Introduction to the article *Degrees of Freedom*.

The article by Walker, H. W. *Degrees of Freedom*. **Journal of Educational Psychology**. 31(4) (1940) 253-269, was transcribed from the original by Chris Olsen, George Washington High School, Cedar Rapids, Iowa. Chris has made every attempt to reproduce the "look and feel" of the article as well as the article itself, and did not attempt in any way to update the symbols to more "modern" notation. Three typographical errors were found in the paper. These errors are noted in the paragraphs below. The article, except for pagination and placement of diagrams, is as it originally appears. The transcribed pages are not numbered to avoid confusion with pagination in the original article.

Typographical errors:

(1) In the section on *t*-distribution (the 7th of these notes) the last sentence should read "The curve is always symmetrical, but is **less** peaked than the normal when n is small."

(2) In the section "(b) Variance of Regressed Values about Total Mean" (the 12th page of these notes) s_x and s_y are reversed in the expression $Y\% - M_y = r\frac{s_x}{s_y}\left(X - M_x\right)$. It should read $Y\% - M_y = r\frac{s_y}{s_x}\left(X - M_x\right)$

(3) In the section "Tests Based on Ratio of Two Variances" (the 14th page of these notes), the second sentence, "we may divide $\frac{s^2r^2}{1}$ by $\frac{s^2\left(1-r\right)^2}{N-2}$ obtaining $\frac{r^2\left(N-2\right)}{1-r^2}$." should read "we may divide $\frac{s^2r^2}{1}$ by $\frac{s^2\left(1-r^2\right)}{N-2}$ obtaining $\frac{r^2\left(N-2\right)}{1-r^2}$."

Another possible confusion to modern ears may come in the section entitled "*F*-distribution and *z*-distribution." The *z*-distribution mentioned is NOT the standardized normal distribution, but is a distribution known as "Fisher's *z* distribution."

A potential problem in reading this file (other than not having Word!) is --[that]-- the equations, which were inserted using MathType from Design Science. Chris used Math Type 4.0, and if you have anything less it could be a problem. A Math Type reader program can be downloaded from the web. --[www.mathtype.com. Follow the paths to "support."]--

Degrees of Freedom. Journal of Educational Psychology. 31(4) (1940) 253-269

DEGREES OF FREEDOM

HELEN M. WALKER

Associate Professor of Education, Teachers College, Columbia University

A concept of central importance to modern statistical theory which few textbooks have attempted to clarify is that of "degrees of freedom." For the mathematician who reads the original papers in which statistical theory is now making such rapid advances, the concept is a familiar one needing no particular explanation. For the person who is unfamiliar with *N*-dimensional geometry or who knows the contributions to modern sampling theory only from secondhand sources such as textbooks, this concept often seems almost mystical, with no practical meaning.

Tippett, one of the few textbook writers who attempt to make any general explanation of the concept, begins his account (p. 64) with the sentence, "This conception of *degrees of freedom* is not altogether easy to attain, and we cannot attempt a full justification of it here; but we shall show its reasonableness and shall illustrate it, hoping that as a result of familiarity with its use the reader will appreciate it." Not only do most texts omit all mention of the concept but many actually give incorrect formulas and procedures because of ignoring it.

In the work of modern statisticians, the concept of degrees of freedom is not found before "Student's" paper of 1908, it was first made explicit by the writings of R. A. Fisher, beginning with his paper of 1915 on the distribution of the correlation coefficient, and has only within the decade or so received general recognition. Nevertheless the concept was familiar to Gauss and his astronomical associates. In his classical work on the *Theory of the Combination of Observations* (Theoria Combinationis Observationum Erroribus Minimis Obnoxiae) and also in a work generalizing the theory of least squares with reference to the combination of observations (Ergänzung zur Theorie der den kleinsten Fehlern unterworfen Combination der Beobachtungen, 1826), he states both in words and by formula that the number of observations is to be decreased by the number of unknowns estimated from the data to serve as divisor in estimating the standard error of a set of observations, or in our terminology $s^2 = \frac{\sum x^2}{N-r}$ where *r* is the number of parameters to be estimated from the data.

The present paper is an attempt to bridge the gap between mathematical theory and common practice, to state as simply as possible what degrees of freedom represent, why the concept is important, and how the appropriate number may be readily determined. The treatment has been made as non-technical as possible, but this is a case where the mathematical notion is simpler than any non-mathematical interpretation of it. The paper will be developed in four sections: (I) The freedom of movement of a point in space when subject to certain limiting conditions, (II) The representation of a statistical sample by a single point in *N*-dimensional space, (III) The import of the concept of degrees of freedom, and (IV) Illustrations of how to determine the number of degrees of freedom appropriate for use in certain common situations.

I. THE FREEDOM OF MOVEMENT OF A POINT IN SPACE WHEN SUBJECT TO CERTAIN LIMITING CONDITIONS

As a preliminary introduction to the idea, it may be helpful to consider the freedom of motion possessed by certain familiar objects, each of which is treated as if it were a mere moving point without size. A drop of oil sliding along a coil spring or a bead on a wire has only one degree of freedom for it can move only on a one-dimensional path, no matter how complicated the shape of that path may be. A drop of mercury on a plane surface has two degrees of freedom, moving freely on a two-dimensional surface. A mosquito moving freely in three-dimensional space, has three degrees of freedom.

Considered as a moving point, a railroad train moves backward and forward on a linear path which is a one-dimensional space lying on a two-dimensional space, the earth's surface, which in turn lies within a three-dimensional universe. A single coördinate, distance from some origin, is sufficient to locate the train at any given moment of time. If we consider a four-dimensional universe in which one dimension is of time and the other three dimensions of space, two coördinates will be needed to locate the train, distance in linear units from a spatial origin and distance in time units from a time origin. The train's path which had only one dimension in a space universe has two dimensions in a space-time universe.

A canoe or an automobile moves over a two-dimensional surface which lies upon a three-dimensional space, is a section of a three-dimensional space. At any given moment, the position of the canoe, or auto, can be given by two coördinates. Referred to a four-dimensional space-time universe, three coördinates would be needed to give its location, and its path would be a space of three dimensions, lying upon one of four.

In the same sense an airplane has three degrees of freedom in the usual universe of space, and can be located only if three coördinates are known. These might be latitude, longitude, and altitude; or might be altitude, horizontal distance from some origin, and an angle; or might be direct distance from some origin, and two direction angles. If we consider a given instant of time as a section through the space-time universe, the airplane moves in a four-dimensional path and can be located by four coördinates, the three previously named and a time coördinate.

The degrees of freedom we have been considering relate to the motion of a point, or freedom of translation. In mechanics freedom of *rotation* would be equally important. A point, which has position only, and no size, can be translated but not rotated. A real canoe can turn over, a real airplane can turn on its axis or make a nose dive, and so these real bodies have degrees of freedom of rotation as well as of translation. The parallelism between the sampling problems we are about to discuss and the movement of bodies in space can be brought out more clearly by discussing freedom of translation, and disregarding freedom of rotation, and that has been done in what follows.

If you are asked to choose a pair of numbers (*x, y*) at random, you have complete freedom of choice with regard to each of the two numbers, have two degrees of freedom. The number pair may be represented by the coördinates of a point located in the *x*, *y* plane, which is a two-dimensional space. The point is free to move anywhere in the horizontal direction parallel to the *xx′* axis, and is also free to move anywhere in the vertical direction, parallel to the *yy′* axis. There are two independent variables and the point has two degrees of freedom.

Now suppose you are asked to choose a pair of numbers whose sum is 7. It is readily apparent that only one number can be chosen freely, the second being fixed as soon as the first is chosen. Although there are two variables in the situation, there is only one independent variable. The number of degrees of freedom is reduced from two to one by the imposition of the condition *x + y = 7*. The point is not now free to move anywhere in the *xy* plane but is constrained to remain on the line whose graph is *x + y = 7*, and this line is a one-dimensional space lying in the original two-dimensional space.

Suppose you are asked to choose a pair of numbers such that the sum of their squares is 25. Again it is apparent that only one number can be chosen arbitrarily, the second being fixed as soon as the first is chosen. The point represented by a pair of numbers must lie on a circle with center at the origin and radius 5. This circle is a one-dimensional space lying in the original two-dimensional plane. The point can move only forward or backward along this circle, and has one degree of freedom only. There were two numbers to be chosen (*N* = 2) subject to one limiting relationship (*r* = 1) and the resultant number of degrees of freedom is $N - r = 2 - 1 = 1$.

Suppose we simultaneously impose the two conditions *x* + *y* = 7 and $x^2 + y^2 = 25$. If we solve these equations algebraically we get only two possible solutions, *x* = 3, *y* = 4, or *x* = 4, *y* = 3. Neither variable can be chosen at will. The point, once free to move in two directions, is now constrained by the equation *x* + *y* = 7 to move only along a straight line, and is constrained by the equation $x^2 + y^2 = 25$ to move only along the circumference of a circle, and by the two together is confined to the intersection of that line and circle. There is no freedom of motion for the point. *N* = 2 and *r = 2*. The number of degrees of freedom is $N - r = 2 - 2 = 0$.

Consider now a point (*x*, *y*, *z*) in three-dimensional space (*N = 3*). If no restrictions are placed on its coördinates, it can move with freedom in each of three directions, has three degrees of freedom. All three variables are independent. If we set up the restriction $x + y + z = c$, where *c* is any constant, only two of the numbers can be freely chosen, only two are independent observations. For example, let $x - y - z = 10$. If now we choose, say, $x = 7$ and $y = 9$, then *z* is forced to be -12. The equation $x - y - z = c$ is the equation of a plane, a two-dimensional space cutting across the original three-dimensional space, and a point lying on this space has two degrees of freedom. $\left(N - r = 3 - 1 = 2.\right)$ If the coördinates of the (*x*, *y*, *z*) point are made to conform to the condition $x^2 + y^2 + z^2 = k$, the point will be forced to lie on the surface of a sphere whose center is at the origin and whose radius is $\sqrt{k}$. The surface of a sphere is a two-dimensional space. (*N* = 3, *r* = 1, $N - r = 3 - 1 = 2$.).

If both conditions are imposed simultaneously, the point can lie only on the intersection of the sphere and the plane, that is, it can move only along the circumference of a circle, which is a one-dimensional figure lying in the original space of three dimensions. ($N - r = 3 - 2 = 1$.) Considered algebraically, we note that solving the pair of equations in three variables leaves us a single equation in two variables. There can be complete freedom of choice for one of these, no freedom for the other. There is one degree of freedom.

The condition *x* = *y* = *z* is really a pair of independent conditions, *x* = *y* and *x* = *z*, the condition *y* = *z* being derived from the other two. Each of these is the equation of a plane, and their intersection gives a straight line through the origin making equal angles with the three axes. If *x* = *y* = *z*, it is clear that only one variable can be chosen arbitrarily, there is only one independent variable, the point is constrained to move along a single line, there is one degree of freedom.

These ideas must be generalized for *N* larger than 3, and this generalization is necessarily abstract. Too ardent an attempt to visualize the outcome leads only to confusion. Any set of *N* numbers determine a single point in *N*-dimensional space, each number providing one of the *N* coördinates of that point. If no relationship is imposed upon these numbers, each is free to vary independently of the others, and the number of degrees of freedom is *N*. Every necessary relationship imposed upon them reduces the number of degrees of freedom by one. Any equation of the first degree connecting the *N* variables is the equation of what may be called a hyperplane (Better not try to visualize!) and is a space of $N - 1$ dimensions. If, for example, we consider only points such that the sum of their coördinates is constant, $\sum X = c,$ we have limited the point to an $N - 1$ space. If we consider only points such that $\sum \left(X - M\right)^2 = k,$ the locus is the surface of a hypershpere with center at the origin and raidus equal to $\sqrt{k}$. This surface is called the locus of the point and is a space of $N - r$ dimensions lying within the original *N* space. The number of degrees of freedom would be $N - r$.

II. THE REPRESENTATION OF A , STATISTICAL SAMPLE BY A POINT IN *N*-DIMENSIONAL SPACE

If any *N* numbers can be represented by a single point in a space of *N* dimensions, obviously a statistical sample of *N* cases can be so represented by a single sample point. This device, first employed by R. A. Fisher in 1915 in a celebrated paper ("Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population") has been an enormously fruitful one, and must be understood by those who hope to follow recent developments.

Let us consider a sample space of N dimensions, with the origin taken at the true population mean, which we will call µ so that $X_1 - m = x_1$, $X_2 - m = x_2$, etc., where $X_1, X_2, \ldots X_N$ are the raw scores of the N individuals in the sample. Let M be the mean and s the standard deviation of a sample of N cases. Any set of N observations determines a single sample point, such as S. This point has N degrees of freedom if no conditions are imposed upon its coördinates.

All samples with the same mean will be represented by sample points lying on the hyper-plane $(X_1 - m) + (X_2 - m) + \ldots + (X_N - m) = N(M - m)$, or $\sum X = NM$, a space of $N - 1$ dimensions.

If all cases in a sample were exactly uniform, the sample point would lie upon the line $(X_1 - m) = (X_2 - m) = (X_3 - m) = \ldots = (X_N - m) = M - m$ which is the line OR in Fig. 1, a line making equal angles with all the coördinate axes. This line cuts the plane $\sum X = NM$ at right angles at a point we may call A. Therefore, A is a point whose coördinates are each equal to $M - m$. By a well-known geometric relationship,

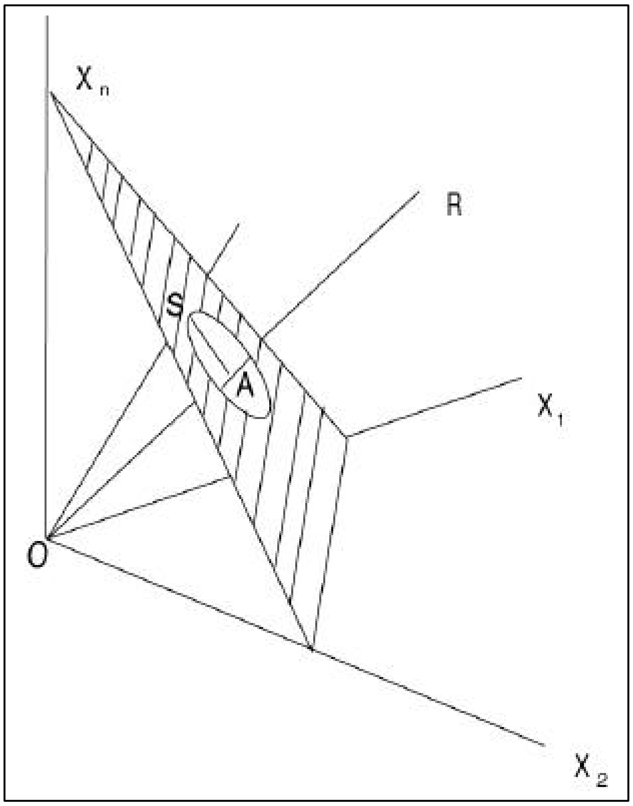

Fig. 1

$$\overline{OS}^2 = (X_1 - m)^2 + (X_2 - m)^2 + ... + (X_N - m)^2$$

$$\overline{OA}^2 = N(M - m)^2$$

$$\overline{OS}^2 = \overline{OA}^2 + \overline{AS}^2$$

$$\overline{AS}^2 = \sum (X - m)^2 - N(M - m)^2 = \sum X^2 - NM^2 = Ns^2$$

Therefore, $OA = (M - m)\sqrt{N}$ and $AS = s\sqrt{N}$. The ratio $\frac{OA}{OS}$ is thus $\frac{M - m}{s}$ and is proportional to the ratio of the amount by which a sample mean deviates from the population mean to its own standard error. The fluctuation of this ratio from sample to sample produces what is known as the *t*-distribution.

For computing the variability of the scores in a sample around a population mean which is known *a priori*, there are available *N* degrees of freedom because the point S moves in *N*-dimensional space about *O*; but for computing the variability of these same scores about the mean of their own sample, there are available only $N - 1$degrees of freedom, because one degree has been expended in the computation of that mean, so that the point *S* moves about *A* in a space of only $N - 1$dimensions.

Fisher has used these spatial concepts to derive the sampling distribution of the correlation coefficient. The full derivation is outside the scope of this paper but certain aspects are of interest here. When we have *N* individuals each measured in two traits, it is customary to represent the *N* pairs of numbers by a correlation diagram of *N* points in two-dimensional space. The same data can, however, be represented by two points in *N*-dimensional space, one point representing the *N* values of *X* and the other the *N* values of *Y*. In this frame of reference the correlation coefficient can be shown to be equal to the cosine of the angle between the vectors to the two points, and to have $N - 2$ degrees of freedom.

III. THE IMPORT OF THE CONCEPT

If the normal curve adequately described all sampling distributions, as some elementary treatises seem to imply, the concept of degrees of freedom would be relatively unimportant, for this number does not appear in the equation of the normal curve, the shape of the curve being the same no matter what the size of the sample. In certain other important sampling distributions -- as for example the Poisson -- the same thing is true, that the shape of the distriution is independent of the number of degrees of freedom involved. Modern statistical analysis, however, makes much use of several very important sampling distributions for which the shape of the curve changes with the effective size of the sample. In the equations of such curves, the number of degrees of freedom appears as a parameter (called *n* in the equations which follow) and probability tables built from these curves must be entered with the correct value of *n*. If a mistake is made in determining *n* from the data, the wrong probability value will be obtained from the table, and the significance of the test employed will be wrongly interpreted. The

Chi-square distribution, the *t*-distribution, and the *F* and *z* distributions are now commonly used even in elementary work, and the table for each of these must be entered with the appropriate value of *n*.

Let us now look at a few of these equations to see the rôle played in them by the number of degrees of freedom. In the formulas which follow, *C* represents a constant whose value is determined in such a way as to make the total area under the curve equal to unity. Although this constant involves the number of degrees of freedom, it does not need to be considered in reading probability tables because, being a constant multiplier, it does not affect the proportion of area under any given segment of the curve, but serves only to change the scale of the entire figure.

Normal Curve.

$$y = C_1 e^{\left(-\frac{x^2}{2s^2}\right)}$$

The number of degrees of freedom does not appear in the equation, and so the shape of the curve is independent of it. The only variables to be shown in a probability table are x/s and y or some function of y such as a probability value.

Chi-square.

$$y = C_2 \left(c^2\right)^{\frac{n-2}{2}} e^{-\frac{x^2}{2}}$$

The number of degrees of freedom appears in the exponent. When $n = 1$, the curve is J-shaped. When $n = 2$, the equation reduces to $y = C_2 e^{-\frac{x^2}{2}}$ and has the form of the positive half of a normal curve. The curve is always positively skewed, but as n increases it becomes more and more nearly like the normal, and becomes approximately normal when n is 30 or so. A probability table must take account of three variables, the size of Chi-square, the number of degrees of freedom, and the related probability value.

t-distribution

$$y = C_3 \left(1 + \frac{t^2}{n}\right)^{-\frac{(n+1)}{2}}$$

The number of degrees of freedom appears both in the exponent and in the fraction t^2/n. The curve is always symmetrical, but is more peaked than the normal when n is small. This curve also approaches the normal form as *n* increases. A table of probability values must be entered with the computed value of *t* and also with the appropriate value of *n*. A few selected values will show the comparison between estimates of significance read from a table of the normal curve and a *t*-table.

For a normal curve, the proportion of area in both tails of the curve beyond 3s is .0027. For a *t*-distribution the proportion is as follows:

n	1	2	5	10	20
p	.204	.096	.030	.014	.007

Again, for a normal curve, the point such that .01 of the area is in the tails is 2.56s from the mean.

For a *t*-distribution, the position of this point is as follows:

n	1	2	3	5	10	20	30
x/s	63.6	9.9	5.8	4.0	3.2	2.8	2.75

F-distribution and z-distribution

$$y = C_4 \frac{F^{\frac{n_1-2}{2}}}{\left(n_1F + n_2\right)^{\frac{n_1+n_2}{2}}} \quad \text{and} \quad y = C_5 \frac{e^{n_1z}}{\left(n_1e^{2z} + n_2\right)^{\frac{n_1+n_2}{2}}}$$

In each of these equations, which provide the tables used in analysis of variance problems, there occurs not only the computed value of *F* (or of *z*), but also the two parameters n_1 and n_2, n_1 being the number of degrees of freedom for the mean square in the numerator of *F* and n_2 the number of degrees of freedom for that in the denominator. Because a probability table must be entered with all three, such a table often shows the values for selected probability values only. The tables published by Fisher give values for *p* = .05, *p* = .01, and *p* = .001; those by Snedecor give *p* = .05 and *p* =01.

Sampling Distribution of r.

This is a complicated equation involving as parameters the true correlation in the population, ?; the observed correlation in the sample, *r*; and the number of degrees of freedom. If $r = 0$ the distribution is symmetrical. If $r \neq 0$ and *n* is large, the distribution becomes normal. If $r \neq 0$ and *n* is small the curve is definitely skewed. David's *Tables of the Correlation Coefficient* (Issued by the Biometrika Office, University College, London, 1938) must be entered with all three parameters.

IV. DETERMINING THE APPROPRIATE NUMBER OF DEGREES OF FREEDOM

A universal rule holds: the number of degrees of freedom is always equal to the number of observations minus the number of necessary relations obtaining among these observations. In geometric terms, the number of observations is the dimensionality of the original space and each relationship represents a section through that space restricting the sample point to a space of one lower dimension. Imposing a relationship upon the observations is equivalent to estimating a parameter from them. For example, the relationship $\sum X = NM$ indicates that the mean of the population has been estimated from observaitons. The number of degrees of freedom is also equal to the number of independent observations, which is the number of original observations minus the number of parmeters estimated from them.

Standard Error of a Mean. --This is $s_{mean} = s / \sqrt{N}$ when s is known for the population. As s is seldom known *a priori*, we are usually forced to make use of the observed standard deviation in the sample, which we will call *s*. In this case s_{mean} **B**$s/\sqrt{N-1}$, one degree of freedom being lost because deviations have been taken around the sample mean, so that we have imposed one limiting relationship, $\sum X = NM$, and have thus restricted the sample point to a hyperplane of $N-1$ dimensions.

Without any reference to geometry, it can be shown by an algebraic solution that $s\sqrt{N}$ **B**$s\ \sqrt{N-1}$. (The symbol **B** is to be read "tends to equal" or "approximates.")

Goodness of Fit of Normal Curve to a Set of Data.--The number of observations is the number of intervals in the frequency distribution for which an observed frequency is compared with the frequency to be expected on the assumption of a normal distribution. If this normal curve has an arbitrary mean and standard deviation agreed upon in advance, the number of degrees of freedom with which we enter the Chi-square table to test goodness of fit is one less than the number of intervals. In this case one restriction is imposed; namely $\sum f = \sum f'$ where *f* is an observed and f' a theoretical frequency. If, however, as is more common, the theoretical curve is made to conform to the observed data in its mean and standard deviation, two additional restrictions are imposed; namely

$\sum fX = \sum f'X$ and $\sum f\left(X-M\right)^2 = \sum f'\left(X-M\right)^2$, so that the number of degrees of freedom is three less than the number of intervals compared. It is clear that when the curves are made to agree in mean and standard deviation, the discrepancy between observed and theoretical frequencies will be reduced, so the number of degrees of freedom in relation to which that discrepancy is interpreted should also be reduced.

Relationship in a Contingency Table.--Suppose we wish to test the existence of a relationship between trait *A*, for which there are three categories, and trait *B*, for which there are five, as shown in Fig. 2. We have fifteen cells in the table, giving us fifteen observations, inasmuch as an "observation" is now the frequency in a single cell. If we want to ask whether there is sufficient evidence to believe that in the population from which this sample is drawn *A* and *B* are independent, we need to know the cell frequencies which would be expected under that hypothesis. There are then fifteen comparisons to be made between observed frequencies and expected frequencies. But. are all fifteen of these comparisons independent?

If we had *a priori* information as to how the traits would be distributed theoretically, then all but one of the cell comparisons would be independent, the last cell frequency being fixed in order to make up the proper total of one hundred fifty, and the degrees of freedom would be $15-1=14$. This is the situation Karl Pearson had in mind when he first developed his Chi-square test of goodness of fit, and Table XII in Vol. I of his *Tables for Statisticians and Biometricians* is made up on the assumption that the number of degrees of freedom is one less than the number of observations. To use it when that is not the case we merely readjust the valuof *n* with which we enter the table.

In practice we almost never have *a priori* estimates of theoretical frequencies, but must obtain them from the observations themselves, thus imposing restrictions on the number of independent observations and reducing the degrees of freedom available for estimating reliability. In this case, if we estimate the theoretical frequencies from the data, we would estimate the frequency $f'_{11} = \left(20\right)\left(40\right)/150$ and others in similar fashion. Getting the expected cell frequencies from the observed marginal frequencies imposes the following relationships:

$$\begin{aligned}
(a)\quad & f_{11}+f_{21}+f_{31}+f_{41}+f_{51}=40 \\
& f_{12}+f_{22}+f_{32}+f_{42}+f_{52}=60 \\
& f_{13}+f_{23}+f_{33}+f_{43}+f_{53}\ =50 \\
(b)\quad & f_{11}+f_{12}+f_{13}=20 \\
& f_{21}+f_{22}+f_{23}=20 \\
& f_{31}+f_{32}+f_{33}=35 \\
& f_{41}+f_{42}+f_{43}=30 \\
& f_{51}+f_{52}+f_{53}=50 \\
(c)\quad & f_{11}+f_{21}+...+f_{51}+f_{12}+...+f_{53}=150
\end{aligned}$$

	A_1	A_2	A_3	
B_1	12	3	5	20
B_2	3	6	11	20
B_3	3	30	2	35
B_4	9	14	7	30
B_5	13	7	25	45
	40	60	50	150

FIG. 2-Observed joint frequency distribution of two traits *A* and *B*.

	A_1	A_2	A_3	
B_1	f_{11}	f_{12}	f_{13}	20
B_2	f_{21}	f_{22}	f_{23}	20
B_3	f_{31}	f_{32}	f_{33}	35
B_4	f_{41}	f_{42}	f_{43}	30
B_5	f_{51}	f_{52}	f_{53}	45
	40	60	50	150

FIG. 3.--Observed marginal frequencies of two traits *A* and *B*.

At first sight, there seem to be nine relationships, but it is immediately apparent that (c) is not a new one, for it can be obtained either by adding the three (a) equations or the five (b) equations. Also any one of the remaining eight can be obtained by appropriate manipulation of the other seven. There are then only seven independent necessary relationships imposed upon the cell frequencies by requiring them to add up to the observed marginal totals. Thus $n = 15 - 7 = 8$ and if we compute Chi-square, we must enter the Chi-square table with eight degrees of freedom. The same result can be obtained by noting that two entries in each row and four in each column can be chosen arbitrarily and there is then no freedom of choice for the remaining entries.

In general in a contingency table, if *c* = number of columns and *r* = number of rows, the number of degrees of freedom is $n = (c-1)(r-1)$ or $n = rc - (r + c - 1)$.

Variance in a Correlation Table.--Suppose we have a scatter diagram with *c* columns, the frequencies in the various columns being $n_1, n_2, ... n_c$, the mean values of *Y* for the columns being $m_1, m_2, ... m_c$, and the regression values of *Y* estimated from *X* being $\tilde{Y}_1, \tilde{Y}_2, ... \tilde{Y}_c$. Thus for any given column, the sum of the *Y*'s is $\sum_1^{n_i} fY = n_i m_i$. For the entire table $N = n_1 + n_2 + ... + n_c$, $NM = \sum_1^c \sum_1^{n_i} fY$, so that $NM = n_1 m_1 + n_2 m_2 + ... + n_c m_c$.

Now we may be interested in the variance of all the scores about the total mean, of all the scores about their own column means, of all the scores about the regression line, of regressed values about the total mean, of column means about the total mean, or of column means about the regression line, and we may be interested in comparing two such variances. It is necessary to know how many degrees of freedom are available for such comparisons.

(a) *Total Variance*.--For the variance of all scores about the total mean, this is $s^2 = \frac{1}{N}\sum_1^N (Y - M)^2$, we have N observations and only one restriction; namely, $\sum fY = NM$. Thus there are $N-1$ degrees, of freedom.

(b) *Variance of Regressed Values about Total Mean*.--The equation for the regressed values being $\tilde{Y} - M_y = r\frac{s_x}{s_y}(X - M_x)$, it is clear that as soon as x is known, y is also known. The sample point can move only on a straight line. There is only one degree of freedom available for the variance of regressed values.

(c) *Variance of Scores about Regression Line*.--There are N residuals of the form $Y - \tilde{Y}$ and their variance is the square of the standard error of estimate, or $s_y{}^2\left(1 - r^2{}_{xy}\right)$. There are N observations and two restrictions; namely,

$$\sum f\left(Y - \tilde{Y}\right) = 0$$

and

$$\sum f\left(Y - \tilde{Y}\right)^2 = Ns^2{}_y\left(1 - r^2{}_{xy}\right).$$

Thus there are $N-2$ degrees of freedom available.

(d) *Variance of Scores about Column Means*.--If from each score we subtract not the regression value but the mean of the column in which it stands, the variance of the residuals thus obtained will be $s_y{}^2\left(1 - E^2\right)$ where E is the correlation ratio obtained from the sample. There are N such residuals. For each column we have the restriction $\sum_1^{n_i} fY = n_i m_i$, making c restrictions in all. The number of degrees of freedom for the variance within columns is therefore $N - c$.

(e) *Variance of Column Means about Total Mean*--To compute this variance we have c observations, *i.e.*, the means of c columns, restricted by the single relation $NM = \sum_1^c n_i m_i$, and therefore have $c-1$ degrees of freedom. The variance itself can be proved to be $s_y{}^2E^2$, and represents the variance among the means of columns,

(f) *Variance of Column Means about Regression Line*.--If for each column we find the difference $m_i - \tilde{Y}_i$ between the column mean and the regression value, and then find

$\frac{1}{N}\sum_{1}^{c} f_i\left(m_i - \tilde{Y}_i\right)^2$, the result will be $s_y{}^2\left(E^2 - r^2\right)$ which is a variance representing the departure of the means from linearity. There is one such difference for each column, giving us c observations, and these observations are restricted by the two relationships $\sum_{1}^{c} f_i\left(m_i - \tilde{Y}_i\right) = 0$ and $\sum_{1}^{c} f_i\left(m_i - \tilde{Y}_i\right)^2 = Ns_y{}^2\left(E^2 - r^2\right)$. Therefore, we have $c-2$ degrees of freedom.

The following scheme shows these relationships in summary form:

Source of variation	Formula	Degrees of Freedom
(d) Scores about column means	$s^2\left(1-E^2\right)$	$N-c$
(e) Means about total mean ....…..	s^2E^2	$c-1$
(a) Total ....………………..	s^2	$N-1$
(c) Scores about regression line...................	$s^2\left(1-r^2\right)$	$N-2$
(b) Regressed values about total mean	s^2r^2	1
(a) Total ....………………	s^2	$N-1$
(d) Scores about column means …….…..	$s^2\left(1-E^2\right)$	$N-c$
(f) Column means about regression line	$s^2\left(E^2-r^2\right)$	$c-2$
(c) Scores about regression line	$s^2\left(1-r^2\right)$	$N-2$
(b) Regressed values about total mean	s^2r^2	1
(f) Column means about regression line	$s^2\left(E^2-r^2\right)$	$c-2$
(e) Column means about total mean.…..	s^2E^2	$c-1$
(*b*) Regressed values about total mean.…..	s^2r^2	1
(*f*) Column means about regression line….	$s^2\left(E^2-r^2\right)$	$c-2$
(*d*) Scores about column means…………..	$s^2\left(1-E^2\right)$	$N-c$
(*a*) Total…………………………………..	s^2	$N-1$

It is apparent that these variances have additive relationships and that their respective degrees of freedom have exactly the same additive relationships.

Tests Based on Ratio of Two Variances.--From any pair of these additive variances, we may make an important statistical test. Thus, to test whether linear correlation exists in the population or not, we may divide $\frac{s^2r^2}{1}$ by $\frac{s^2(1-r)^2}{N-2}$ obtaining $\frac{r^2(N-2)}{1-r^2}$. To test whether a relationship measureable by the correlation ratio exists in the population, we may divide $\frac{s^2E^2}{c-1}$ by $\frac{s^2(1-E^2)}{N-c}$ obtaining $\frac{E^2}{1-E^2}\cdot\frac{N-c}{c-1}$. To test whether correlation is linear, we may divide $\frac{s^2(E^2-r^2)}{c-2}$ by $\frac{s^2r^2}{1}$ obtaining $\frac{E^2-r^2}{r^2(c-2)}$ or may divide $\frac{s^2(E^2-r^2)}{c-2}$ by $\frac{s^2(1-E^2)}{N-c}$ obtaining $\frac{E^2-r^2}{1-E^2}\cdot\frac{N-c}{c-2}$. In each case, the resulting value is referred to Snedecor's F-table which must be entered with the appropriate number of degrees of freedom for each variance. Or we may find the logarithm of the ratio to the base *e*, take half of it, and refer the result to Fisher's *z*-table, which also must be entered with the appropriate number of degrees of freedom for each variance.

Partial Correlation.--For a coefficient of correlation of zero order, there are $N-2$ degrees of freedom. This is obvious, since a straight regression line can be fitted to any two points without residuals, and the first two observations furnish no estimate of the size of r. For each variable that is held constant in a partial correlation, one additional degree of freedom is lost, so that for a correlation coefficient of the *p*th order, the degrees of freedom are $N-p-2$. This places a limit upon the number of meaningful interrelationships which can be obtained from a small sample. As an extreme illustration, suppose twenty-five variables have been measured for a sample of twenty-five cases only, and all the intercorrelations computed, as well as all possible partial correlations--the partials of the twenty-third order will of necessity be either +1 or -1, and thus are meaningless. Each such partial will be associated with $25-23-2$ degrees of freedom. If the partial were not +1 or -1 the error variance $\frac{s^2(1-r^2)}{N-p-2}$ would become infinite, a fantastic situation.

BIBLIOGRAPHY

Dawson, S.: *An Introduction to the Computation of Statistics*. University of London Press, *1933, p. 114. No* general discussion. Gives rule for c^2 only.

Ezekiel, M.: *Methods of Correlation Analysis*. John Wiley & Sons, 1930, p. 121.

Fisher, R. A.: "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population." *Biometrika*, Vol. x, 1915, pp. 507-521. First application of n-dimensional geometry to sampling theory.

Fisher, R. A.: *Statistical Methods for Research Workers.* Oliver and Boyd. This has now gone through seven editions. The term "degrees of freedom" does not appear in the index, but the concept occurs constantly throughout the book.

Goulden, C. H.: *Methods of Statistical* Analysis. John Wiley and Sons, Inc., 1939. See index.

Guilford, J. P.: *Psychometric Methods.* McGraw-Hill, 1936, p. 308.

Mills, F. C.: *Statistical Methods Applied to Economics and Business*. Henry Holt & Co., 2nd ed., 1938. See index.

Rider, P.R.: "A survey of the theory of small samples." *Annals of Mathematics*. Vol. xxxi, 1930, pp. 577-628. Published as a separate monograph by Princeton University Press, $1.00. Gives geometric approach to sampling distributions.

Rider, P. R.: *An Introduction to Modern Statistical Methods.* John Wiley and Sons, Inc., 1939. See index. While there is no general explanation of the meaning of degrees of freedom, this book gives a careful and detailed explanation of how the number of degrees of freedom is to be found in a large variety of situations.

Snedecor, G. W.: *Statistical Methods.* Collegiate Press, Inc., 1937, 1938. See index.

Snedecor, G. W.: *Calculation and Interpretation of Analysis of Variance and Covariance.* Collegiate Press, Inc., 1934, pp. 9-10.

Tippett, L. H. C.: *The Methods of Statistics.* Williams and Norgate, Ltd., 1931. One of the few attempts to treat the concept of degrees of freedom in general terms, but without geometric background, is made on pages 64-65.

Yule and Kendall: *Introduction to the Theory of Statistics.* Charles Griffin & Co. London, 1937, pp. 415-416, 436.