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.41	2.32	2.22	1.87	1.85	1.82	1.79	1.79	1.79	2.41
16	4.49	5.62	4.95	4.45	2.93	3.24	2.71	2.58	4.44	4.20	4.03	3.89	3.78	3.69	2.35	2.35	2.28	2.25	3.16	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.42	2.33	2.23	1.88	1.86	1.83	1.80	1.80	1.80	2.42
15	4.54	5.66	4.98	4.48	2.94	3.28	2.72	2.59	4.54	4.30	4.13	3.99	3.88	3.80	2.36	2.36	2.29	2.26	3.18	2.10	2.07	2.04	1.99	1.96	1.91	1.87	1.81	2.43	2.34	2.24	1.89	1.87	1.84	1.81	1.81	1.81	2.43



Table 3.—Distribution of F

Degrees of freedom for greater mean square

14	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00
13	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.52	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
12	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.56	2.52	2.48	2.44	2.40	2.37	2.34	2.32	2.29	2.27	2.24	2.21
11	4.84	3.98	3.59	3.36	3.20	3.07	2.97	2.90	2.85	2.80	2.76	2.72	2.68	2.64	2.60	2.56	2.52	2.49	2.46	2.44	2.41	2.39	2.36	2.34
10	4.96	4.10	3.71	3.48	3.32	3.19	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.54	2.52	2.54
9	5.12	4.26	3.86	3.63	3.47	3.34	3.23	3.18	3.13	3.10	3.07	3.03	2.99	2.95	2.92	2.88	2.85	2.82	2.80	2.77	2.76	2.73	2.72	2.71
8	5.28	4.46	4.07	3.84	3.69	3.56	3.44	3.39	3.34	3.31	3.28	3.24	3.21	3.17	3.14	3.11	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.54	3.51	3.48	3.45	3.41	3.38	3.34	3.32	3.29	3.27	3.24	3.23
6	5.99	5.14	4.75	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.66	4.62	4.58	4.54	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36	4.36
4	7.71	6.94	6.56	6.36	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.68	5.66	5.64	5.64	5.63	5.63
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.92	26.83	26.76	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.12
2	18.51	19.00	19.17	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.48	19.49	19.49	19.50	19.50	19.50
1	4.052	4.999	5.408	5.764	6.022	6.237	6.414	6.559	6.678	6.779	6.859	6.921	6.969	7.007	7.037	7.060	7.078	7.093	7.107	7.119	7.129	7.137	7.144	7.149
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	
	Error df																							

Table 3.—Distribution of F (continued)<sup>1</sup>

Error df	Degrees of freedom for greater mean square																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.50	1.43	1.41
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.86	1.81	1.74	1.70	1.65	1.59	1.51	1.44	1.37	1.35
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.02	1.97	1.92	1.83	1.78	1.71	1.67	1.62	1.56	1.48	1.41	1.34	1.32
38	4.10	3.25	2.85	2.62	2.46	2.35	2.27	2.20	2.14	2.09	2.05	2.01	1.96	1.91	1.82	1.77	1.70	1.66	1.61	1.55	1.47	1.40	1.33	1.31
40	4.08	3.23	2.83	2.61	2.45	2.34	2.26	2.19	2.12	2.07	2.04	2.00	1.95	1.90	1.81	1.76	1.69	1.65	1.60	1.54	1.46	1.39	1.32	1.30
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.97	1.92	1.87	1.78	1.73	1.66	1.62	1.57	1.50	1.43	1.36	1.29	1.27
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.93	1.88	1.79	1.74	1.67	1.63	1.58	1.51	1.44	1.37	1.30	1.28
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.78	1.73	1.66	1.62	1.57	1.50	1.43	1.36	1.29	1.27
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	2.00	1.96	1.91	1.86	1.77	1.72	1.65	1.61	1.56	1.49	1.42	1.35	1.28	1.26
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.89	1.84	1.75	1.70	1.63	1.59	1.54	1.47	1.40	1.33	1.26	1.24

Reproduced by permission of the author and publishers from table 10.5.3 of Snedecor's *Statistical Methods* (ed. 5), Iowa State University Press. Permission has also been granted by the literary executor of the late Professor Sir Ronald A. Fisher and Oliver and Boyd, Ltd., publishers, for the portion of the table computed from Dr. Fisher's table VI in *Statistical Methods for Research Workers*.

<sup>1</sup> First line of figures in each pair is for the 5% level; second line in each pair is for the 1% level.

Table 5.—Confidence intervals for binomial distribution

Number observed	95-percent interval						Fraction observed $\bar{p}$	Size of sample			
	Size of sample, $n$							100	250	1000	
	10	15	20	30	50						
0	0	0	0	0	0	0	0.0	4	0	1	0
1	0	45	0	22	0	17	0.07	4	0	1	0
2	3	56	2	40	1	31	0.11	5	0	4	0
3	7	65	4	48	3	38	0.14	6	1	5	3
4	12	74	8	55	6	44	0.17	7	1	6	2
5	19	81	12	62	9	49	0.21	8	2	7	4
6	26	88	16	68	12	54	0.25	9	3	7	5
7	35	93	21	73	15	59	0.29	10	3	10	5
8	44	97	27	79	19	64	0.34	11	4	11	6
9	55	100	32	84	23	68	0.39	12	5	12	6
10	69	100	38	88	27	73	0.44	13	6	13	7
11	86	100	45	92	32	77	0.49	14	7	14	8
12	105	100	52	96	36	81	0.54	15	8	15	9
13	126	100	60	98	41	85	0.59	16	9	16	10
14	148	100	68	100	46	88	0.64	17	10	17	11
15	172	100	78	100	51	91	0.69	18	11	18	12
16	200	100	88	100	56	94	0.74	19	12	19	13
17	230	100	100	100	62	97	0.79	20	13	20	14
18	265	100	110	100	69	99	0.84	21	14	21	15
19	305	100	120	100	75	100	0.89	22	15	22	16
20	350	100	130	100	83	100	0.94	23	16	23	17
21	400	100	140	100	90	100	0.98	24	17	24	18
22	455	100	150	100	97	100	1.00	25	18	25	19
23	515	100	160	100	104	100	1.00	26	19	26	20
24	580	100	170	100	111	100	1.00	27	20	27	21
25	650	100	180	100	119	100	1.00	28	21	28	22
26	725	100	190	100	127	100	1.00	29	22	29	23
27	805	100	200	100	136	100	1.00	30	23	30	24
28	890	100	210	100	145	100	1.00	31	24	31	25
29	980	100	220	100	155	100	1.00	32	25	32	26
30	1075	100	230	100	165	100	1.00	33	26	33	27
31	1175	100	240	100	175	100	1.00	34	27	34	28
32	1280	100	250	100	186	100	1.00	35	28	35	29
33	1390	100	260	100	197	100	1.00	36	29	36	30
34	1505	100	270	100	209	100	1.00	37	30	37	31
35	1625	100	280	100	221	100	1.00	38	31	38	32
36	1750	100	290	100	234	100	1.00	39	32	39	33
37	1880	100	300	100	247	100	1.00	40	33	40	34
38	2015	100	310	100	261	100	1.00	41	34	41	35
39	2155	100	320	100	275	100	1.00	42	35	42	36
40	2300	100	330	100	290	100	1.00	43	36	43	37
41	2450	100	340	100	305	100	1.00	44	37	44	38
42	2605	100	350	100	320	100	1.00	45	38	45	39
43	2765	100	360	100	336	100	1.00	46	39	46	40
44	2930	100	370	100	352	100	1.00	47	40	47	41
45	3100	100	380	100	369	100	1.00	48	41	48	42
46	3275	100	390	100	386	100	1.00	49	42	49	43
47	3455	100	400	100	404	100	1.00	50	43	50	44
48	3640	100	410	100	422	100	1.00	51	44	51	45
49	3830	100	420	100	441	100	1.00	52	45	52	46
50	4025	100	430	100	460	100	1.00	53	46	53	47

1 If  $\bar{p}$  exceeds 0.50, read  $1.00 - \bar{p}$  = fraction observed and subtract each confidence limit from 100.

Table 4.—Accumulative distribution of chi-square

Degrees of freedom	Probability of a greater value												
	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
1	0.01	0.02	0.05	0.10	0.20	0.40	0.80	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.20	0.40	0.80	2.77	4.01	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.34	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.36	13.36	15.51	17.53	20.09	21.98
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.58	14.68	17.02	19.02	21.67	23.58
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.59	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.58	44.16
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.39	42.96	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.98
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.13	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.59	15.31	16.91	18.91	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.33	16.07	17.71	19.71	23.56	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.40	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.80	55.70	59.34	63.69	66.77
50	27.99	29.71	32.30	34.75	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17

Reproduced by permission of the author and publishers from table 1.14.1 of Snedecor's Statistical Methods (ed. 5), Iowa State University Press. Permission has also been given by the editor and trustees of Biometrics.

Table 5.—Confidence intervals for binomial distribution (continued)

Number observed	99-percent interval												
	Size of sample, n						Fraction observed $\bar{p}$ <sup>1</sup>	Size of sample					
	10	15	20	30	50	100		250	1000				
0	41	0	30	0	23	0	.00	0	5	0	2	0	1
1	54	0	40	0	32	0	.01	0	7	0	5	0	2
2	65	1	49	1	39	0	.02	0	9	1	6	1	3
3	74	4	56	2	45	1	.03	0	10	1	7	2	4
4	81	5	63	4	51	3	.04	0	12	2	9	3	6
5	87	8	69	6	56	4	.05	0	13	2	10	3	7
6	92	12	74	8	61	6	.06	2	14	3	11	4	8
7	96	16	79	11	66	8	.07	2	16	3	13	5	9
8	99	21	84	15	70	10	.08	3	17	4	14	6	10
9	100	26	88	18	74	12	.09	3	18	5	15	7	12
10	100	31	92	22	78	14	.10	4	19	6	16	8	13
11	100	37	95	26	82	16	.11	4	20	6	17	9	14
12	100	44	98	30	85	18	.12	5	21	7	18	9	15
13	100	51	99	34	88	20	.13	6	23	8	19	10	16
14	100	59	100	39	92	24	.14	6	24	9	20	11	17
15	100	67	100	44	94	26	.15	7	26	9	22	12	18
16	100	75	100	49	96	29	.16	8	27	10	23	13	19
17	100	83	100	55	98	32	.17	9	29	11	24	14	20
18	100	91	100	61	99	35	.18	9	30	12	25	15	21
19	100	99	100	68	100	38	.19	10	31	13	26	16	22
20	100	107	100	77	100	42	.20	11	32	14	27	17	23
21	100	115	100	84	100	45	.21	12	33	15	28	18	24
22	100	123	100	90	100	48	.22	12	34	16	30	19	26
23	100	131	100	95	100	52	.23	13	35	17	31	20	27
24	100	139	100	100	100	56	.24	14	36	18	32	21	28
25	100	147	100	105	100	60	.25	15	38	19	33	22	29
26	100	155	100	110	100	64	.26	16	39	20	34	23	30
27	100	163	100	115	100	68	.27	16	40	20	35	23	31
28	100	171	100	120	100	72	.28	17	41	21	36	24	32
29	100	179	100	125	100	76	.29	18	42	22	37	25	33
30	100	187	100	130	100	80	.30	19	43	23	38	26	34
31	100	195	100	135	100	84	.31	20	44	24	39	27	35
32	100	203	100	140	100	88	.32	21	45	25	40	28	36
33	100	211	100	145	100	92	.33	21	46	26	41	29	37
34	100	219	100	150	100	96	.34	22	47	26	42	30	38
35	100	227	100	155	100	100	.35	23	48	27	43	31	39
36	100	235	100	160	100	104	.36	24	49	28	44	32	40
37	100	243	100	165	100	108	.37	25	50	29	45	33	41
38	100	251	100	170	100	112	.38	26	51	30	46	34	42
39	100	259	100	175	100	116	.39	27	52	31	47	35	43
40	100	267	100	180	100	120	.40	28	53	32	48	36	44
41	100	275	100	185	100	124	.41	29	54	33	50	37	45
42	100	283	100	190	100	128	.42	30	55	34	51	38	46
43	100	291	100	195	100	132	.43	30	56	35	52	39	47
44	100	299	100	200	100	136	.44	31	57	36	53	40	48
45	100	307	100	205	100	140	.45	32	58	37	54	41	49
46	100	315	100	210	100	144	.46	33	59	38	55	42	50
47	100	323	100	215	100	148	.47	34	60	39	56	43	51
48	100	331	100	220	100	152	.48	35	61	40	57	44	52
49	100	339	100	225	100	156	.49	36	62	41	58	45	53
50	100	347	100	230	100	160	.50	37	63	42	59	46	54

<sup>1</sup> If  $\bar{p}$  exceeds 0.50, read  $1.00 - \bar{p}$  = fraction observed and subtract each confidence limit from 100.  
 Reproduced by permission of the author and publishers from table 1.3.1 of Snedecor's *Statistical Methods* (ed. 5), Iowa State University Press.

Table 6.—Arc sine transformation

%	0	1	2	3	4	5	6	7	8	9
0.0	0	0.57	0.81	0.99	1.15	1.28	1.40	1.52	1.62	1.72
0.1	1.81	1.90	1.99	2.07	2.14	2.22	2.28	2.36	2.43	2.50
0.2	2.56	2.63	2.69	2.75	2.81	2.87	2.92	2.98	3.03	3.09
0.3	3.14	3.19	3.24	3.29	3.34	3.39	3.44	3.49	3.53	3.58
0.4	3.63	3.67	3.72	3.76	3.80	3.85	3.89	3.93	3.97	4.01
0.5	4.05	4.09	4.13	4.17	4.21	4.25	4.29	4.33	4.37	4.40
0.6	4.44	4.48	4.52	4.55	4.59	4.62	4.66	4.69	4.73	4.76
0.7	4.80	4.83	4.87	4.90	4.93	4.97	5.00	5.03	5.07	5.10
0.8	5.13	5.16	5.20	5.23	5.26	5.29	5.32	5.35	5.38	5.41
0.9	5.44	5.47	5.50	5.53	5.56	5.59	5.62	5.65	5.68	5.71
1	5.74	6.02	6.29	6.55	6.80	7.04	7.27	7.49	7.71	7.92
2	8.13	8.33	8.53	8.72	8.91	9.10	9.28	9.46	9.63	9.81
3	9.98	10.14	10.31	10.47	10.63	10.78	10.94	11.09	11.24	11.39
4	11.54	11.68	11.83	11.97	12.11	12.25	12.39	12.52	12.66	12.79
5	12.92	13.05	13.18	13.31	13.44	13.56	13.69	13.81	13.94	14.06
6	14.18	14.30	14.42	14.54	14.65	14.77	14.89	15.00	15.12	15.23
7	15.34	15.45	15.56	15.68	15.79	15.89	16.00	16.11	16.22	16.32
8	16.43	16.54	16.64	16.74	16.85	16.95	17.05	17.16	17.26	17.36
9	17.46	17.56	17.66	17.76	17.85	17.95	18.05	18.15	18.24	18.34
10	18.44	18.53	18.63	18.72	18.81	18.91	19.00	19.09	19.19	19.28
11	19.37	19.46	19.55	19.64	19.73	19.82	19.91	20.00	20.09	20.18
12	20.27	20.36	20.44	20.53	20.62	20.70	20.79	20.88	20.96	21.05
13	21.13	21.22	21.30	21.39	21.47	21.56	21.64	21.72	21.81	21.89
14	21.97	22.06	22.14	22.22	22.30	22.38	22.46	22.55	22.63	22.71
15	22.79	22.87	22.95	23.03	23.11	23.19	23.26	23.34	23.42	23.50
16	23.58	23.66	23.73	23.81	23.89	23.97	24.04	24.12	24.20	24.27
17	24.35	24.43	24.50	24.58	24.65	24.73	24.80	24.88	24.95	25.03
18	25.10	25.18	25.25	25.33	25.40	25.48	25.55	25.62	25.70	25.77
19	25.84	25.92	25.99	26.06	26.13	26.21	26.28	26.35	26.42	26.49
20	26.56	26.64	26.71	26.78	26.85	26.92	26.99	27.06	27.13	27.20
21	27.27	27.35	27.42	27.49	27.56	27.63	27.69	27.76	27.83	27.90
22	27.97	28.04	28.11	28.18	28.25	28.32	28.38	28.45	28.52	28.59
23	28.66	28.73	28.79	28.86	28.93	29.00	29.06	29.13	29.20	29.27
24	29.33	29.40	29.47	29.53	29.60	29.67	29.73	29.80	29.87	29.93
25	30.00	30.07	30.13	30.20	30.26	30.33	30.40	30.46	30.53	30.59
26	30.66	30.72	30.79	30.85	30.92	30.98	31.05	31.11	31.18	31.24
27	31.31	31.37	31.44	31.50	31.56	31.63	31.69	31.76	31.82	31.88
28	31.95	32.01	32.08	32.14	32.20	32.27	32.33	32.39	32.46	32.52
29	32.58	32.65	32.71	32.77	32.83	32.90	32.96	33.02	33.09	33.15
30	33.21	33.27	33.34	33.40	33.46	33.52	33.58	33.65	33.71	33.77
31	33.83	33.89	33.96	34.02	34.08	34.14	34.20	34.27	34.33	34.39
32	34.45	34.51	34.57	34.63	34.70	34.76	34.82	34.88	34.94	35.00
33	35.06	35.12	35.18	35.24	35.30	35.37	35.43	35.49	35.55	35.61
34	35.67	35.73	35.79	35.85	35.91	35.97	36.03	36.09	36.15	36.21
35	36.27	36.33	36.39	36.45	36.51	36.57	36.63	36.69	36.75	36.81
36	36.87	36.93	36.99	37.05	37.11	37.17	37.23	37.29	37.35	37.41
37	37.47	37.52	37.58	37.64	37.70	37.76	37.82	37.88	37.94	38.00
38	38.06	38.12	38.17	38.23	38.29	38.35	38.41	38.47	38.53	38.59
39	38.65	38.70	38.76	38.82	38.88	38.94	39.00	39.06	39.11	39.17
40	39.22	39.27	39.32	39.38	39.44	39.50	39.56	39.62	39.67	39.73
41	39.82	39.87	39.93	39.99	40.05	40.11	40.16	40.22	40.28	40.34
42	40.40	40.46	40.51	40.57	40.63	40.69	40.74	40.80	40.86	40.92
43	40.98	41.03	41.09	41.15	41.21	41.27	41.32	41.38	41.44	41.50
44	41.55	41.61	41.67	41.73	41.78	41.84	41.90	41.96	42.02	42.07
45	42.13	42.19	42.25	42.30	42.36	42.42	42.48	42.53	42.59	42.65
46	42.71	42.76	42.82	42.88	42.94	42.99	43.05	43.11	43.17	43.22
47	43.28	43.34	43.39	43.45	43.51	43.57	43.62	43.68	43.74	43.80
48	43.85	43.91	43.97	44.03	44.08	44.14	44.20	44.25	44.31	44.37
49	44.43	44.48	44.54	44.60	44					

Table 6.—Arc sine transformation (continued)

%	0	1	2	3	4	5	6	7	8	9
55	47.87	47.93	47.98	48.04	48.10	48.16	48.22	48.27	48.33	48.39
56	48.45	48.50	48.56	48.62	48.68	48.73	48.79	48.85	48.91	48.97
57	49.02	49.08	49.14	49.20	49.26	49.31	49.37	49.43	49.49	49.54
58	49.60	49.66	49.72	49.78	49.84	49.89	49.95	50.01	50.07	50.13
59	50.18	50.24	50.30	50.36	50.42	50.48	50.53	50.59	50.65	50.71
60	50.77	50.83	50.89	50.94	51.00	51.06	51.12	51.18	51.24	51.30
61	51.35	51.41	51.47	51.53	51.59	51.65	51.71	51.77	51.83	51.88
62	51.94	52.00	52.06	52.12	52.18	52.24	52.30	52.36	52.42	52.48
63	52.53	52.59	52.65	52.71	52.77	52.83	52.89	52.95	53.01	53.07
64	53.13	53.19	53.25	53.31	53.37	53.43	53.49	53.55	53.61	53.67
65	53.73	53.79	53.85	53.91	53.97	54.03	54.09	54.15	54.21	54.27
66	54.33	54.39	54.45	54.51	54.57	54.63	54.70	54.76	54.82	54.88
67	54.94	55.00	55.06	55.12	55.18	55.24	55.30	55.37	55.43	55.49
68	55.55	55.61	55.67	55.73	55.80	55.86	55.92	55.98	56.04	56.11
69	56.17	56.23	56.29	56.35	56.42	56.48	56.54	56.60	56.66	56.73
70	56.79	56.85	56.91	56.98	57.04	57.10	57.17	57.23	57.29	57.35
71	57.42	57.48	57.54	57.61	57.67	57.73	57.80	57.86	57.92	57.99
72	58.05	58.12	58.18	58.24	58.31	58.37	58.44	58.50	58.56	58.63
73	58.69	58.76	58.82	58.89	58.95	59.02	59.08	59.15	59.21	59.28
74	59.34	59.41	59.47	59.54	59.60	59.67	59.74	59.80	59.87	59.93
75	60.00	60.07	60.13	60.20	60.27	60.33	60.40	60.47	60.53	60.60
76	60.67	60.73	60.80	60.87	60.94	61.00	61.07	61.14	61.21	61.27
77	61.34	61.41	61.48	61.55	61.62	61.68	61.75	61.82	61.89	61.96
78	62.03	62.10	62.17	62.24	62.31	62.37	62.44	62.51	62.58	62.65
79	62.72	62.80	62.87	62.94	63.01	63.08	63.15	63.22	63.29	63.36
80	63.44	63.51	63.58	63.65	63.72	63.79	63.87	63.94	64.01	64.08
81	64.16	64.23	64.30	64.38	64.45	64.52	64.60	64.67	64.75	64.82
82	64.90	64.97	65.05	65.12	65.20	65.27	65.35	65.42	65.50	65.57
83	65.65	65.73	65.80	65.88	65.96	66.03	66.11	66.19	66.27	66.34
84	66.42	66.50	66.58	66.66	66.74	66.81	66.89	66.97	67.05	67.13
85	67.21	67.29	67.37	67.45	67.54	67.62	67.70	67.78	67.86	67.94
86	68.03	68.11	68.19	68.28	68.36	68.44	68.53	68.61	68.70	68.78
87	68.87	68.95	69.04	69.12	69.21	69.30	69.38	69.47	69.56	69.64
88	69.73	69.82	69.91	70.00	70.09	70.18	70.27	70.36	70.45	70.54
89	70.63	70.72	70.81	70.91	71.00	71.09	71.19	71.28	71.37	71.47
90	71.56	71.66	71.76	71.85	71.95	72.05	72.15	72.24	72.34	72.44
91	72.54	72.64	72.74	72.84	72.95	73.05	73.15	73.26	73.36	73.46
92	73.57	73.68	73.78	73.89	74.00	74.11	74.21	74.32	74.44	74.55
93	74.66	74.77	74.88	75.00	75.11	75.23	75.35	75.46	75.58	75.70
94	75.82	75.94	76.06	76.19	76.31	76.44	76.56	76.69	76.82	76.95
95	77.08	77.21	77.34	77.48	77.61	77.75	77.89	78.03	78.17	78.32
96	78.46	78.61	78.76	78.91	79.06	79.22	79.37	79.53	79.69	79.86
97	80.02	80.19	80.37	80.54	80.72	80.90	81.09	81.28	81.47	81.67
98	81.87	82.08	82.29	82.51	82.73	82.96	83.20	83.45	83.71	83.98
99.0	84.26	84.29	84.32	84.35	84.38	84.41	84.44	84.47	84.50	84.53
99.1	84.56	84.59	84.62	84.65	84.68	84.71	84.74	84.77	84.80	84.84
99.2	84.87	84.90	84.93	84.97	85.00	85.03	85.07	85.10	85.13	85.17
99.3	85.20	85.24	85.27	85.31	85.34	85.38	85.41	85.45	85.48	85.52
99.4	85.56	85.60	85.63	85.67	85.71	85.75	85.79	85.83	85.87	85.91
99.5	85.95	85.99	86.03	86.07	86.11	86.15	86.20	86.24	86.28	86.33
99.6	86.37	86.42	86.47	86.51	86.56	86.61	86.66	86.71	86.76	86.81
99.7	86.86	86.91	86.97	87.02	87.08	87.13	87.19	87.25	87.31	87.37
99.8	87.44	87.50	87.57	87.64	87.71	87.78	87.86	87.93	88.01	88.10
99.9	88.19	88.28	88.38	88.48	88.60	88.72	88.85	89.01	89.19	89.43
100.0	90.00	—	—	—	—	—	—	—	—	—

Reproduced by permission of the author and publishers from table 11.12.1 of Snedecor's *Statistical Methods* (ed. 5), Iowa State University Press. Permission has also been granted by the original author, Dr. C. I. Bliss of the Connecticut Agricultural Experiment Station.

Table 7.—Significance of correlation coefficients [CORRELATION COEFFICIENTS AT THE 5% AND 1% LEVELS OF SIGNIFICANCE]

Degrees of freedom	Degrees of freedom	
	5%	1%
1	.997	1.000
2	.950	.990
3	.878	.959
4	.811	.917
5	.754	.874
6	.707	.834
7	.666	.798
8	.632	.765
9	.602	.735
10	.576	.708
11	.553	.684
12	.532	.661
13	.514	.641
14	.497	.623
15	.482	.606
16	.468	.590
17	.456	.575
18	.444	.561
19	.433	.549
20	.423	.537
21	.413	.526
22	.404	.515
23	.396	.505

Reproduced by permission of the author and publishers from table 7.6.1 of Snedecor's *Statistical Methods* (ed. 5), Iowa State University Press. Permission has also been granted by the literary executor of the late Professor Sir Ronald A. Fisher and Oliver and Boyd, Ltd., publishers, for the portion of the table taken from Dr. Fisher's table V.A. in *Statistical Methods for Research Workers*.