

REC:6:49

Outline of Material Covered in
Course of Lectures on Experimental Statistics
Presented in Fall, 1948 to
AEC Group at Duke University

by Ralph E. Comstock

Mimeo. Series #10
For Limited Distribution

I. Experimentation - The experimenter is interested in the attributes of populations. He may be primarily concerned with precise estimation of attributes of a single population or with comparison of analagous attributes of two or more populations. For example, a sociologist might wish to know with maximum accuracy the mean income of negro tenant farmers in North Carolina. On the other hand, he might wish to know how the mean incomes and variation in income compare for white and negro tenant farmers.

Because of variation within populations (all negro tenant farmers do not receive equal incomes) samples drawn from a single population will vary among themselves. It is apparent, therefore, that from samples the investigator can only estimate the attributes of populations; he will never know them exactly. (This, of course, does not apply for finite populations on which complete and accurate information can be obtained. However, while the collection of information on all members of a finite population is a possibility, exactness in that information is more often than not unattainable. How will one determine the precise income of a tenant farmer?)

The science of statistics is concerned with the problems imposed by the errors involved in estimating population attributes from sample information.

II. First step in the design of a good investigation - Define carefully the population or populations about which it is desired to draw conclusions. When this is not done, the sample studied may frequently be inappropriate. For example, if one wishes to know the incidence in corn of mutations caused by irradiation of seed with gamma rays, yellow corn from a single field would constitute an inadequate sample. Moisture content, color, starch content, etc. may all be factors affecting the mutation rate. Hence, it would not be safe to draw conclusions to be applied to corn in general unless the sample used for experimentation were representative of the general corn population.

III. Parameters and statistics.

- A. Constants (which may vary from one population to another) involved in the mathematical description of a population are called parameters.
- B. Estimates of parameters based on sample information are called statistics.

Example: The "normal" distribution is described by the following equation:

$$df = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x-m)^2}{\sigma^2}} dx$$

where df is the frequency of population members in any infinitesimal range, dx , of the measured character of the population. The parameters involved are σ^2 , the variance of the population; and m , the population mean. On the other hand the mean of a sample drawn from the population is a statistic which estimates the parameter, m .

IV. Uses of statistics.

- A. Summarization of data.
- B. Estimation of parameters.
- C. Determination of the precision of such estimates and tests of "significance" of the deviation of estimates from hypothetical values.

V. The mean, variance, and standard deviation of populations.

Let quantitative measures on the individuals of a population be symbolized as

$$X_1, X_2, X_3, \dots, X_N.$$

For example, the population might be white men 21 years of age or older in the U.S. and the measure, height in inches. X_1 would then be the height of one member of this population, X_2 the height of another, etc.

The population mean =

$$m = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{S(X)}{N}$$

The population variance = σ^2

$$= \frac{(X_1 - m)^2 + (X_2 - m)^2 + \dots + (X_N - m)^2}{N} = \frac{S(X - m)^2}{N}$$

The population standard deviation is the square root of the variance = σ

$$= \sqrt{\frac{S(X-m)^2}{N}}$$

Note: The population variance and standard deviation will also be referred to as the variance and standard deviation of the individuals of the population.

VI. Sample means and variances.

A. The sample mean = $\bar{x} = \frac{S(X)}{n}$,

where X_1, X_2, \dots, X_n are quantitative measures on the individuals of a sample. The sample mean is an unbiased statistic, i.e. on the average sample means equal the population mean, and as sample size is increased the sample mean may be expected to deviate less and less from the population mean.

B. The sample variance = $s^2 = \frac{S(X - \bar{X})^2}{N-1}$

s^2 is also an unbiased statistic, estimating the population variance. It would be biased (a little too small on the average) if the denominator were N instead of $N-1$.

VII. Coding - Two sorts are common. One involves subtraction of a constant value from each of a series of numbers. The other consists of dividing (or multiplying) each of a series of numbers by a constant. The purpose is usually to obtain numbers which render arithmetic computations less laborious.

A. Effect of subtraction coding on the mean and variance.

Let c be a constant subtracted from each value of a variable, X . Then

$$x_1 = X_1 - c$$

$$x_2 = X_2 - c$$

etc.

Where x_1, x_2, \dots, x_n are the coded values of X_1, X_2, \dots, X_n .

$$\begin{aligned} \bar{x} &= \frac{(X_1 - c) + (X_2 - c) + \dots + (X_n - c)}{n} = \\ &= \frac{S(X) - nc}{n} = \bar{X} - c \end{aligned}$$

i.e. the mean of $x = X - c$ is the mean of the original values minus the coding constant.

$$\begin{aligned} s_x^2 &= \frac{S[(X - c) - (\bar{x} - c)]^2}{n - 1} = \frac{S[X - c - \bar{X} + c]^2}{n - 1} \\ &= \frac{S[X - \bar{X}]^2}{n - 1} = s_X^2 \end{aligned}$$

i.e. the variance of the coded values is the same as the variance of the original values.

B. Effect of division (or multiplication) coding. Let c be a constant by which each value of X is divided. Stated another way we are letting each X be multiplied by $1/c$. Then,

$$x_1 = X_1/c$$

$$x_2 = X_2/c$$

etc.

$$\bar{x} = \frac{X_1/c + X_2/c + \dots + X_n/c}{n} =$$

$$\frac{X_1}{cn} + \frac{X_2}{cn} + \dots + \frac{X_n}{cn} = \frac{S(X)}{cn} = \frac{\bar{X}}{c}$$

i.e. the mean of the coded values is equal to the mean of the original value divided by the constant used as divisor in coding (or multiplied

by the constant used as multiplier in coding).

$$s_x^2 = \frac{S(X/c - \bar{X}/c)^2}{n-1} = \frac{S(X - \bar{X})^2}{c^2(n-1)} = \frac{s_X^2}{c^2}$$

i.e. the variance of the coded values is equal to the variance of the original values divided by the square of the constant used as divisor in coding (or multiplied by the square of the constant used as multiplier in coding).

C. Special case of subtraction coding in which $c = \bar{X}$. Then

$$x_1 = X_1 - \bar{X}$$

$$x_2 = X_2 - \bar{X}$$

etc.

$$\bar{x} = \bar{X} - c = \bar{X} - \bar{X} = 0$$

$$s_x^2 = s_X^2$$

From this point on small case letters will be used only as coded values obtained by subtraction coding with c equal to the mean of the uncoded variables.

VIII. Working formula for the variance

$$S(X - \bar{X})^2 = S(X^2 - 2X\bar{X} + \bar{X}^2) = SX^2 - 2\bar{X}S(X) + n\bar{X}^2.$$

Now since $\bar{X} = \frac{S(X)}{n}$,

$$n\bar{X}^2 = n \frac{S(X)}{n} \cdot \frac{S(X)}{n} = \frac{S(X)}{n} \cdot S(X) = \bar{X}S(X).$$

Therefore,

$$S(X - \bar{X})^2 = SX^2 - \bar{X}S(X) = SX^2 - \frac{[SX]^2}{n}.$$

(SX^2 is commonly called the uncorrected sum of squares; and $[SX]^2/n$, the correction factor.)

$$s_x^2 \text{ is then } \frac{S(X - \bar{X})^2}{n-1} = \frac{S(X^2) - [SX]^2/n}{n-1}.$$

Since the variance of values obtained by subtraction coding is the same as that of original values, we can write

$$s_X^2 = s_x^2 = \frac{S(x^2) - [Sx]^2/n}{n-1}.$$

However, since $\bar{x} = 0$, $S(x)$ must equal zero, so

$$s_X^2 = s_x^2 = \frac{S(x^2)}{n-1}.$$

Numerical Example of Computation of s_X^2 and s_X .

The figures in the X column are the number of lines in each of six paragraphs chosen at random from "Statistical Methods" by R.A. Fisher.

<u>X</u>	<u>X²</u>	$\frac{(SX)^2}{N} = \frac{(105)^2}{6} = 1837.5$
14	196	
17	289	$s_X^2 = \frac{SX^2 - (SX)^2/N}{N-1}$
23	529	
7	49	$= \frac{2319 - 1837.5}{5} = 96.3$
10	100	
<u>34</u>	<u>1156</u>	$s_X = \sqrt{96.3} = 9.81$
SX = 105	SX ² = 2319	

From our sample we estimate the average paragraph length of the book as 17.5 lines and the standard deviation of paragraph length as 9.81 lines.

Degrees of Freedom associated with an estimated variance or standard deviation are N-1 in cases like the above. Thus in this example the variance estimate is based on 5 degrees of freedom.

IX. Population variances of sums and differences. Consider two populations:

A_1, A_2, \dots, A_N with mean \bar{A} , and B_1, B_2, \dots, B_N with mean \bar{B} , and let the sum of an A and a B drawn at random from their respective populations be designated as Z. Then

$$Z_1 = A_1 + B_1$$

$$Z_2 = A_2 + B_2$$

etc.

Where the subscript numbers are applied at random to the members of the A and B populations.

$$\bar{Z} = \frac{A_1 + B_1 + A_2 + B_2 + \dots + A_N + B_N}{N} = \bar{A} + \bar{B}$$

$$z_1 = Z_1 - \bar{Z} = A_1 + B_1 - \bar{A} - \bar{B} = a_1 + b_1$$

$$z_2 = Z_2 - \bar{Z} = A_2 + B_2 - \bar{A} - \bar{B} = a_2 + b_2$$

etc.

(Remember note under VII, C on significance of small case letters.)

$$\sigma_Z^2 = \sigma_z^2 = \frac{S(a+b)^2}{N} = \frac{Sa^2}{N} + \frac{2Sab}{N} + \frac{Sb^2}{N}$$

However, since the A's and B's were chosen at random in the formation of the Z's, Sab will approach zero as N becomes large, i.e. in infinite populations, and we can write

$$\sigma_Z^2 = \sigma_z^2 = \frac{Sa^2}{N} + \frac{Sb^2}{N} = \sigma_a^2 + \sigma_b^2 = \sigma_A^2 + \sigma_B^2.$$

Now suppose Z were formed by summing the values for individuals drawn at random from three populations so that

$$Z_1 = A_1 + B_1 + C_1.$$

By procedure analogous to that above it can then be shown that

$$\sigma_Z^2 = \sigma_A^2 + \sigma_B^2 + \sigma_C^2.$$

It will readily be perceived by extending the analogy that the variance of the sum of any number of randomly selected variates will equal the sum of the variances of those variates.

Now consider the differences between variates. Let

$$Z_1 = A_1 - B_1, Z_2 = A_2 - B_2, \text{ etc.}$$

Then

$$z_1 = a_1 - b_1, \text{ etc.}$$

$$\sigma_Z^2 = \sigma_z^2 = \frac{S(a-b)^2}{N} = \frac{Sa^2}{N} - \frac{2Sab}{N} + \frac{Sb^2}{N}$$

Since Sab approaches zero for large N, we have as before

$$\sigma_Z^2 = \sigma_A^2 + \sigma_B^2$$

Extending this it becomes clear that if

$$Z = \pm A \pm B \pm C \pm D + \dots$$

$$\sigma_Z^2 = \sigma_A^2 + \sigma_B^2 + \sigma_C^2 + \sigma_D^2 + \dots$$

i.e. that when variates are combined either by addition or subtraction (or addition of some and subtraction of others) the variance of the resultant is the sum of the variances of the variates combined. Remember that the above holds only when the things summed are chosen randomly from their respective populations.

X. Variance of the mean of n values from the same population.

$$\bar{X} = \frac{X_1}{n} + \frac{X_2}{n} + \dots + \frac{X_n}{n}$$

Think of a series of such means. Since the X_1 's, the X_2 's, etc. are all from the same population, we can write

$$\sigma_{X_1}^2 = \sigma_{X_2}^2 = \dots = \sigma_{X_n}^2 = \sigma_X^2$$

Then remembering (from VII) the relation between the variance of a variable and the variance of that variable divided by a constant (n in this case) it

is clear from what was learned above about the variance of sums that

$$\sigma_{\bar{X}}^2 = \frac{\sigma_{X_1}^2}{n^2} + \frac{\sigma_{X_2}^2}{n^2} + \dots + \frac{\sigma_{X_n}^2}{n^2} = \frac{n \sigma_X^2}{n^2} = \frac{\sigma_X^2}{n} .$$

Thus the estimated variance of our estimate of the mean paragraph length in Fisher's book (Section VIII) is $96.3/6 = 16.5$ and the estimated standard deviation of means of 6 paragraphs is $\sqrt{16.5} = 4.06$ lines.

XI. The probability of specified deviations from the population mean.

From the expression for the "normal" frequency distribution (see III) the proportion of a "normally distributed" population that will deviate in magnitude by more than any specified amount from the population mean, can be determined. (This is accomplished by integration between limits.) Table 8.6, p. 180 of the text gives the proportions of a normal population that have magnitudes equal or greater than the mean but not greater than the mean plus various multiples of the standard deviation (symbolized by t in the table). Since the normal distribution is symmetrical about the mean the tabled proportions can also be taken as those equal or less than the mean but not less than the mean minus the indicated multiples of the standard deviation.

For example for $t = 2.01$, the tabled value is .4778 so it can be stated that $2 \times .4778 = .9556$ of the individuals of a normally distributed population will have magnitudes in the range, $m \pm 2.01\sigma$.

As another example assume an observed difference of 16 units between two sample means and that the standard deviation of such differences is known to be 10 units. Then if the true difference, i.e. the difference between the means of the two populations from which the samples were drawn, were zero; the observed difference would be 1.6 standard deviations larger than the true difference which can be considered the mean of a population of which the observed

difference is a member. From table 8.6 we learn that $1 - (2 \times .4452) = .1096$ or 10.96% of the members of a normal population will deviate in magnitude by 1.6 or more standard deviations. This could be stated as follows: The probability of observing a difference as large or larger than 16 units in any single experiment comparable to the one in question is 10.96%, if the true difference is zero.

Ordinarily the true standard deviation is not known but an estimate of it based on sample information is available. This may either be too large or too small. The net result is that if this estimate is used as the real standard deviation probabilities obtained as described above will be somewhat too small. The ratio between an estimate (of the difference between a sample value and a hypothetical population value) and its estimated standard deviation has been designated as t (frequently referred to as Student's t). For example

$$t = d/s_d \quad \text{where } \underline{d} \text{ is the difference between two sample means.}$$

Values of t that will be exceeded in various proportions of a large number of cases have been computed from the normal frequency distribution and the distribution of s and have been tabulated (see table 3.8, p. 65 of the text). For example if the estimate of a difference is 15 and its estimated standard deviation is 6, $t = 2.5$. Now suppose the estimate of the standard deviation was based on 10 degrees of freedom. Then from the table we note that the probability of this t is less than .05 if the true difference is zero.

XII. Numerical examples of the use of the t-test.

A. Our estimate of average paragraph length of Fisher's book was 16.5 and its estimated standard deviation was 4.06. The latter figure is also an estimate of the standard deviation of differences between means of samples of size 6 and the true mean paragraph length for the book since the true

mean is a fixed value (these differences would be of the nature of the values $x = X - a$ of VII, A whose variance is equal to σ_X^2). Hence

$$t = \frac{\bar{X} - m}{s_X} \quad \text{and} \quad ts_X = \bar{X} - m$$

From table 3.8 we note that with 5 d.f. (degrees of freedom) a t as large or larger than 2.57 will occur 5% of the time as a result of random variation. Thus if we multiply s_X by 2.57 we have a value of $\bar{X} - m$ which in the average of a large number of cases would be exceeded in the case of 5% of such samples. This knowledge enables us to set what are called Fiducial Limit about our estimate of the population mean. In this case our Fiducial Limits are:

$$16.5 - (2.57 \times 4.06) = 6.1$$

$$\text{and } 16.5 + (2.57 \times 4.06) = 26.9 .$$

In only 1 case of 20, i.e. 5% of the cases, will the population mean be outside the Fiducial Limits. Thus we may conclude that unless a one in twenty chance has come off the mean paragraph length is not less than 6.1 nor more than 26.9 lines. It is really not a very good estimate.

- B. In the following data X_1 is the number of lines in 10 paragraphs chosen at random in Snedecor's book and X_2 is number in 10 random paragraphs of Fisher's book.

$\underline{x_1}$	$\underline{x_2}$	$\bar{x}_1 = 11.3$	$\bar{x}_2 = 17.3$
5	14	$s_1^2 = 48.23$	$s_2^2 = 73.79$
9	17	$\bar{x}_1 - \bar{x}_2 = -6.0$	
8	23		
10	7		
7	10	$s^2(\bar{x}_1 - \bar{x}_2) = \frac{48.23}{10} + \frac{73.79}{10} = 12.202$	
9	34		
29	24	$s(\bar{x}_1 - \bar{x}_2) = 3.49$	
9	13		
10	8	$t = \frac{-6.0}{3.49} = 1.72$	
$\underline{17}$	$\underline{23}$		
113	173		

Unless there is some a priori reason for suspecting a difference between the variances of the two populations the routine procedure is to use the sum of the degrees of freedom for the two samples as the d.f. for t. In this case there would be 18. The probability of $t \geq 1.72$ with 18 d.f. is slightly over 10% assuming the population value of $\bar{x}_1 - \bar{x}_2 = 0$. The data suggest there may be a difference between the books in paragraph length but are hardly conclusive since a t as large or larger than 1.72 would occur one time in 10 as a result of sampling error.

- C. Paired data - Example 2.5, p. 35 of the text furnishes an example. If the yield of oats varies from year to year as a consequence of weather variation it is to be anticipated that a paired comparison (concurrent observation of the yield of both varieties in a series of years) will furnish a better estimate of differences in yield of the two varieties than would one which allowed differences between years to effect the estimate of the varietal difference. The t-test is then based on differences between the paired

observations. The data of example 2.5 are as follows:

Year	Yield in Bushels		Difference	(Difference) ²
	Variety A	Variety B		
1	34	30	4	16
2	30	15	15	225
3	41	33	8	64
4	25	25	0	0
5	<u>45</u>	<u>25</u>	<u>20</u>	<u>400</u>
			47	705

$$\bar{d} = \frac{47}{5} = 9.4$$

$$s_d^2 = \frac{705 - (47)^2/5}{4} = 65.8$$

$$s_d^2 = 65.8/5 = 13.16$$

$$s_d = 3.63$$

$$t = 9.4/3.63 = 2.59$$

A t of this size will occur only about 6% of the time if there is no difference between the means of the populations sampled. Thus we would be on fairly safe grounds if we conclude that variety A were truly a better yielder than variety B. For another example of this sort of data see Table 2.2, p.44.

It should be noted that the tabulated probabilities of t are derived from the normal distribution and hence are strictly applicable only to data from normally distributed populations. However, deviation from the normal distribution must be rather great before the t-test becomes greatly in error.

XIII. Analysis of variance. In its simplest form the analysis of variance involves two variance estimates made in such a manner that both will be estimates of the same population effects if an appropriate null hypothesis is true. Evidence for non-validity of the null hypothesis will be expressed by a difference

between the two estimates that is of a magnitude to be expected only infrequently as a consequence of sampling error.

Example 1

The following data were obtained in an experiment comparing three diets. The experimental animals used were rats and the effects of the diets were measured in terms of increase in body weight. Ten rats were assigned at random to each diet and weight increase measured over the same period of time for all 30 rats.

Weight gains of individuals rats

<u>Diet A</u>	<u>Diet B</u>	<u>Diet C</u>	
73	98	94	
102	74	79	
118	56	96	
104	111	98	
81	95	102	
107	88	102	
100	82	108	
87	77	91	
117	86	120	
<u>111</u>	<u>92</u>	<u>105</u>	
1000	859	995	Grand total = 2854

The null hypothesis (N.H.) in this case is that the treatments do not differ in their effect on rat growth. If the N.H. is true the variation among treatment sums and the variation among individuals accorded the same treatment can be used for two independent estimates of the population variance.

The population variance of sums of 10 random individuals is $10\sigma^2$ (see IX). Thus if the variance of the three sums is divided by 10 it will be an estimate of σ^2 as will the average within treatment variance.

The computations are as follows:

A correction factor (C.F.) is computed as follows:

$$C.F. = \frac{[S X]^2}{N} = \frac{(2854)^2}{30} = 271,511.$$

$$\text{Total Sum of Squares (S.S.)} = \sum X^2 - C.F.$$

$$= (73)^2 + (102)^2 + \dots + (105)^2 - 271,511 = 6,445.0$$

$$\text{Treatment S.S.} = \frac{\sum [S_t]^2}{n_t} - C.F.$$

$$= \frac{(1000)^2 + (859)^2 + (995)^2}{10} - 271,511 = 1,279.6$$

Within treatment or error S.S.

$$= \text{Total S.S.} - \text{Treatment S.S.} = 6,445.0 - 1,279.6 = 5,165.4 .$$

These values together with associated degrees of freedom are ordinarily tabulated as follows:

<u>Source of variation</u>	<u>d.f.</u>	<u>S.S.</u>	<u>m.s.</u>	<u>F</u>
Treatments	2	1,279.6	639.8	3.34
<u>Within treatments</u>	<u>27</u>	<u>5,165.4</u>	<u>191.3</u>	
Total	29	6,445.0		

The two variance estimates are obtained by dividing the two S.S.'s by their respective degrees of freedom. These values are termed mean squares and are recorded in the m.s column. In this case their relative sizes suggest that there was a treatment difference, i.e. that the N.H. is not true.

An exact test of the discrepancy between the two mean squares is made using F, the ratio of the largest to the smallest. In this case $F = 639.8/191.3 = 3.34$. Table 10.7, pp. 222-225 in the text gives values of F that will be equalled or exceeded 5% and 1% of the time if the N.H. is true. The table is entered in accordance with the numbers of degrees of freedom associated with the two mean squares. When the larger m.s. has 2 d.f. and the smaller 27 d.f. we note from the table that the 5% value of F is 3.35. Thus the result observed in the case of this experiment is one that would

occur only very very slightly more than 5% of the time if there are no real effects of the treatments. Since the result observed was one to be expected so infrequently in the absence of real treatment effects it strongly suggests that there are real differences between the treatments for support of growth in rats.

Returning to the data we note that mean growth on diets A and C was 100 grams and 99.5 grams, respectively, as compared to 85.9 grams for diet B. It would appear the suggested treatment difference is between B on the one hand and A and C on the other. The nature of the diets makes this inference logical on nutritional grounds; diet B was high in cereals while A contained beef and C pork.

Example 2

This is data from an experiment designed to learn whether thinning of peach trees results in production of larger fruit. Since trees may differ in fruit size it was decided to compare a thinned and unthinned branch on each of several trees so that differences between trees would not be a source of error in measuring the effect of the treatment (thinning). The data are as follows. Each figure is the average weight in grams of 50 peaches:

<u>True</u>	<u>Thinned branch</u>	<u>Unthinned branch</u>	<u>Sum</u>	<u>Difference</u>
1	104	83	187	21
2	80	58	138	22
3	89	62	151	27
4	88	80	168	8
5	90	70	160	20
6	85	62	147	23
7	94	88	182	6
8	75	76	151	-1
9	87	64	151	23
10	100	86	186	14
11	93	85	178	8
12	102	90	192	12
13	91	80	171	11
14	89	65	154	24
15	87	70	157	17
16	<u>86</u>	<u>81</u>	<u>167</u>	<u>5</u>
Sum	1440	1200	2640	240
Mean	90	75		15

There was an average difference of 15 grams in favor of the thinned branches. Moreover, there was a degree of consistency in the results. In 15 of the 16 trees, the thinned branches yielded the larger fruit.

This experiment was designed in such a way that the total sum of squares is divisible into three portions; one due to variation between trees, one to the effect of treatment, and a remainder to unassignable causes (experimental error). Note that the sum for each tree involves data from one thinned and one unthinned branch; and that each tree contributed equally to each treatment sum. As a consequence, tree differences do not contribute to the difference between treatment sums, nor does the difference (if any) between treatments contribute to differences between sums for trees. In technical language we say that the effects of trees and treatments are unconfounded. Lack of confounding is obviously a desirable feature of an experiment.

The treatment effect can obviously be tested using t computed from the differences; this is a paired experiment.

$$s_d^2 = \frac{(21)^2 + (22)^2 + \dots + (5)^2 - (240)^2/16}{15} = 68.53$$

$$s_d = \sqrt{68.53/16} = 2.07$$

$$t = 15/2.07 = 7.25$$

From the t-table (p.65) we learn that when there are 15 d.f., as in this case, a t as large or larger than 2.95 will occur only 1% of the time as a consequence of sampling error. Thus a t of 7.25 can only be interpreted as evidence for a difference between the treatments.

While the treatment difference can be tested using t , the test can be made using analysis of variance and this would be preferred if one also wished to test the reality of differences between trees. The computations are as follows:

$$C.F. = (2640)^2/32 = 217,800$$

$$\text{Total S.S.} = (104)^2 + (80)^2 + \dots + (81)^2 - C.F. = 4320$$

$$\text{Tree S.S.} = \frac{(187)^2 + (138)^2 + \dots + (167)^2}{2} - C.F. = 2006$$

$$\text{Treatment S.S.} = \frac{(1440)^2 + (1200)^2}{16} - C.F. = 1800$$

$$\text{Error (tree x treatment) S.S.} = \text{Total S.S.} - \text{Tree S.S.}$$

$$- \text{Treatment S.S.} = 514.$$

The analysis of variance table is as follows:

<u>Source of variation</u>	<u>d.f.</u>	<u>S.S.</u>	<u>m.s.</u>	<u>F.</u>
Trees	15	2006	133.73	3.89
Treatments	1	1800	1800.00	52.52
Error	15	514	34.27	
Total	31	4320		

From the F-table (p. 222) we find that when there are 1 and 15 d.f. an F value as large or larger than 8.68 will occur by chance only one time in 100. The observed F of 52.52 therefore indicates a real treatment difference as in the case of the t-test. With 15 and 15 d.f. an F as large or larger than 3.52 has a random probability of only one in 100. Thus the observed F for the tree m.s. indicates real tree differences. This result was not available from the t-test.

Note that when there is 1 d.f. for treatments $F = t^2$ ($52.52 = 7.25 \times 7.25$). When only two things are being compared the t and F-tests are merely different forms of the same test and always give identical results.

The computation of the analysis of variance when there are more than two treatments are analogous to those in the above example. The divisor used in computing a sum of squares is always the number of observations summed in the figures being used in getting the S.S. (for example, note that in the case of the treatment S.S. in this example the divisor was 16, the number of trees on which observations were made for each treatment).

The analysis of variance is applicable to data involving more sources of variation than in the above examples. Section 11.14, p. 304, of the text contains three examples, giving computational details for the first two.

It should be noted that the F table like the t table is appropriate to normally distributed material. Again, however, moderate deviations from normality is not likely to lead to serious error.

XIV. Linear regression.

There is frequently reason to believe or to suspect that one variable is dependent on another. For example, the number of tooth cavities may be related to milk consumption, money spent for groceries by a family should be related

to the family income, weight of children to their height, etc. In some instances there will be reason to believe that if such a relation exists it will be of a linear form, that is, that the extent of the response in the dependent variable will be the same at all levels of the independent variable. In some instances the relation will be linear (or essentially so) in a portion of the range of variation but non-linear outside that portion of the range. In the second example listed above income is the independent variable, money spent for groceries the dependent. In this case the relation is probably close to linear until medium to high incomes are reached after which money spent for groceries per unit income probably decreases as income gets higher.

When the relation between two variables is linear in form it can be expressed symbolically as follows:

$$Y - \bar{Y} = \beta(X - \bar{X}) + e$$

where Y is the value for the dependent variable,

\bar{Y} the mean of all Y 's,

X the value of the independent variable,

\bar{X} the mean of all X 's,

e the portion of $Y - \bar{Y}$ not associated with variation in X , and

β is the change in Y per unit change in X , the linear regression coefficient.

We will be concerned with three problems.

1. How to estimate β .
2. How to determine the probability of an estimate of β assuming $\beta = 0$.
3. The precision of estimates of β .

The assumption throughout will be that the relationship between the variables is linear or essentially linear. This will of course not be the case in many

problems. There are technics for testing the hypothesis of linearity and for dealing with non-linear cases but time prevents us from considering them.

There are many cases where the assumption of linearity can be justified from prior knowledge.

Example:

The following data are for the State of Ohio. X is rainfall in excess of 8 inches and Y is average corn yield in bushels. Our questions are (1) is average corn yield dependent on rainfall, and (2) if so, how much is yield increased by an inch of rain?

<u>Year</u>	<u>X</u>	<u>Y</u>
1883	19	19
84	11	18
85	13	14
86	0	12
87	6	13
88	3	11
89	15	17
90	4	10
91	5	12
92	7	14
93	6	16
94	1	12
95	1	11
96	26	20
97	7	16
98	8	17
99	6	14
1900	9	14
01	9	16
02	12	17
03	12	16
04	15	18
05	11	17
06	10	16
07	<u>9</u>	<u>15</u>
	225	375

The quantity, $\frac{S(X - \bar{X})(Y - \bar{Y})}{S(X - \bar{X})^2}$, is an unbiased estimate of β , The statistic

$$b = \frac{S(X - \bar{X})(Y - \bar{Y})}{S(X - \bar{X})^2}$$

in addition to being unbiased, is also efficient, i.e. the data can yield no estimate of β that is more precise than b .

The quantity, $S(X - \bar{X})(Y - \bar{Y})$ is called the sum of products for X and Y. It can be converted to a computation form analogous to that for $S(X - \bar{X})^2$.

The working form is

$$S(XY) - \frac{S(X) S(Y)}{N}$$

For our example

$$S(X - \bar{X})^2 = (19)^2 + (11)^2 + \dots (9)^2 - \frac{(225)^2}{25} = 826$$

$$S(X - \bar{X})(Y - \bar{Y}) = (19 \times 19) + (11 \times 18) + \dots (9 \times 15) - \frac{(225)(375)}{25} = 319$$

$$S(Y - \bar{Y})^2 = (19)^2 + (18)^2 + \dots (15)^2 - \frac{(375)^2}{25} = 172$$

$$b = \frac{319}{826} = .39$$

According to this data then, corn production per acre goes up .39 bushel for each increase of one inch in rainfall. We want to know whether $b = .39$ is significantly different from zero and more specifically just how precise it is as an estimate of β , the true effect of rainfall. The test of significance can be made using either t or the analysis of variance. If t is to be used it will be computed as

$$t = b/s_b$$

and s_b must first be computed.

$$s_b = \sqrt{\frac{S(Y - \bar{Y})^2 - \frac{[S(X - \bar{X})(Y - \bar{Y})]^2}{S(X - \bar{X})^2}}{(N - 2) S(X - \bar{X})^2}} = \sqrt{\frac{172 - \frac{(319)^2}{826}}{23(826)}} = .051$$

$$t = .39/.05 = 7.62$$

The degrees of freedom are $(N - 2)$. From table 3.8, p. 65 we find that a $t \geq 2.807$ would occur only 1 time in 100 by chance. We conclude that β is not zero, that yield of corn in Ohio does vary with rainfall. Using the technic of XII, A we can state further that unless a one in twenty chance has come off the value of β is between

$$.39 + (.051 \times 2.069) = .495$$

$$\text{and } .39 - (.051 \times 2.069) = .285$$

The analysis of variance procedure for testing whether b deviates significantly from zero is as follows: The sum of squares for Y is divided into two parts, that due to regression on X, and that due to deviation from regression.

$$\text{S.S. due to regression} = \frac{[\sum(X - \bar{X})(Y - \bar{Y})]^2}{\sum(X - \bar{X})^2} = 123$$

$$\text{S.S. due to deviation from regression} = \sum(Y - \bar{Y})^2 - \frac{[\sum(X - \bar{X})(Y - \bar{Y})]^2}{\sum(X - \bar{X})^2} = 49$$

The analysis of variance table is as follows:

<u>Source of variation</u>	<u>d.f.</u>	<u>S.S.</u>	<u>m.s.</u>	<u>F.</u>
Regression	1	123	123	58.11
Deviation from regression	23	49	2.12	
Total	24	172		

We note from the F-table that 58.11 is far above F for $P = .01$ when d.f. are 1 and 23. Note that as before the same result is obtained as with the t-test. Again the two are the same test in different form and if sufficient decimals had been carried throughout we would have had $F = t^2$.

Deviation of the formula for b .

b is the least squares estimate of β , i.e. its computation is such that the sum of squares of deviations of the observed Y values from the line

$$Y_p = \bar{Y} + b(X - \bar{X})$$

is a minimum. Y_p is the predicted value of Y for the value of X inserted in the right hand side of the equation. The formula for the deviation of a specific Y from the regression line is

$$Y - Y_p = Y - \bar{Y} - b(X - \bar{X})$$

and the sum of squares of these deviations is

$$S(Y - Y_p)^2 = S[(Y - \bar{Y}) - b(X - \bar{X})]^2.$$

Taking the derivative of the right hand side with respect to b we have

$$\frac{d}{db} S(Y - Y_p)^2 = -2 S(X - \bar{X})(Y - \bar{Y}) + 2b S(X - \bar{X})^2.$$

Setting the derivative equal to zero and solving for b we have

$$b = \frac{S(X - \bar{X})(Y - \bar{Y})}{S(X - \bar{X})^2}$$

XV. The product moment correlation coefficient.

It is commonly referred to as simply the correlation coefficient. It is by definition

$$r = \frac{S(X - \bar{X})(Y - \bar{Y})}{\sqrt{S(X - \bar{X})^2 S(Y - \bar{Y})^2}}$$

In the example of the preceding section

$$r = \frac{319}{\sqrt{(826)(172)}} = .85$$

A correlation coefficient will always have the same probability of chance occurrence as the linear regression coefficient associated with it. Hence

there is no need to specify a significance test for the correlation coefficient.

The important attributes of the correlation coefficient are as follows:

1. It may vary from -1 to +1. A value of 1.0, either plus or minus, signifies perfect correlation. If the sign is plus the two variables vary in the same direction. If the sign is minus one is increasing where the other is decreasing.
2. The sum of squares in Y due to regression on X is

$$\frac{[\sum(X - \bar{X})(Y - \bar{Y})]^2}{\sum(X - \bar{X})^2} = r^2 \sum(Y - \bar{Y})^2$$

The sum of squares for deviation of Y from regression on X is

$$(1 - r^2) \sum(Y - \bar{Y})^2 .$$

The correlation coefficient is falling out of use since it furnishes no new information once b and the two portions of the sum of squares of Y are known. Since one almost always wants to know those quantities r becomes superfluous. Furthermore, in a great deal of experimental data the values of the independent variable are selected. For example, if one were experimenting in the effect of X-rays on mutation rate of a specified gene he would in all probability use graded doses of X-rays. The correlation coefficient changes with the distribution of X-values involved, whereas the regression coefficient does not so long as one works within the range where regression is linear.

XVI. The binomial distribution.

The t and F tests are precise when the variable being studied follows the normal distribution. While many variables which are measured on a continuous scale are normal or reasonably close to normal in distribution, discontinuous ones such as result from enumeration are frequently not. A distribution frequently involved in such data is the binomial.

Suppose that in a very large population of seed corn the proportion capable of germinating when planted is p . Then if four, taken at random, are planted in each hill we can compute the probability of having 0, 1, 2, 3, and 4 capable of germination in the same hill. It should be noted that since five results are the only ones that can occur the sum of their probabilities must be one. The probabilities of the five possibilities are the terms of the expansion of $(p + q)^4$ where p is the proportion capable of germination and $q = 1 - p$. Since $p + q = 1$, $(p + q)^4 = 1.0$, i.e. the sum of the five probabilities is one as it should be. The expansion is

$$p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4 .$$

The exponent of p serves to identify the result for which each term is the probability. Thus p^4 is the probability of four seeds taken at random all being capable of germination, $4p^3q$ is the probability of 3 capable of germination and one not, etc. If, for example $p = .8$, the theoretical distribution is as follows:

Seed capable of germination out of four	Frequency
4	$p^4 = (.8)^4 = .4096$
3	$4p^3q = .4096$
2	$6p^2q^2 = .1536$
1	$4pq^3 = .0256$
0	$q^4 = .0016$
Total	$(p + q)^4 = 1.0000$

Two points should be noted about the example.

1. The probability of a randomly drawn seed being one capable of germination is always the same.
2. Sets on which counts are made are of constant size. Uniform probability of the event being enumerated and enumeration in sets of

constant size are the conditions leading to the binomial distribution.

It will be noted further that the binomial distribution is completely specified by two quantities, p and n . In practical problems n is usually known, leaving only p to be estimated.

The population mean is np and the variance npq . These can be verified for the example given above.

$$\begin{aligned}\bar{X} &= 4(.4096) + 3(.4096) + 2(.1536) + 1(.0256) = 3.2 = np \\ \sigma^2 &= 16(.4096) + 9(.4096) + 4(.1536) + 1(.0256) - (3.2)^2 \\ &= .64 = npq.\end{aligned}$$

When n is moderately large and p is between .2 and .8 one will not fall into serious error by treating a binomial distribution as though it had the normal form.

Example.

Suppose one wished to test the resistance of two varieties of corn to smut and had inoculated 100 plants of each variety and found 30 susceptible plants in one variety, 50 in the other. Under the null hypothesis, i.e. that the varieties did not differ in susceptibility, the fraction susceptible (p) is estimated as $\frac{30 + 50}{200} = .4$. Then the variance of the number susceptible out of 100 is $100(.4)(.6) = 24$, the variance of the difference between two counts is $2 \times 24 = 48$, and the standard error of the difference between two counts is 6.9. Since the observed difference 20 is almost 3 times as large as its standard error it is unlikely that it is a random deviate from a true difference of zero. We would conclude therefore that under the conditions of the experiment that the varieties actually differ in susceptibility.

There are also many instances in which knowledge of the binomial distribution will be of value when the assumption of approximate normality is not involved.

Example:

Suppose one wishes to know which of a group of phenotypically normal plants are carrying a certain recessive mutant gene that has a visible effect when homozygous. The obvious way to test them is to look for homozygotes among their offspring produced by self-fertilization. How many offspring should be observed per plant to insure that one or more homozygotes will be observed among the progeny of any plant that is actually carrying the recessive gene? Assuming equal viability of the normal and mutant types they will be produced in the ratio 3:1 by any heterozygous plant. If p is the proportion of mutant types, it is obviously equal to $\frac{1}{4}$. Then if n progeny are observed per plant, the expected frequency or probability of observing no homozygous recessives is $(1 - p)^n = (3/4)^n$ if the parent plant carries the gene. Thus if 4 progeny are observed per plant

$$(3/4)^4 = .316$$

of the heterozygous parent plants would be expected to have no mutant type offspring and would be wrongly classified. If n were increased to 10, however, this fraction of the heterozygous parent plants would be reduced to

$$(3/4)^{10} = .056.$$

This doesn't mean that the number observed should necessarily be high. It might well be profitable to test more parent plants observing a rather small number of offspring per plant and use information of the above type to adjust the results for the proportion of heterozygous parents expected to have no mutant types among the small number of progeny observed.

There are a variety of applications of the binomial in the investigation of frequencies of either natural or induced mutations.

XVII. The Poisson distribution.

The ratio of the mean to the variance in the binomial distribution is $\frac{np}{npq}$ or $\frac{np}{np(1-p)}$. As p approaches zero this ratio approaches 1.0 as a limit. Thus in binomial distributions the mean and variance are for all practical purposes equal when p is very small. This special case of the binomial distribution is known as the Poisson distribution.

Even an extremely rare event will be noted occasionally if a large number of cases are checked. The number of occurrences of such an event in a specified large number of opportunities for occurrence will be distributed as follows:

<u>No. of occurrences</u>	<u>Frequency</u>
0	$1/e^m$
1	m/e^m
2	$m^2/2e^m$
3	$m^3/(2)(3)e^m$
4	$m^4/(2)(3)(4)e^m$
etc.	

where m equals mean number of occurrences in the specified number of opportunities.

It will be noted that this distribution is completely specified by only one parameter, m , which is both the mean and variance. This, of course, assumes something close to constancy in the number of opportunities for the event.

The Poisson distribution is known to be important in connection with counts of emissions from radioactive materials. Assuming good technic successive counts over standard time intervals are considered to be distributed in the Poisson fashion. Apparently the probability of any single emission

being counted is very low, and whether it is actually counted or not is a purely random event. Further, the total rate of emission is sufficiently constant to satisfy the assumption of approximate equality of total number per unit time.

The fact that such counts are distributed Poisson-wise allows two useful applications of our knowledge of the distribution. (1) The counting technique can be checked by finding whether the variance of individual counts on the same material is of the order of the mean number of emissions counted. If it is larger than this mean it constitutes evidence for extraneous (other than the effect of randomness in whether a particular emission is actually counted) sources of error in the counts. (2) The variances and standard errors of individual counts, mean counts, and differences between individual or mean counts can be inferred directly from the number of emissions counted if it has been demonstrated that extraneous errors are inconsequential.

Examples.

(1) A simple check for extraneous variance can be made using counts on a series of samples from the same material. If there is no extraneous variance the variance of the counts computed directly from the data should be of the order of the mean of the several counts. As a statistical test for agreement it will suffice to use F with $N - 1$ and infinite degrees of freedom. Suppose one had obtained the following counts:

6000, 5590, 6200, 5810, 6120, 6300
6080, 5670, 6340, 5960, 6100, 5920 .

The mean is 6007.5. The computed variance is 54,257. F is $\frac{54,257}{6,007.5}$
 $= 9.03$ which is much larger than $F_{.01} = 2.24$ for $N - 1 = 11$ and ∞ degrees of freedom (see F table, p. 225, Snedecor).

The conclusion to be drawn would be that sources of error other than randomness in individual emissions being counted or not, had contributed to the variance of the counts. Such factors as non-homogeneity of the material from which the samples were taken or variation in sample preparation could be responsible for such results.

- (2) Assuming all the variance in counts to be of the Poisson source, i.e. due to the random element with respect to individual emissions being counted, variances of individual counts, means, and differences are obtained directly from the magnitude of the count and can be used as a guide to counts necessary for a given purpose. Suppose one were comparing material from two sources and wished to be able to classify a difference as significant at the 5% level if it were as large as 10% of the mean for the two. This means that the standard error of the difference between two counts, s_d , is to be one-half of $\frac{m}{10}$ or less. (Note: If the distribution is truly Poisson and the mean is used as the estimate of the variance, the distribution of the relative deviate rather than the t distribution is used to test significance of differences. For this purpose Table 8.6, p. 180 is used as in Section XI.) The variance of a single count is m ; that of the difference between two counts (see Section IX) is $2m$, and hence the standard error of the difference is $\sqrt{2m}$.

Thus our requirement is that

$$\sqrt{2m} = \frac{1}{2} \left(\frac{m}{10} \right)$$

or less. Solving the above for m we have

$$2m = \frac{m^2}{400}$$

$$800m - m^2 = 0$$

$$m(800 - m) = 0$$

of which the pertinent solution is

$$m = 800 .$$

Thus if the mean of two counts is 800 the standard error of the difference between them is one-half of 80 or 40. This is an extremely simple application but the process can obviously be extended to more intricate situations.

XVIII. Chi-square, χ^2 .

Chi-square will be considered only with respect to its use in testing agreement in observed frequencies of a series of events with frequencies expected on the basis of some a priori hypothesis.

$$\chi^2 = \sum \left[\frac{(o - e)^2}{e} \right]$$

Where e is the expected number in a class and o is the number observed.

Suppose, for example, that a mendelian trait is theoretically controlled by a single gene pair of which one is completely dominant to the other. Then the progeny of heterozygous parents should be 3/4 of one type, 1/4 of the alternative type. Suppose that of 144 individuals observed, 100 are of the dominant type, 44 of the recessive. The corresponding expected numbers are $3/4 \times 144 = 108$ and $1/4 \times 144 = 36$

$$\chi^2 = \frac{(108 - 100)^2}{108} + \frac{(44 - 36)^2}{36} = 2.371$$

The total number of degrees of freedom for chi-square will always be one less than the number of classes unless the expected frequencies are based on the data in any other respect than the requirement that the totals of the expected and observed frequencies be equal. In the above example there is, therefore, 1 degree of freedom. From the chi-square table (9.2, p. 190) we note that with 1 d.f. chi-square will equal or exceed 2.706 in

1 case out of 10 as a result of random variation. Hence in the above example, the data do not furnish critical evidence against the hypothetical 3:1 ratio of the alternative traits. Said another way, the data are in reasonable agreement with the theory.

As another example suppose we wish to make a test for linkage between two mutant genes and have for the purpose crossed double heterozygous (AaBb) individuals with double recessive (aabb) individuals. Four types of progeny (AaBb, Aabb, aaBb, aabb) are expected in the ratio 1:1:1:1, assuming no linkage. Suppose numbers observed were as follows:

<u>Genotype</u>	<u>No. observed</u>	<u>No. expected</u>
AaBb	130	100
Aabb	114	100
aaBb	68	100
aabb	<u>88</u>	<u>100</u>
	400	400

The expected values are computed on the hypothesis of independent segregation between the two loci and equal viability of the four types of gametes and zygotes.

$$\chi^2 = \frac{(30)^2}{100} + \frac{(14)^2}{100} + \frac{(32)^2}{100} + \frac{(12)^2}{100} = 22.64 .$$

Going to table 9.2, we find that with 3 d.f. χ^2 will equal or exceed 11.341 only one time in 100 as a result of chance. The data are clearly in disagreement with at least a part of the hypothesis.

A more specific hypothesis could be tested abandoning the assumption of equal viability. Expected frequencies are then computed so that they will agree with the observed with respect to (1) the fraction of individuals carrying B, and (2) the fraction carrying A, as well as with regard

to total number. The expected frequencies are then

$$\begin{array}{l}
 AaBb \quad \frac{244}{400} \times \frac{198}{400} \times 400 = 120.78 \\
 Aabb \quad \frac{244}{400} \times \frac{202}{400} \times 400 = 123.22 \\
 aaBb \quad \frac{156}{400} \times \frac{198}{400} \times 400 = 77.22 \\
 aabb \quad \frac{156}{400} \times \frac{202}{400} \times 400 = 78.78
 \end{array}$$

$$\chi^2 = \frac{(9.22)^2}{120.78} + \frac{(9.22)^2}{123.22} + \frac{(9.22)^2}{77.22} + \frac{(9.22)^2}{78.78} = 3.527$$

There is only 1 d.f. in this case because the expected numbers were made to agree in three ways to the four observed numbers. From the χ^2 table we learn that with 1 d.f. a χ^2 of 3.527 will occur a little more than 5% of the time by chance. Hence while the data are not in real good agreement with the hypothesis of independent segregation between the two loci, there is a moderate possibility that the discrepancy was due to chance instead of linkage.

The agreement of the observed frequencies of A's and a's to a 1:1 ratio and of the B's and b's to a 1:1 ratio could also be tested individually. For the ratio of A to a types, the observed and expected numbers are as follows:

<u>Type</u>	<u>Observed</u>	<u>Expected</u>
A	244	200
a	156	200

$$\chi^2 = \frac{(44)^2}{200} + \frac{(44)^2}{200} = 19.36, \quad \text{d.f.} = 1.$$

This χ^2 is highly significant.

For the ratio of B to b types the numbers are as follows:

<u>Type</u>	<u>Observed</u>	<u>Expected</u>
B	198	200
b	202	200

$$\chi^2 = \frac{(2)^2}{200} + \frac{(2)^2}{200} = .04, \quad \text{d.f.} = 1.0.$$

It could be concluded that this data was in accord with the hypothesis of equal probability of B and b types, that it did not furnish strong evidence for linkage between the two loci, but that it did not agree with the hypothesis of equal probability of the A and a types.

To summarize, χ^2 can always be used to test agreement of observed frequencies in a set of categories with frequencies that would occur on the average if a certain hypothesis were true. A simple rule for counting degrees of freedom is to note the number of categories in which the observed frequency could be any number and still allow the observed and expected frequencies to agree in the ways specified by the hypothesis. For example, in the above instance if expected numbers are to agree with the observed in total and in ratio of A's to a's and of B's to b's, once the number of one of the four genotypes is set, there is only one value that can be taken by each of the other three frequencies. Hence there is but one degree of freedom.

It has been my experience that the value of lectures on a subject is almost always increased by the possession of an outline of the material prepared by the lecturer. I have found such outlines to serve a valuable purpose in tying the lectures to material available in textbooks even though the subject is covered in much greater detail in existing books. I have prepared this material in hopes that it will contribute to what the members of the group will retain about statistics from the lectures presented to them.

The intent was not to cover the subject matter of the lectures in detail but rather to summarize some of its more important aspects in a similar order and manner to that in which they were originally presented. The purpose was to give the members of the group something to which they could return to refresh their memories on the content of the lectures.

It was possible to use certain illustrative examples from the field of genetics because the group was at the same time receiving lectures in that subject.

R. E. Comstock