STATISTICAL GUIDES

by

Harry V. Wiant, Jr.
Professor of Forestry

1981

Division of Forestry West Virginia University

Morgantown, WV 26506

Contents

Parametric Statistics

Sum of squares of deviations..... 1
Sample standard deviation..... 2
Sample standard error..... 5
Paired t-test..... 6
Group t-test..... 8
Analysis of variance (one-way)..... 10
Multiple comparisons..... 14
Analysis of variance (two-way)..... 15
Factorial analysis..... 18
Linear regression..... 22
Covariance analysis..... 27
Multiple regression..... 30
Curvilinear regression..... 35
Testing accuracy..... 39

Nonparametric Statistics

Chi-square (hypothetical ratio)..... 41
Chi-square (test of independence)..... 42
Chi-square test of dependence..... 45
Chi-square measure of dependence..... 46
Median test..... 47
McNemar test of change..... 48
Binomial test..... 49
Wilcoxon test..... 50
Mann-Whitney test..... 51
Kruskal-Wallis test..... 54
Friedman test..... 56
Correlation of ranked data..... 59

SUM OF SQUARES OF DEVIATIONS

In statistical calculations, one is frequently calculating the sum of squares of deviations from a mean value (symbolized as $\varepsilon(X - \bar{x})^2$ or εx^2). The operation is easily understood when done as follows:

	X	Deviation $(X - \bar{x})$	Squared Deviation $(X - \bar{x})^2$
	10	0	0
	5	-5	25
	15	5	25
	10	0	0
Total (εX)	40	0	εx^2 50
Mean $(\bar{x})$	10		

The same value can be obtained much more readily with a calculator as follows:

$$\varepsilon x^2 = \varepsilon X^2 - (\varepsilon X)^2/n = (10)^2 + (5)^2 + (15)^2 + (10)^2 - (40)^2/4 =$$

Note: (1) n = number of observations.
(2) $(\varepsilon X)^2/n$ is known as the correction term = C.

SAMPLE STANDARD DEVIATION

The sample standard deviation (S) provides a measure of dispersion with which to qualify the average of a normally distributed population (a population yielding a bell-shaped curve peaking at the mean). In such a population, about 66 percent of the individuals would be expected to fall within the values of the mean plus or minus one standard deviation; the mean plus or minus two standard deviations should include approximately 95 percent of the population (Fig. 1).

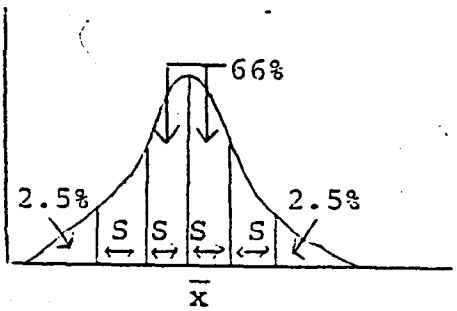


Fig. 1. Normal distribution with mean and standard deviation.

An example of calculation procedures is as follows:

Height of randomly selected seedlings (in.)
(X)
10
5
15
10
Total (εX) = 40
Mean ($\bar{x}$) = 10

$$S = \sqrt{\frac{\varepsilon x^2}{n-1}}$$

$$\varepsilon x^2 = \varepsilon X^2 - (\varepsilon X)^2/n =$$

$$(10)^2 + (5)^2 + \ldots (10)^2 - \frac{(40)^2}{4} = 50$$

$$n-1 = 4-1 = 3$$

$$S = \sqrt{\frac{50}{3}} = 4.08$$

In this example, we would expect about 66 percent of the population of seedlings to fall between 10 ± 4.1 inches, or 95 percent between 10 ± 8.2 inches.

SAMPLE STANDARD ERROR OF THE MEAN

The sample standard error of the mean ($S_{\bar{x}}$) calculated from a random sample of a given size provides an estimate of the spread within which a sample of means will fall. The mean plus or minus one standard error should contain 66 percent of other mean estimates; the mean plus or minus two standard errors includes about 95 percent of the other mean estimates (Fig. 1).

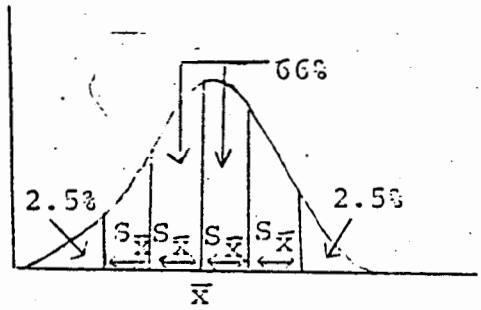


Fig. 1. Normal distribution with mean and standard error.

An example of calculation procedures is as follows:

Height of randomly selected seedlings (in.)

(X)
10
5
15
10
$\varepsilon X = 40$
$\bar{x} = 10$

$$S_{\bar{x}} = \sqrt{\frac{S^2}{n}}$$

$$S^2 = \frac{\varepsilon x^2}{n-1}$$

$$\varepsilon x^2 = \varepsilon X^2 - (\varepsilon X)^2/n =$$

$$(10)^2+(5)^2+\ldots(10)^2-\frac{(40)^2}{4} = 50$$

$$n-1 = 4-1 = 3$$

$$S^2 = \frac{50}{3} = 16.667$$

$$S_{\bar{x}} = \sqrt{\frac{16.667}{4}} = 2.04$$

In this example, we would expect the means from other samples of the same size to fall within 10 ± 2.0 inches approximately 66 percent of the time, or 10 ± 4.1 inches about 95 percent of the time.

PAIRED t-TEST

The t-Test is valid with a random sample from a normally distributed population. It may be used to test the probability that differences found are real or merely variations one could expect in sampling from a normal population. Data are paired only when there is sufficient reason to do so, as there would be when sampling the same population with different techniques, or in different places; when using twin animals in an experiment, etc.

An example of calculation procedures is as follows:

Growth of various tree species with and without fertilizer applications (in.)

Species	With (X_1)	Without (X_2)	Difference ($d=X_1-X_2$)
1	10	5	5
2	15	10	5
3	20	10	10
4	20	20	0
5	40	45	-5
Total	105	90	15
Mean	21	18	

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_d^2}{n}}}$$

$$s_d^2 = \frac{\varepsilon d^2 - \frac{(\varepsilon d)^2}{n}}{n-1} = \frac{(5)^2+(5)^2+\ldots(-5)^2-\frac{(15)^2}{5}}{5-1} = 32.5$$

$$t = \frac{21-18}{\sqrt{\frac{32.5}{5}}} = \frac{3}{2.55} = 1.176$$

Degrees of freedom (d.f.) = (number of pairs) - 1 = 5 - 1 = 4

With d.f. = 4, our calculated t-value would have to equal or exceed 2.776 for the difference between the means to be significant at the 5 percent level. Therefore, we conclude the two means did not differ significantly.

GROUP t-TEST

A collection of individuals may be assigned at random to two groups, each group receiving a different treatment. In this case, a difference between the two sample means may be assessed by the t-test to determine significance.

An example of calculation procedures is as follows:

Annual growth (in.) of a tree species with and without fertilizer applications

With (X_1)	Without (X_2)
10	5
15	10
20	10
20	20
40	45
$\varepsilon X_1 = 105$	$\varepsilon X_2 = 90$
$\bar{x}_1 = 21$	$\bar{x}_2 = 18$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S^2(n_1+n_2)}{(n_1)(n_2)}}}$$ Note: n_1 and n_2 need not be equal.

$$S^2 = \frac{\varepsilon x_1^2 + \varepsilon x_2^2}{(n_1-1)+(n_2-1)}$$

$$\varepsilon x_1^2 = \varepsilon X_1^2 - (\varepsilon X_1)^2/n = (10)^2+(15)^2+\ldots(40)^2-\frac{(105)^2}{5} = 520$$

$$\varepsilon x_2^2 = \varepsilon X_2^2 - (\varepsilon X_2)^2/n = (5)^2+(10)^2+\ldots(45)^2-\frac{(90)^2}{5} = 1030$$

$$S^2 = \frac{520 + 1030}{(5-1)+(5-1)} = 193.75$$

$$t = \frac{21 - 18}{\sqrt{\frac{193.75(5+5)}{(5)(5)}}} = 0.34$$

Degrees of freedom (d.f.) $= (n_1-1)+(n_2-1) = (5-1)+(5-1) = 8$

With d.f. = 8, our calculated t-value would have to equal or exceed 2.306 for the difference between the means to be significant at the 5 percent level. Therefore, we conclude the two means did not differ significantly.

ANALYSIS OF VARIANCE
(one-way classification)

Studies such as that described for the group t-test may involve more than two groups. An analysis of variance may then be made.

Species	Annual diam. growth (in.) of various species on the same site				Totals	Averages
A	2	3	2	1	8	2
B	1	1	1	1	4	1
C	3	3	4	2	12	3
					24	

1. Correction term (C) = $(\varepsilon X)^2/an = (24)^2/12 = 48$
   a = number of treatments (species) = 3
   n = number of observations per treatment = 4

2. Total sum of squares = $\varepsilon X^2-C = (2)^2+(3)^2+\ldots(2)^2-48 = 12$

3. Treatment (species) sum of squares =

$$\frac{\varepsilon(\text{treatment totals}^2)}{n}-C =$$

$$\frac{(8)^2+(4)^2+(12)^2}{4}-48 = 8$$

Note: In these operations the divisor is always the number of observations that make up each total being squared.

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatment (species)	a-1=2	8	8/2=4	4/0.44=9.09*
Error	11-2=9	12-8=4	4/9=0.44	
Total	an-1=11	12		

We conclude there was a highly significant difference in diameter growth of these species.

Note: Had there been unequal numbers of observations for the treatments (species), calculation procedures would have been changed as follows:

1. $C = (\epsilon X)^2 / n_A + n_B + n_C$

   e.g., n_A = number of observations for species A

2. Treatment (species) sum of squares =

$$\frac{(\text{treatment total A})^2}{n_A} + \frac{(\text{treatment total B})^2}{n_B} + \frac{(\text{treatment total C})^2}{n_C} - C$$

3.

Source of variation	Degrees of freedom
Treatment (species)	a - 1
Error	total d.f. - treat. d.f.
Total	$n_A+n_B+n_C-1$

MULTIPLE COMPARISONS

Significant differences among means in the previous example could be detected by the Tukey test. Means must differ by a value equal to or greater than $QS_{\bar{x}}$ to be significant. Q is read from a table according to the number of treatments (3 in our example) and degrees of freedom in the error term (9 in our example).

$Q = 3.95$

$$S_{\bar{x}} = \sqrt{\frac{\text{error mean square}}{n}} = \sqrt{\frac{0.44}{4}} = 0.332$$

$$QS_{\bar{x}} = (3.95)(0.332) = 1.31$$

Species	$\bar{x}$	$\bar{x} - 1$	$\bar{x} - 2$
C	3	2*	1
A	2	1	
B	1		

We conclude the diameter growth of species C and B differed significantly. Other mean comparisons were not significant.

Had there been unequal numbers of observations for treatments (species), the difference between two means must equal or exceed $tS_{\bar{d}}$ to be significant, where:

t = t-value for chosen significance level and error d.f

$$S_{\bar{d}} = \sqrt{S^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

S^2 = error mean square

n_i and n_j = number of observations in the two means being compared

ANALYSIS OF VARIANCE
(two-way classification)

Treatments randomized in each of several blocks may be analyzed by the paired t-test if there are only two treatments or the following analysis of variance if there are two or more treatments.

Annual diam. growth (in.) of various species on different blocks

Species	Block I	II	III	IV	Totals
A	2	3	2	1	8
B	1	1	1	1	4
C	3	3	4	2	12
	6	7	7	4	24

1. Correction term (C) = $(\varepsilon X)^2/an = (24)^2/12 = 48$

2. Total sum of squares = $\varepsilon X^2 - C = (2)^2+(3)^2+\ldots(2)^2-48 = 12$

3. Treatment (species) sum of squares =

$$\frac{\varepsilon(\text{treatment totals}^2)}{n} - C =$$

$$\frac{(8)^2+(4)^2+(12)^2}{4} - 48 = 8$$

4. Blocks sum of squares = $\frac{\varepsilon(\text{block totals}^2)}{a} - C =$

$$\frac{(6) + (7)^2 +\ldots(4)^2}{3} - 48 = 2$$

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Blocks	n-1=3	2	2/3=0.667	
Treatment (species)	a-1=2	8	8/2=4	4/0.333=12.0
Error	2x3 or 11-(3+2)=6	12-(8+2)= 2	2/6=0.333	
Total	an-1=11	12		

We conclude there was a highly significant difference in diameter growth of these species. Means could now be compared to determine which ones differ significantly.

FACTORIAL ANALYSIS

Factorial analysis is a type of analysis of variance which detects the interaction of one factor on another when treatments are applied at different levels. This design can provide much information and is commonly used.

For example, consider an experiment where two tree species are grown at two spacings on three sites (blocks) and height growth for a certain period is to be analyzed.

Species	Spacing	Blocks 1	2	3	Totals
A	I	2	1	2	5
	II	3	2	3	8
B	I	1	2	1	4
	II	1	1	1	3
Totals		7	6	7	20

1. $C = (20)^2/12 = 33.3$

2. Total ss $= (2)^2 + \ldots (1)^2 - C = 6.7$

3. Treatment ss (over-all) $= \frac{(5)^2 + \ldots (3)^2}{3} - C = 4.7$

4. Blocks ss $= \frac{(7)^2 + \ldots (7)^2}{4} - C = 0.2$

5. Error ss = total ss - (treat. ss + rep. ss) =

   $6.7 - (4.7 + 0.2) = 1.8$

Summary Table A

Spacing	Species A	Species B	Totals
I	5	4	9
II	8	3	11
Totals	13	7	20

1. Total ss in A $= \frac{(5)^2 + \ldots (3)^2}{3} - C = 4.7$

2. Species ss $= \frac{(13)^2 + (7)^2}{6} - C = 3.0$

3. Spacing ss $= \frac{(9)^2 + (11)^2}{6} - C = 0.4$

4. Species x spacing ss = interaction = total ss in A - (species ss + spacing ss) = 4.7 - (3.0+0.4) = 1.3

Source of Variation	d.f.	s.s.	m.s.	F
Total	11	6.7		
Blocks	2	0.2	0.1	$\frac{0.1}{0.3}$ = 0.3 n.s.
Treat.	(3)	(4.7)	(1.6)	($\frac{1.6}{0.3}$ = 5.3 [1]*)
Species	1	3.0	3.0	$\frac{3.0}{0.3}$ = 10*
Spacing	1	0.4	0.4	$\frac{0.4}{0.3}$ = 1.3 n.s.
Species x Spacing	1	1.3	1.3	$\frac{1.3}{0.3}$ = 4.3 n.s.
Error	6	1.8	0.3	

[1] Significance here is not very meaningful. One is more interested in specific treatments (species, etc.).

Therefore, this study revealed a significant difference in the growth of the species (species A grew significantly more than species B); no other factors were significant.

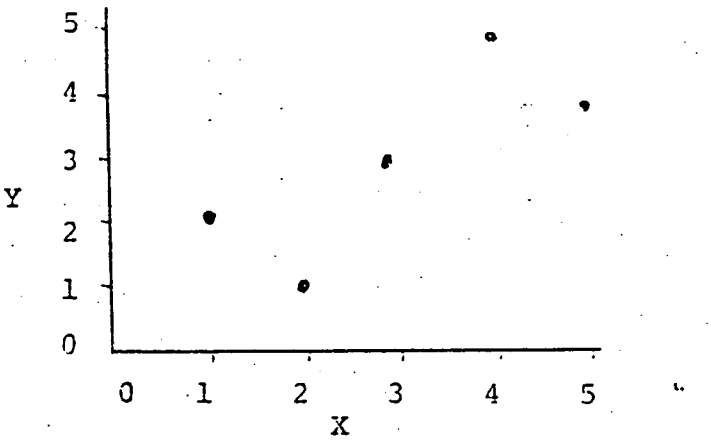
LINEAR REGRESSION

Linear regressions are often used for predicting one quantity by its relation to another.

An example of calculation procedures is as follows:

Annual height growth of a tree species with different amounts of nitrogen fertilizer

Independent Variable Pounds of fertilizer (X)		LOG-	Dependent Variable Height growth (in.) (Y)		LOG
	1	0		2	.693147
	2	.693147		1	0
	3	1.09861		3	1.09861
	4	1.38629		5	1.60944
	5	1.60944		4	1.38629
Total	15			15	
Mean	3			3	



Plotting of height growth over amount of fertilizer

1. Deviations from regression:

$$\varepsilon x^2 = \varepsilon(X^2) - (\varepsilon X)^2/n = (1^2 + 2^2 + \ldots 5^2) - (15)^2/5 =$$

$$55 - 225/5 = 55 - 45 = 10$$

$$\varepsilon y^2 = \varepsilon(Y^2) - (\varepsilon Y)^2/n = (2^2 + 1^2 + \ldots 4^2) - (15)^2/5 =$$

$$55 - 45 = 10$$

$$\varepsilon xy = \varepsilon XY - (\varepsilon X)(\varepsilon Y)/n = (1)(2)+(2)(1)+\ldots(5)(4) -$$

$$(15)(15)/5 = 53 - 225/5 = 53 - 45 = 8$$

2. The sample regression coefficient indicates the rate at which the regression line slopes upward or downward.

$$b = \varepsilon xy/\varepsilon x^2 = 8/10 = 0.8$$

This means the height growth increased 0.8 inches for each additional pound of fertilizer. In other words, b indicates the change in the dependent variable for an unit change in the independent variable.

3. a = mean of the population when X = 0

$$a = \bar{y} - b\bar{x} = 3 - (0.8)(3) = 3 - 2.4 = 0.6$$

4. The regression equation is:

$$Y = a + bX = 0.6 + 0.8X$$

5. To plot the regression line, the mean points (3 pounds and 3 inches) are fixed. With 1 pound of fertilizer,

$$Y = 0.6 + 0.8\ (1) = 1.4$$

With 5 pounds of fertilizer,

$$Y = 0.6 + 0.8\ (5) = 4.6$$

By plotting these points, the line can be drawn.

6. The proportion of the total variation in height growth accounted for by the fertilizer relationship can be determined.

$$\text{Coefficient of determination } (r^2) = \frac{(\varepsilon xy)^2/\varepsilon x^2}{\varepsilon y^2} = \frac{(8)^2/10}{10} = 0.64$$

In other words, we can say that 64 percent of the variation in height growth was accounted for by the rate of fertilizer application.

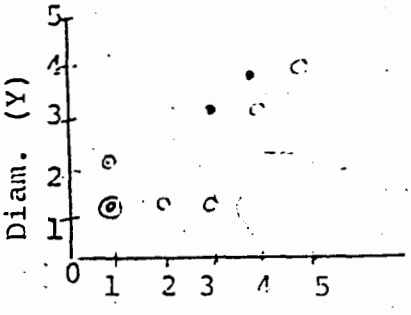
7. An analysis of variance can be made to determine if the linear relationship is significant.

Source of variation	d.f.	Sum of squares	Mean square	F
Due to regression	1	$(\varepsilon xy)^2/\varepsilon x^2=(8)^2/10=6.4$	6.4	5.3 n.s.
Error (unexplained)	3	εy^2-above=10-6.4=3.6	1.2	$=s_{y\cdot x}^2$
Total	4	$\varepsilon y^2 = 10$		

The F-value for the degrees of freedom in this experiment would have to exceed 10.13 to be significant at the 5 percent level. Therefore, we conclude the linear relationship is not significant.

COVARIANCE ANALYSIS

Covariance analysis provides a means for testing the differences between slopes (parallelism) and elevations (heights) of regression lines. Consider the following example:



Because of the variation of points around the lines, is there statistical reason to have more than one regression line (with one slope, one elevation)? (Actually, one would doubt that these relations are linear, but we'll assume they are.)

Figure 1. Diameter (cm) at ground level as related to age (yrs.) of pine (·) and spruce (o) seedlings.

1.

	pine (·)		spruce (o)	
	X	Y	X	Y
	1	0	1	1
	1	1	2	1
	1	2	3	1
	3	3	4	3
	4	4	5	4
sum	10	10	15	10
mean	2	2	3	2

2. Regression for pine

$\varepsilon X^2 = (1)^2+\ldots(4)^2 = 28$ $\qquad \varepsilon Y^2 = (0)^2+\ldots(4)^2 = 30$

$C = (10)^2/5 = \underline{20}$ $\qquad C = (10)^2/5 = \underline{20}$

$\varepsilon x^2 = 8$ $\qquad \varepsilon y^2 = 10$

$\varepsilon XY = (1)(0)+\ldots(4)(4) = 28$

$C = (10)(10)/5 = \underline{20}$

$\varepsilon xy = 8$

Although not necessary, the regression equation for pine may be calculated.

$b = \varepsilon xy/\varepsilon x^2 = 8/8 = 1$

$a = \bar{y} - b\bar{x} = 2 - (1)2 = 0$

$Y = a + bX$

$Y = 0 + 1X$

3. Regression for spruce.

$\varepsilon X^2 = (1)^2+\ldots(5)^2 = 55$ $\qquad$ $\varepsilon Y^2 = (1)^2+\ldots(4)^2 = 28$

$C = (15)^2/5 = \underline{45}$ $\qquad$ $C = (10)^2/5 = \underline{20}$

$\varepsilon x^2 = 10$ $\qquad$ $\varepsilon y^2 = 8$

$\varepsilon XY = (1)(1)+\ldots(5)(4) = 30$

$C = (15)(10)/5 = \underline{30}$

$\varepsilon xy = 8$

Regression equation for spruce.

$b = 8/10 = 0.8$

$a = 2-(0.8)3 = -0.4$

$Y = -0.4 + 0.8X$

4. Regression with pine and spruce pooled.

$\varepsilon X^2 = 28 + 55 = 83$ $\qquad$ $\varepsilon Y^2 = 30 + 20 = 50$

$C = (25)^2/10 = \underline{62.5}$ $\qquad$ $C = (20)^2/10 = \underline{40}$

$\varepsilon x^2 = 20.5$ $\qquad$ $\varepsilon y^2 = 10$

$\varepsilon XY = 28 + 30 = 66$

$C = (25)(20)/10 = \underline{50}$

$\varepsilon xy = 16$

5. Analysis of covariance.

Line	Species	d.f.	εx^2	εxy	y^2	Deviations from regression: d.f.	$\varepsilon d.y.x^2$ [1]	$S\ y.x^2$ [2]
1.	pine	4	8	8	10	3	2	0.667
2.	spruce	4	10	8	8	3	1.6	0.533
3.	within					6	3.6	0.6 [3]
4.	reg. coef.					1 [8]	0.2 [5]	0.2
5.	common	8	18	16	18	7	3.8 [4]	0.543
6.	adj. means					1 [9]	1.7 [7]	1.7
7.	total	9	20.5	16	18	8	5.5 [6]	

Do calculations in following order of lines:
1, 2, 7, 3, 5, 4, 6.

[1] $\varepsilon y^2 - (\varepsilon xy)^2/\varepsilon x^2$

[2] $\varepsilon dy.x^2 \div$ d.f.

[3] 3.6/6 = 0.6

[4] $18 - (16)^2/18$

[5] common minus within

[6] $18 - (16)^2/20.5$

[7] 5.5 - 3.8

[8] (no. of regressions)-1

[9] (no. of regressions)-1

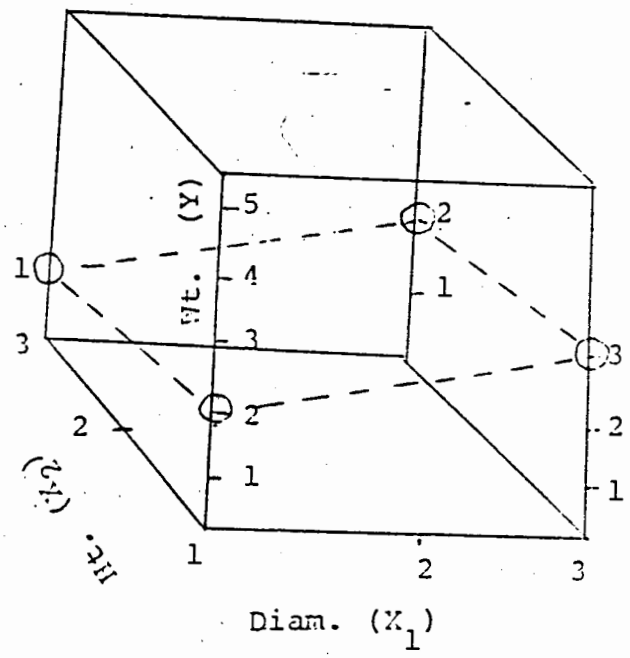
To test slope: $F = 0.2/0.6 = 0.33$ n.s. (d.f. = 1,6)

To test elevation: $F = 1.7/0.543 = 3.13$ n.s. (d.f. = 1,7)

Thus, there is no statistical evidence that the two species differ in the relationship examined. Based on this small sample, one regression equation (pooled results) is sufficient.

MULTIPLE REGRESSION

A multiple regression may be useful to determine a relationship between two or more independent variables (X_1, X_2, etc.) and one dependent variable (Y) for prediction or testing purposes. Straight-line relationships are assumed. Consider this example:



Note: This example has been designed to fit a plane perfectly. With more than two independent variables, such visual presentation is not possible.

1.

	Seedling diam. (X_1)	Seedling ht. (X_2)	Seedling wt. (Y)
	1	1	2
	1	3	1
	3	1	3
	3	3	2
sum	8	8	8
mean	2	2	2

2. Calculations:

$\varepsilon X_1^2 = (1)^2 + \ldots (3)^2 = 20$ $\qquad$ $\varepsilon X_2^2 = (1)^2 + \ldots (3)^2 = 20$

$C = (8)^2/4 = \underline{16}$ $\qquad$ $C = (8)^2/4 = \underline{16}$

$\varepsilon x_1^2 = 4$ $\qquad$ $\varepsilon x_2^2 = 4$

$\varepsilon Y^2 = (2)^2 + \ldots (2)^2 = 12$ $\qquad$ $\varepsilon X_1 Y = (1)(2) + \ldots (3)(2) = 12$

$C = (8)^2/4 = \underline{16}$ $\qquad$ $C = (8)(8)/4 = \underline{16}$

$\varepsilon y^2 = 2$ $\qquad$ $\varepsilon x_1 y = 2$

$\varepsilon X_1 X_2 = (1)(1) + \ldots (3)(3) = 16$ $\qquad$ $\varepsilon X_2 Y = (1)(2) + \ldots (3)(2) = 14$

$C = (8)(8)/4 = \underline{16}$ $\qquad$ $C = (8)(8)/4 = \underline{16}$

$\varepsilon x_1 x_2 = 0$ $\qquad$ $\varepsilon x_2 y = -2$

3. The multiple regression equation:

$$\hat{Y} = \bar{y} + b_1 (X_1 - \bar{x}_1) + b_2 (X_2 - \bar{x}_2)$$

$$b_1 = \frac{(\varepsilon x_2^2)(\varepsilon x_1 y) - (\varepsilon x_1 x_2)(\varepsilon x_2 y)}{D} = \frac{(4)(2) - (0)(-2)}{16} = 0.5$$

$$(D = (\varepsilon x_1^2)(\varepsilon x_2^2) - (\varepsilon x_1 x_2)^2 = (4)(4) - (0)^2 = 16)$$

$$b_2 = \frac{(\varepsilon x_1^2)(\varepsilon x_2 y) - (\varepsilon x_1 x_2)(\varepsilon x_1 y)}{D} = \frac{(4)(-2) - (0)(2)}{16} =$$

$$\frac{-8}{16} = -0.5$$

Therefore, $\hat{Y} = 2 + 0.5 (X_1 - 2) - 0.5 (X_2 - 2)$

$$\hat{Y} = 2 + 0.5X_1 - 0.5X_2$$

4. To determine the fraction of the total sum of squares attributable to regression, R^2 is calculated.

$$R^2 = \varepsilon\hat{y}_{12}^2/\varepsilon y^2 = 2/2 = 1.0$$ (All variation accounted for in this example.)

$$(\varepsilon\hat{y}_{12}^2 = b_1\, \varepsilon x_1 y + b_2 \varepsilon x_2 y$$

$$= (0.5)(2) + (-0.5)(-2)$$

$$= 1 + 1 = 2$$

5. An over-all test of the significance of the regression can now be made.

Source of variation	d.f.	s.s.	m.s.	F
Total	3⌊1	$\varepsilon y^2=2$		
Due regression	2⌊3	$\varepsilon\hat{y}_{12}^2=2$	1	$\frac{1}{0}$ = infinity
Deviations (error)	1⌊2	$\overline{0}$	0	

⌊1 n-1

⌊2 n-(no. of variables, Y and X's)

⌊3 3-1=2

CURVILINEAR REGRESSION

If a plot of regression data (Y over X) indicates a curved relation, the use of the linear regression formula is incorrect and significance is reduced over that obtainable with the appropriate curvilinear regression.

Rectification of Data. It is often possible to rectify the data into something near a linear relation by transforming the Y- and / or X- values. Once this is done, a linear regression can be calculated and tests of significance made based on the transformed data. Among the many transformations that may prove useful is the logarithmic form.

The Second Degree Polynomial. In many instances an acceptable curve can be fitted to a nonlinear relation by the second degree polynomial:

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

Calculations are the same as in a multiple regression, with X and X^2 being the two independent variables.

As an example, assuming a second degree polynomial fits the following data, we can conduct the calculations as described under multiple regression.

1.

	Seedling diam. (X_1 or X)	(X^2 or X_2)	seedling wt. (Y)
	1	1	2
	2	4	6
	3	9	12
	4	16	20
sum	10	30	40
mean	2.5	7.5	10

$\varepsilon x_1^2 = 5$ $\quad \varepsilon y^2 = 104$ $\quad \varepsilon x_1 x_2 = 25$

$\varepsilon x_2^2 = 129$ $\quad \varepsilon x_1 y = 30$ $\quad \varepsilon x_2 y = 154$

$D = 20$ $\quad b_1 = b = 1$ $\quad b_2 = c = 1$

$$Y = 10 + 1\,(X_1-2.5) + 1\,(X_2-7.5)$$

$$Y = 0 + 1X_1 + 1X_2 \quad \text{or} \quad Y = 0 + 1X + 1X^2$$

2. To test the significance of departure from linear regression.

Source of Variation	d.f.	s.s.	m.s.	F
Deviations from linear reg.	2 [1]	4 [4]		
Deviations from curved reg.	1 [2]	0 [5]	0	
Curvilinearity of reg.	1 [3]	4	4	$\frac{4}{0}$ = infi

[1] n-2
[2] n-(no. of variables, Y and X's)
[3] 2-1 = 1

$$|\underline{4}\ \varepsilon y^2 - (\varepsilon xy)^2/\varepsilon x^2 = 184 - (30)^2/5 = 4$$

$$|\underline{5}\ \varepsilon y^2 - [b\ (\varepsilon x_1\ y) + c\ (\varepsilon x_2\ y)] =$$

$$184 - [1\ (30) + 1\ (154)] = 0$$

TESTING ACCURACY

Foresters must often compare the accuracy of a new measuring technique against an accepted standard. Chi-square tests are recommended.

Suppose a new instrument estimates dbh of trees to the nearest inch from the center of sample plots and one wishes to compare these to d-tape measurements.

Observation	Estimated dbh$^{(e)}$	Actual dbh$^{(a)}$	d = e-a
1	4	5	-1
2	4	4	0
3	4	3	1
4	3	4	-1
5	5	4	1
Total	20	20	0
Mean	4	4	0

$$\chi^2 = \frac{\varepsilon(e-a)^2}{\sigma^2} \qquad \sigma^2 = \frac{E^2}{(1.96)^2}$$

E = accuracy one specifies unless a 1-in-20 chance has occurred (let us use 1 inch in our example).

$$\sigma^2 = \frac{(1)^2}{(1.96)^2} = 0.260$$

$$\chi^2 = \frac{(4-5)^2 + (4-4)^2 + \ldots(5-4)^2}{0.260}$$

$$= \frac{4}{0.260} = 15.385$$

Our chi-square exceeds the table value (11.07) with 5 (n) degrees of freedom, thus we conclude the accuracy we desired was not achieved.

One might specify that estimates be within a certain percent (P) of true values unless a 1-in-20 chance has occurred (let us use 10% in our example).

$$\chi^2 = \left(\frac{(196)^2}{P^2}\right)\left(\varepsilon\ \left(\frac{e}{a} - 1\right)^2\right)$$

$$= \left(\frac{(196)^2}{(10)^2}\right)\left(\varepsilon\ \left(\frac{4}{5} - 1\right)^2 + \left(\frac{4}{4} - 1\right)^2 + \ldots\left(\frac{5}{4} - 1\right)^2\right)$$

$$= (384.16)(0.277)$$

$$= 106.412$$

Again, with 5 degrees of freedom, we conclude the accuracy desired was not achieved.

Note: A more useful chi-square table than is usually found for these tests is on page 144 in Forest Science volume 6. For v degrees of freedom, chi-square values can be approximated by:

$$\chi^2 = .853 + v + 1.645\sqrt{2v-1}$$

A useful modification is to find error limits within which deviations between two techniques will fall unless a 1-in-20 chance has occurred:

in units:

$$E = \left[\frac{(1.96)^2\ \varepsilon(e-a)^2}{\chi^2}\right]^{\frac{1}{2}}$$

in percent:

$$E = \left[\frac{(1.96)^2\ \varepsilon\left(\frac{e}{a} - 1\right)^2}{\chi^2}\right]^{\frac{1}{2}} \quad (100)$$

where χ^2 is for n degrees of freedom.

CHI-SQUARE
(hypothetical ratio)

The chi-square technique permits one to compare sample counts of individuals that do or do not possess a certain attribute (as survival and non-survival) with an expected ratio. With this method, it can be determined whether the deviation of observed from expected is probably due or not due to chance alone.

An example of calculations is as follows:

A researcher is of the opinion that half of the white pine trees in a given area is infected with white pine blister rust. From a sample of 500 trees, he finds 200 are infected and 300 are not. Do these data disprove his theory?

	Infected	Not Infected
Expected	$F_1 = 250$	$F_2 = 250$
Observed	$f_1 = 200$	$f_2 = 300$

$$x^2 = \frac{(f_1 - F_1)^2}{F_1} + \frac{(f_2 - F_2)^2}{F_2}$$

$$x^2 = \frac{(200-250)^2}{250} + \frac{(300-250)^2}{250}$$

$$x^2 = \frac{(50)^2}{250} + \frac{(50)^2}{250} = \frac{2500}{250} + \frac{2500}{250} = 10 + 10 = 20^{**}$$

(highly significant at the 1 percent level, and beyond)

We have one degree of freedom in this example; if a tree were not infected, it must be infected. With d.f. = 1, a calculated chi-square value of 3.84 or above indicates we must reject the null hypothesis (the hypothesis that there is no real difference). Therefore, we conclude there was a highly significant difference between observed and expected ratios.

CHI-SQUARE
(test of independence)

Chi-square may be used also to test two or more criteria of classification for independence when there is no hypothetical expected ratio. For example, one checks the survival of shortleaf pine seedlings on two soil types and finds:

	Soil type A	Soil type B	Row totals
Living	a = 30	b = 10	40
Dead	c = 30	d = 30	60
Column totals	60	40	100

Is there a significant difference in the proportions of living to dead on the two soil types? To test this null hypothesis of independence (if accepted, the proportions may be samples from equal population ratios):

$$\chi^2 = \varepsilon\frac{(f-F)^2}{F} \text{ where F for any cell } = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

$$\text{F for a, for example} = \frac{(40)(60)}{100} = 24$$

$$\chi^2 = \frac{(30-24)^2}{24} + \frac{(10-16)^2}{16} + \frac{(30-36)^2}{36} + \frac{(30-24)^2}{24} = 6.25^*$$

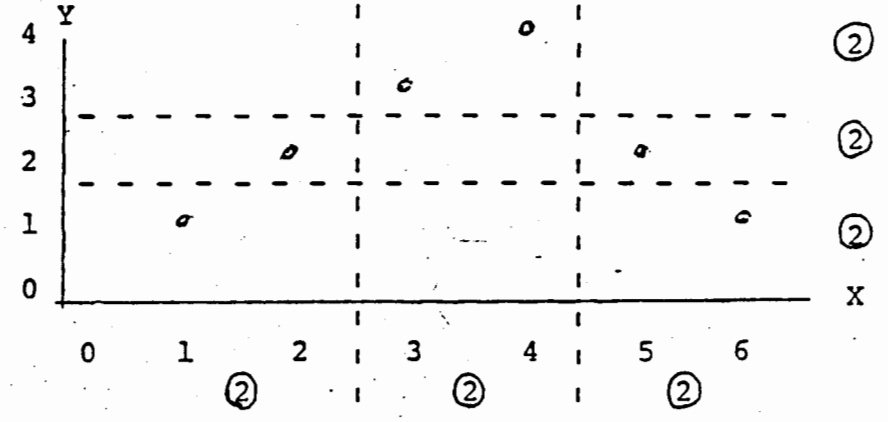
Degrees of freedom (d.f.) = (rows - 1)(columns - 1) =

(2 - 1)(2 - 1) = 1

The null hypothesis is rejected. The proportions were significantly different on the two soil types.

CHI-SQUARE TEST OF DEPENDENCE

Suppose a plot of Y over X gives the following:



The graph is divided (dashed lines) to give approximately equal totals (circled values). The table developed is:

	Column			Totals
Row	1	2	3	
1	1	0	1	2
2	1	0	1	2
3	0	2	0	2
Totals	2	2	2	6

In this example, the expected value in each cell is (2) (2) /6 = .667. The chi-square value, with (r-1)(c-1) = (3-1)(3-1) = 4 degrees of freedom is:

$$x^2 = \frac{(1-.667)^2}{.667} + \frac{(0-.667)^2}{.667} + \frac{(1-.667)^2}{.667} + \frac{(1-.667)^2}{.667} + \frac{(0-.667)^2}{.667} +$$

$$\frac{(1-.667)^2}{.667} + \frac{(0-.667)^2}{.667} + \frac{(2-.667)^2}{.667} + \frac{(0-.667)^2}{.667} = 6.00$$

As the chi-square value must exceed 9.49 to be significant at the 5 percent level, we did not detect a significant dependence of Y on X.

CHI-SQUARE MEASURE OF DEPENDENCE

There are several measures of dependence in chi-square contingency tables, where r = number of rows, c = number of columns, and q is the smaller of the two, T is the calculated chi-square value, and N is total sample size. Some are:

Cramer $$R_1 = \frac{T}{N(q - 1)}$$

(Varies between 0 and 1 but depends on r and c for interpretation.)

Pearson $$R_2 = \sqrt{\frac{T}{N + T}}$$

Mean-square $$R_3 = \frac{T}{N}$$

* Phi coefficient $$R_4 = \frac{ad - bc}{\sqrt{r_1 r_2 c_1 c_2}}$$

(Indicates direction of dependence; varies between + 1 and - 1.)

* Yule and Kendall $$R_5 = \frac{ad - bc}{ad + bc}$$

* Ives and Gibbons $$R_6 = \frac{(a + d) - (b + c)}{a + b + c + d}$$

None of the above is really best.

*For 2 x 2 tables of the form:

	c_1	c_2
r_i	a	b
r_2	c	d

MEDIAN TEST

To test whether several populations have the same median, calculate the grand median and proceed as follows:

Sample	Observations
1	2,3,4,8,
2	2,5,6
3	3,7,14,20

The grand median equals the middle ranked value if there is an odd number of observations or the average of the middle two if an even number. In this example, it is 5. Then:

	Sample			Totals
	1	2	3	
Exceeds median (O_{1i})	1	1	3	a = 5
Less or equal to median (O_{2i})	3	2	1	b = 6
Totals (n_i)	4	3	4	N = 11

The hypothesis of equal medians is rejected if T exceeds the chi-square value with degrees of freedom = number of columns - 1 = 3 - 1 = 2, where:

$$T = \frac{N^2}{ab} \sum \frac{\left(O_{1i} - \frac{n_i a}{N}\right)^2}{n_i}$$

$$= \frac{(11)^2}{(5)(6)} \left(\frac{\left(1 - \frac{(4)(5)}{11}\right)^2}{4} + \frac{\left(1 - \frac{(3)(5)}{11}\right)^2}{3} + \frac{\left(3 - \frac{(4)(5)}{11}\right)^2}{4} \right)$$

$$= 4.033\ (.167 + .044 + .349) = 2.258$$

The hypothesis is not rejected at the 5 percent level as T does not exceed 5.99. In a randomized block design, a different median is used for each block.

McNEMAR TEST OF CHANGE

When two variables for each individual can be placed in one of two classes, this test is useful. Suppose 60 people are asked to classify a clearcut as acceptable or unacceptable both before and after viewing a sign explaining this silvicultural system. Of the 60, 30 say it is unacceptable before viewing the sign and 5 of the 30 after. The others did not change their opinions. Was the change caused by the sign significant at the 5 percent level?

		After viewing the sign	
		Acceptable	Unacceptable
Before viewing the sign	Acceptable	a = 30	b = 0
	Unacceptable	c = 25	d = 5

The change is significant at the 5 percent level if T exceeds a chi-square value of 3.841, where:

$$T = \frac{(b - c)^2}{b + c}$$

$$= \frac{(0 - 25)^2}{0 + 25}$$

$$= 25$$

Therefore, the change is significant. If b + c is 20 or less, a binomial table is used instead of the chi-square approximation.

BINOMIAL TEST

If in n independent trials, each outcome can be in either "class 1" or "class 2," the probability of being in either class can be tested. For example, suppose a researcher is of the opinion that half of the white pine trees in a given area is infected with white pine blister rust (p = .5). From a sample of 500 trees:

Infected (class 1)	Not infected (class 2)
200	300

The hypothesis of equal probability is rejected at the 5 percent level if the number observed in class 1 is less than or equal to t_1 or greater than t_2, where:

$$t_1 = n\,p - 1.96\sqrt{n\,p\,(1-p)}$$

$$= (500)(.5) - 1.96\sqrt{(500)(.5)(1 - .5)}$$

$$= 228$$

$$t_2 = n\,p + 1.96\sqrt{n\,p\,(1-p)}$$

$$= (500)(.5) + 1.96\sqrt{(500)(.5)(1 - .5)}$$

$$= 272$$

As 200 is less than 228, the hypothesis of equal probability is rejected. Tables are available for various values of p and for an n of 20 or less.

WILCOXON SIGNED RANKS TEST

This procedure can be used to test whether paired data (X, Y) tend to have equal values. Test scores made by 22 forestry students are used to illustrate the procedure:

Student no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Test 1 (X)	37	66	52	74	56	58	78	68	55	29	54	49	65	70	85	48	80	68	95	32	75	83
Test 2 (Y)	10	80	60	90	60	60	70	60	40	50	30	60	50	70	60	50	30	50	100	50	70	80
Y-X	-27	14	8	16	4	2	-8	-8	-15	21	-24	11	-15	0	-25	2	-50	-18	5	18	-5	-3
Rank of |Y-X|*	20	11	8	14	4	15	8	8	12.5	17	18	10	12.5	-	19	1.5	21	15.5	5.5	15.5	5.5	3
R_i**	0	11	8	14	4	15	0	0	0	17	0	10	0	-	0	1.5	0	0	5.5	15.5	0	0

*When Y - X = 0, that pair is ignored; the average of tied ranks is used; n in this example is 21 (note sign is ignored in this ranking).

**When Y - X is negative, $R_i = 0$.

To be significant at the 5 percent level, the $\Sigma R_i = 88$ must exceed

$\left[n(n+1)/4\right] + 1.96\sqrt{n(n+1)(2n+1)/24} = \left[21(21+1)/4\right] + 1.96\sqrt{21(21+1)(2(21)+1)/24} = 172$ or be less than $\left[n(n+1)/4\right] - 1.96\sqrt{n(n+1)(2n+1)/24} =$

$\left[21(21+1)/4\right] - 1.96\sqrt{21(21+1)(2(21)+1)/24} = 59$. Therefore we conclude that test scores tended to be equal. A special table is used for an n of 20 or less.

MANN-WHITNEY TEST

This procedure can be used for two random samples to test whether X's, with sample size n, and Y's, with sample size m, tend to be equal. Test scores made in different sections of a forestry course will be used to illustrate the procedure.

Scores in section 1 (X)64,56,24,24,64,20,68,52,76,50,56,44,56,58,60,40,60,40,44,72,28
Scores in section 2 (Y)37,66,52,74,56,58,78,68,55,29,54,49,65,70,85,48,80,68,95,32,75,83

Therefore, n = 21, m = 22. Scores are now ranked, using the average of tied ranks:

X	Y	Rank	X	Y	Rank	X	Y	Rank
20		1		54	17		70	34
24		2.5		55	18	72		35
24		2.5	56		20.5		74	36
28		4	56		20.5		75	37
	29	5	56		20.5	76		38
	32	6		56	20.5		78	39
	37	7	58		23.5			
40		8.5		58	23.5		80	40
40		8.5	60		25.5			
44		10.5	60		25.5		83	41
44		10.5	64		27.5			
	48	12	64		27.5			
	49	13		65	29		85	42
50		14		66	30		95	43
	52	15.5	68		32			
52		15.5		68	32			
				68	32			

$$T = (\text{sum of ranks for X's}) - \frac{n(n+1)}{2} = 373.5 - \frac{21(21+1)}{2} = 142.5$$

To be significant at the .5 percent level, T must exceed

$$\frac{nm}{2} + 1.96\sqrt{\frac{nm(n+m+1)}{12}} = \frac{(21)(22)}{2} + 1.96\sqrt{\frac{(21)(22)(21+22+1)}{12}} = 254 \text{ or}$$

be less than $\frac{nm}{2} - 1.96\sqrt{\frac{nm(n+m+1)}{12}} = \frac{(21)(22)}{2} - 1.96\sqrt{\frac{(21)(22)(21+22+1)}{12}} = 208.$

We conclude test scores tended to be unequal. A special table is used when n and m are 20 or less.

KRUSKAL-WALLIS TEST

When random samples are selected from several (k) populations, this procedure, similar to a one-way analysis of variance, can be used to determine if at least one population tends to yield larger observations than at least one of the other populations. Suppose the height of seedlings of three species in a completely randomized design is:

Species A 15, 10, 4, 12, 11, 10
Species B 6, 11, 3, 2
Species C 1, 6, 7

Observations are ranked from the smallest to the largest with tied values receiving the average ranks:

Species A		Species B		Species C		
Obs.	Rank	Obs.	Rank	Obs.	Rank	
4	4	2	2	1	1	
10	8.5	3	3	6	5.5	
10	8.5	6	5.5	7	7	
11	10.5	11	10.5			
12	12					
15	13					
R_i	56.5		21.0		13.5	
n_i	6		4		3	N=13

The test statistic is $T = \left(\frac{12}{N(N+1)} \Sigma \frac{Ri^2}{n_i}\right) - 3(N+1) =$

$$\left[\left(\frac{12}{13(13+1)}\right)\left(\frac{(56.5)^2}{6} + \frac{(21.0)^2}{4} + \frac{(13.5)^2}{3}\right)\right] - 3(13+1) = 4.354.$$

There is a significant difference at the 5 percent level if T exceeds the chi-square approximation with k - 1 degrees of freedom (chi-square = 5.99 in our case). We did not detect a significant difference. There are procedures for multiple comparisons and exact tables for k = 3 and n_i = 5 or less.

FRIEDMAN TEST

Data from a randomized block design with k treatments and b blocks are analyzed as follows to determine if at least one treatment tends to yield larger observed values than at least one other treatment:

	Treatment		
Block	A	B	C
1	140	40	10
2	110	30	30
3	110	20	4
4	100	35	6

Rank within blocks, assigning average rank to tied values.

	Treatment		
Block	A	B	C
1	3	2	1
2	3	1.5	1.5
4	3	2	1
4	3	2	1
Ri	12.0	7.5	4.5

The test statistic is $T = \left(\frac{12}{bk\ (k + 1)} \Sigma\ Ri^2\right) - 3b(k + 1) =$

$$\left[\left(\frac{12}{(4)(3)(3 + 1)}\right)\left((12.0)^2 + (7.5)^2 + (4.5)^2\right)\right] - (3)(4)(3 + 1) = 7.125.$$

To be significant at the 5 percent level, T must exceed the chi-square approximation with k - 1 degrees of freedom (chi-square = 5.99 in our case). We detected a significant difference. Multiply comparison procedures are available.

CORRELATION OF RANKED DATA

When quantitative measurement is not feasible, rankings may be compared by a correlation technique. Suppose two foresters are asked to rank the relative tolerance of 7 species (A to G) and the following data are obtained:

Species	Ranking by Forester 1	Ranking by Forester 2	Difference d	d^2
A	1	2	-1	1
B	2	1	1	1
C	3	3	0	0
D	4	4	0	0
E	5	6	-1	1
F	6	7	-1	1
G	7	5	2	4
			$\varepsilon d = 0$	$\varepsilon d^2 = 8$

Spearman's rank correlation coefficient (r_s) is calculated:

$$r_s = 1 - \frac{6\varepsilon d^2}{n(n^2 - 1)}$$

$$r_s = 1 - \frac{6 \times 8}{7(49 - 1)} = 0.86$$

For samples of 10 or fewer pairs, r_s must exceed the following values for significance at the 5% level:

Size of sample	5% level
4 or less	none
5	1.000
6	0.886
7	0.750
8	0.714
9	0.683
10	0.648

For samples of more than 10 pairs, use a table for testing r (remember that the degrees of freedom equal 2 less than the number of pairs).

We conclude there was a significant correlation of ranking of tolerance by our two foresters.