Vol. 46, No. 6 November, 1949

Psychological Bulletin

THE USE AND MISUSE OF THE CHI-SQUARE TEST

DON LEWIS
State University of Iowa
AND
C. J. BURKE
Indiana University

It has become increasingly apparent over a period of several years that psychologists, taken in the aggregate, employ the chi-square test incorrectly. The number of applications of the test does not seem to be increasing, but the number of misapplications does. This paper has been prepared in hopes of counteracting the trend. Its specific aims are to show the weaknesses in various applications that have been made and to set forth clearly the circumstances under which χ^2 can be legitimately applied in testing different hypotheses.

To confirm a general impression that the number of misuses of χ^2 has become surprisingly large, a careful survey was made of all papers published in the *Journal of Experimental Psychology*[1] during the three years 1944, 1945, and 1946. Fourteen papers were found which contained one or more applications of the chi-square test. The applications in only three of these papers (1, 15, 18) were judged to be acceptable. In one other paper (21), the several applications could be called "correct in principle" but they involved extremely small theoretical frequencies. In nine of the fourteen papers (2, 4, 10, 16, 17, 20, 28, 29, 30) the applications were clearly unwarranted. In the remaining case,[2] it was not possible to determine what had been done; and the author, when questioned twice by letter, did not choose to reply.

The principal sources of error (or inaccuracy) in the fourteen papers

[1] The choice of this particular journal resulted from a belief that the psychologists who publish in it are probably better versed, on the average, in statistical methodologies than are those publishing in other journals. No criticism of the journal nor of individual authors nor of experimental findings is intended. The sole purpose is to illustrate correct and incorrect applications of the chi-square test.

[2] This paper, by Pronko (27), fails to provide the reader with any basis whatever for forming an independent judgment relative to the correctness or incorrectness of the two applications of χ^2 which were made. In this respect it is a good example of the current trend in papers published in psychological journals to reduce explanations of methods of analyzing data to a point where they are quite unintelligible.

just referred to, as well as in papers published in other journals, are as follows:

1. Lack of independence among the single events or measures[3]
2. Small theoretical frequencies
3. Neglect of frequencies of non-occurrence
4. Failure to equalize ΣF_o (the sum of the observed frequencies) and ΣF_t (the sum of the theoretical frequencies)
5. Indeterminate theoretical frequencies
6. Incorrect or questionable categorizing
7. Use of non-frequency data
8. Incorrect determination of the number of degrees of freedom
9. Incorrect computations (including a failure to weight by N when proportions instead of frequencies are used in the calculations)

These errors will be explained in detail and illustrated with examples taken for the most part from books and published papers.

It is not surprising that errors of the types listed are frequently made; several of the standard texts to which psychologists turn for statistical guidance contain faulty illustrations. For example, Peters and Van Voorhis (26) make four applications of χ^2, only one of which is without flaws; and in one of the applications made by Guilford (13, p. 91), there is a failure to equalize ΣF_o and ΣF_t and to calculate the number of degrees of freedom correctly.

A single application made by Peters and Van Voorhis contains the first four errors in the above list. Table I is based on their Table Table XXXV (26, p. 411). Twelve dice were thrown fourteen times, and a record was kept of the number of aces appearing at each throw. The observed frequencies F_o are entered in the second column of the table. A value of χ^2, given in the last column, was calculated for each of the fourteen throws. A composite value of χ^2 was obtained by summing the separate values. The first of the four errors in this application is that the observed frequencies lack independence. They lack independence because the same twelve dice were thrown each time. This means that, when the frequencies are grouped, it is impossible to take into account the effects of individual differences in the dice and possible compensating effects from one die to another. As a consequence, no statements can be made about the behavior of an individual die, nor

[3] The term independence, as here used, has reference to individual or single events. In contrast, the hypothesis of independence that is tested by means of χ^2 specifies a lack of relationship (that is, an absence of interaction) between the variates represented in a contingency table. The events that occur to yield the frequencies of a contingency table must be mutually independent even though the variates are related.

is it possible to generalize the findings to any population of dice from which the twelve can be considered a sample. Therefore, only hypotheses which relate specifically to the twelve dice as a group can be tested. More will be said later about this kind of error.

The second flaw in the application comes from using theoretical frequencies of 2. These values are too low to yield a quantity whose

TABLE I

APPLICATION OF THE CHI-SQUARE TEST BY PETERS AND VAN VOORHIS (26)

Throw	F_o	F_t	$(F_o - F_t)^2$	$\frac{(F_o - F_t)^2}{F_t}$
1	1	2	1	0.5
2	3	2	1	0.5
3	2	2	0	0.0
4	3	2	1	0.5
5	1	2	1	0.5
6	4	2	4	2.0
7	2	2	0	0.0
8	4	2	4	2.0
9	1	2	1	0.5
10	0	2	4	2.0
11	3	2	1	0.5
12	2	2	0	0.0
13	3	2	1	0.5
14	1	2	1	0.5
Σ	30	28		10

distribution approximates the χ^2 distribution. The third mistake is the failure to equalize ΣF_o and ΣF_t, which are shown in the table as 30 and 28 respectively. This mistake is related to a fourth one—a failure to take account of the frequencies of non-occurrence of aces. Any of the four errors is sufficient to invalidate this use of the χ^2 test.[4]

[4] If the observed and theoretical frequencies of non-occurrence of aces had been used in the calculations, the composite value of χ^2 would have been 12 instead of 10. This difference happens not to be large. But in another illustration used by Peters and Van Voorhis (26, Table XXXVI, p. 414), the difference is large. The value of 14.52 is given in the text. When χ^2 is correctly computed by taking account of the frequencies of non-occurrence, the resulting value is 29.72. The number of degrees of freedom remains the same, but the calculated value of χ^2 is more than doubled.

Fundamental Theory

The two most basic requirements in any application of the chi-square test are (a) independence among the separate measures and (b) theoretical frequencies of reasonable size. These requirements can be shown in an elementary way by examining a two-category distribution of measures. But first, an unequivocal definition of χ^2 is needed.

If z is a normal deviate in standard form defined in relation to population parameters m and σ, then

$$z = \frac{X - m}{\sigma}$$

and

$$\chi^2 = z^2 = \frac{(X - m)^2}{\sigma^2}, \text{ with } df = 1. \qquad [1][5]$$

Chi-square with 1 degree of freedom is thus defined as the square of a deviation from the population mean divided by the population variance.

If there are r *independent* measures of the variate X, there will be r independent values of z, and the resulting formula for χ^2 is

$$\chi^2 = \sum_{i=1}^{r} z_i^2 = \sum \frac{(X_i - m)^2}{\sigma^2}, \text{ with } df = r. \qquad [2]$$

Values of χ^2 may range from 0 to ∞, and they have frequency distributions which depend upon the value of r. The distribution function of χ^2, in general form, will be given later. It will then be made clear that the chi-square tests of independence and goodness of fit can be applied unequivocally only to frequency data (or to proportions derived from frequency data). For the present, the plausibility of the two basic requirements stated above will be shown through an examination of the two-category case.

The two-category case. Consider a population of N *independent* events (things; measures), each of which may fall either into category A or into category B. It is assumed that these categories are clearly defined before samples are drawn and that the category in which a given event falls can be unequivocally determined. A sample is drawn from the population and the sample data are to be employed in determining whether or not a certain hypothesis regarding the proportion of cases in each category is tenable. If p is the expected (theoretical) proportion

[5] Equation [1] verifies the statement that the square root of χ^2 with $df = 1$ is distributed as z (or Student's t) with $df = \infty$.

for category A and q the expected proportion for category B, it follows that

$$p + q = 1.$$

The probability $P(n)$ that n of the N events will fall into category A is given by the binomial distribution function

$$P(n) = \frac{N!}{n!(N-n)!} p^n q^{N-n}, \qquad [3]$$

which is the general expression for obtaining the successive terms arising from the expansion of the binomial $(p+q)^N$. The limiting form of equation [3], as N becomes indefinitely large, is a normal distribution function having a mean of Np and a variance of Npq. In symbols,

$$\lim_{N\to\infty} P(n) = \frac{1}{\sqrt{2\pi}\sqrt{Npq}} e^{-(n-Np)^2/2Npq}. \qquad [4][6]$$

Np is the *population* mean of category A. It is the expected value of n, that is, the theoretical frequency to be associated with category A. Nq is the corresponding theoretical frequency to be associated with category B. Npq is the *population* variance. As stated in equation [1], the square of a deviation from the population mean divided by the population variance is distributed as χ^2 with 1 *df*. Thus, from equation [4], it is seen that the quantity

$$\chi^2 = \frac{(n-Np)^2}{Npq} \qquad [5]$$

would be distributed exactly as χ^2 with 1 *df*, provided that N is indefinitely large. In the two-category case, equation [5] may be employed to calculate an approximate value of χ^2. It should be noted that both p and q appear in the denominator of the right-hand term. No restriction is placed during the calculation; the equation gives an approximate solution for any value of N.

The formula that is commonly used in the two-category case to obtain a value of χ^2 is

$$\chi'^2 = \frac{(n-Np)^2}{Np} + \frac{(N-n-Nq)^2}{Nq}, \qquad [6]$$

where Np and Nq are theoretical frequencies and n and $(N-n)$ are the

[6] If the investigator is concerned with the probability $P(N-n)$ that $(N-n)$ events will fall into category B, the limiting form of equation [3] would be written

$$\lim_{N\to\infty} P(N-n) = \frac{1}{\sqrt{2\pi}\sqrt{Npq}} e^{-(N-n-Nq)^2/2Npq}. \qquad 4a]$$

corresponding observed frequencies. The prime symbol is placed on χ^2 in [6] to distinguish between χ^2 as *defined* by equation [5] and χ^2 as ordinarily *calculated* with formula [6]. When [6] is used, the number of *df* is 1 less than the number of categories because one restriction ($\Sigma F_t = N$) is imposed on the theoretical frequencies. Therefore, the number of *df* for [6] is 1, just as it is for equation [5].

It can readily be shown that χ^2 and χ'^2 are identical quantities. It is for this reason, and this reason alone, that formula [6] may be used in obtaining an estimate of χ^2 in the two-category case.[7]

The foregoing discussion reveals the two limitations that hold in any application of the chi-square test. The first limitation is that χ^2 is correctly used only if the N events or measures are independent. Equations [3] and [4] are valid statements only when independence exists. The second basic limitation relates to the size of the theoretical frequencies. If Np (or Nq) remains small as N becomes large, the limiting form of the binomial distribution function is not a normal distribution, as was assumed in writing equation [4]. If, for any reason, either Np or Nq is small, the limiting form of equation [3] is the Poisson distribution function. Under such circumstances, the quantity on the right of equation [5] would not be the square of a normal deviate divided by the population variance and, consequently, would not be distributed as χ^2 with 1 *df*.

It should be emphasized that the categories are assumed to be designated in the population before the individual sample is drawn. It should also be emphasized that the equating of the sums of observed and theoretical frequencies *and* the use of the frequency of non-occurrence (in this case, the frequency with which measures fall in category B) are necessary to establish the identity between the quantities defined in equation [5] and equation [6].

[7] If observed and theoretical proportions are used in calculating values of χ^2, equations [5] and [6] become

$$\chi^2 = N\frac{(p_o - p_t)^2}{p_t q_t} \qquad [5a]$$

or

$$\chi^2 = N\frac{(q_o - q_t)^2}{p_t q_t}; \qquad [5b]$$

and

$$\chi^2 = N\frac{(p_o - p_t)^2}{p_t} + N\frac{(q_o - q_t)^2}{q_t}. \qquad [6a]$$

In these equations, N is the total number of cases while p_o and q_o are the observed proportions and p_t and q_t are the theoretical proportions for categories A and B, respectively. The equations may be derived from [5] and [6], and they reveal that if proportions are used instead of frequencies, the values calculated from the proportions must be multiplied by N.

The more general case. It is common to have frequency data that fall into several categories instead of just two. This fact requires an extension of the ideas discussed in the two-category case to encompass any number of categories. The basic features of this extension will now be presented. Actually, no new ideas enter into the development. The proof is mathematically more complex, but the underlying ideas are the same.

Consider a population of N *independent* events, with k possible outcomes, $v_1, v_2, v_3 \cdots v_k$. Assume that, in the population,

v_1 occurs with a probability of p_1
v_2 occurs with a probability of p_2
v_3 occurs with a probability of p_3
. . . . . .
v_k occurs with a probability of p_k

The *joint* probability $P(n_i)$ that out of N events exactly n_1 will fall in category v_1, n_2 will fall in category v_2, n_3 in v_3, *and* $\cdots n_k$ in v_k, is given by the multinomial distribution function

$$P(n_i) = \frac{N!}{n_1!n_2!n_3! \cdots n_k!} p_1^{n_1} p_2^{n_2} p_3^{n_3} \cdots p_k^{n_k}, \qquad [7]$$

where

$$\sum_{i=1}^{k} n_i = N.$$

This is the fundamental expression from which the distribution function of χ^2 is derived. It confirms the statement that the measures (frequencies) in the various cells (categories, classes, etc.) of a multidimensional table must be mutually independent to enable a legitimate application of χ^2 in testing any hypothesis concerning the table. And because equation [7] is written in terms of the frequencies $n_1, n_2, n_3 \cdots n_k$, the chi-square tests of independence and goodness of fit, based as they are on a distribution function derived from [7], may be used unequivocally only in relation to frequency data.[8]

The distribution function of χ^2, here symbolized by $g_r(\chi^2)$, may be written

$$g_r(\chi^2) = C_r(\chi^2)^{(r+2)/2} e^{-(\chi^2/2)}, \qquad [8]$$

where C_r, a coefficient which changes with r, is given by

[8] This statement limiting the use of χ^2 to frequency data is not meant to exclude certain special applications such as finding confidence limits for a population variance from a known sample variance or testing several sample variances for homogeneity. These special applications are mentioned again toward the end of the paper.

$$C_r = \frac{1}{2^{(r/2)}\,\Gamma(r/2)}. \qquad [9]^9$$

It should be emphasized that equation [8] is an exact distribution function for the quantity defined in equations [2] and [5] but only an approximation of the distribution of the quantity defined by equation [7]. Its used in relation to equation [7] requires three separate approximations, each of which assumes a theoretical frequency of reasonable size. The three approximations are:

1. Replacing each of the factorials in equation [7] by its Stirling approximation.
2. Taking a step similar to the one whereby $(1+[X/n])^n$ is replaced by e^X when n is large.
3. Substituting a continuous integral for a summation of discrete quantities.

All of these approximations are quite acceptable and lead to inconsequential errors so long as Np is reasonably large. This reaffirms the fundamental requirement that *the theoretical frequencies must not be small*, if any calculated value of χ^2 is to be distributed as χ^2.[10]

General computational formula. The formula that is commonly employed in calculating values of χ^2 is

$$\chi^2 = \sum \frac{(F_o - F_t)^2}{F_t}, \qquad [10]$$

[9] This expression for C_r contains the gamma function $\Gamma(r/2)$. A gamma function is a function which reduces to a factorial whenever the argument is an integer. In the general case, $(n-1)! = \Gamma(n)$. Equation [9] may be written

$$C_r = \frac{1}{2^{(r/2)}\left(\frac{r-2}{2}\right)!}. \qquad [9a]$$

Whenever r is an even number, the factorial in [9a] is an integral number, and its value along with the value of C_r, can be determined in a straight forward manner. On the other hand, if r is an odd number, the "factorial" is fractional, and its value must be determined either by referring to a table of the gamma function or by using the equation

$$\left(\frac{r-2}{2}\right)! = \left(\frac{r-2}{2}\right)\left(\frac{r-4}{2}\right)\left(\frac{r-6}{2}\right)\cdots\left(\frac{r-[r-1]}{2}\right)(\sqrt{\pi}).$$

A reader who is interested in plots of the χ^2 distribution function for various values of r is referred to Lewis (19).

[10] A derivation of equation [8] from equation [7] is presented in considerable detail by Greenhood (11) and can be followed by persons familiar with advanced calculus. Greenhood's development, which is more complete but similar to one given by Fry (8), indicates clearly the limitations which hold in applications of the chi-square test to frequency data. It is not necessary, of course, to derive equation [8] from the multinomial distribution function; it may be derived directly from the joint normal distribution function in n variables.

where F_o and F_t, as usual, are observed and theoretical frequencies and the summation extends over all cells (categories) of the table. The equation holds for any number of categories and reduces to the form of [6] when the number of categories is 2. It serves as a constant reminder that the chi-square tests of independence and goodness of fit can be applied unequivocally only to frequency data.

Applications: I. The Goodness of Fit of Distribution Functions

One of the commonest applications of the chi-square test is in evaluating the hypothesis that a set of frequency data can be satisfactorily represented by some specified distribution function. It makes little difference what the function is, so long as its fundamental properties are known. The goodness of fit of binomial, Poisson, and normal distribution functions is often tested.

Two-category case. The correct use of χ^2 in connection with a symmetrical binomial distribution function (where $p = q$) can be illustrated with data obtained in a coin-guessing "experiment." A coin was tossed, and 96 students of elementary psychology each guessed whether the coin came up "heads" or "tails." The hypothesis to be tested is that the guess of each student, like the fall of the coin itself, was a purely chance occurrence and that each student was as likely to say heads as to say tails. The results are shown in Table II.

TABLE II

Data from a Coin-Guessing Experiment

H_o ($= F_o$ for heads)	68
H_t ($= F_t$ for heads)	48
T_o ($= F_o$ for tails)	28
T_t ($= F_t$ for tails)	48

As seen in the first and third rows, 68 students guessed heads and 28 guessed tails. The theoretical frequencies are 48 and 48. The value of χ^2 is obtained as follows:

$$\chi^2 = \frac{(H_o - H_t)^2}{H_t} + \frac{(T_o - T_t)^2}{T_t} = \frac{(68 - 48)^2}{48} + \frac{(28 - 48)^2}{48} = 16.67.^{11}$$

With 1 df, this value is significant at better than the 0.1% level of confidence;

[11] Two alternative formulas that may be used in the two-category case when $p = q$ are as follows:

$$\chi^2 = (H_o - H_t)^2(2/H_t);$$

$$\chi^2 = \frac{(H_o - T_o)^2}{H_o + T_o}.$$

These formulas are the exact equivalents of the one used above.

so it must be concluded that the guesses of the 96 students were somehow biased in favor of heads.

Additive property of χ^2, illustrated with two-category data. A fundamental property of χ^2 is indicated by the rule which states that the sum of any number of separate and independent values of χ^2 is distributed as χ^2, the number of *df* being the sum of the separate *df*'s. The rule will now be applied in relation to coin-guessing data. Two applications will be made, the first of which is incorrect. *This incorrect application is purposely included* as a means of showing how it differs from a correct application and also of revealing the source of many of the errors made by investigators when employing the chi-square test.

Ninety-six students of elementary psychology were given five successive "trials" in coin guessing. A single coin was tossed five times. After each toss, the 96 students each guessed whether the coin came up "heads" or "tails." Each student wrote his five guesses in order on a sheet of paper. The turn of the coin was never revealed. The results are summarized in Table III. The

TABLE III

DATA FROM A SECOND COIN-GUESSING EXPERIMENT

	Tosses					
	1	2	3	4	5	
H_o	68	49	39	54	54	
H_t	48	48	48	48	48	
T_o	28	47	57	42	42	
T_t	48	48	48	48	48	
χ^2	16.67	.04	3.37	1.50	1.50	$\sum = 23.08$

frequencies for toss 1 are the same as those given in Table II.

With the ten pairs of observed and theoretical frequencies in Table III (two pairs for each toss), it is possible to test five separate hypotheses—that the guesses of the 96 students on *any one* of the five tosses were chance occurrences and were as apt to be heads as tails. Five values of χ^2 are given in the bottom row of the table. Each of the values was computed with the formula employed with the data of Table II. The number of degrees of freedom is 1 in each case. For 1 *df*, the value of χ^2 at the 5% level of confidence is 3.841. Four of the five calculated values are less than this and provide no satisfactory basis for rejecting the hypothesis of chance occurrence.

To secure an estimate of χ^2 which makes the probability of rejecting a false hypothesis large, it is desirable, when conditions warrant, to summate separate estimates of χ^2 and obtain a composite estimate. The number of *df* for the composite estimate is always the sum of the separate *df*'s. The sum of the five values in Table III is 23.08. The number of *df* is 5. The hypothesis under test is that the guesses of students *on five successive tosses* of a coin are purely chance occurrences, with the probability of a guess of heads (by any student on any toss) equal to the probability of a guess of tails. The composite value of

23.08 is significant at better than the 0.1% level of confidence, and would warrant a rejection of the hypothesis if the application of the test were correct.

But it is not correct to summate the five values of χ^2 in Table III. The reason is that the responses of the 96 students from one toss to the next cannot be assumed to have been independent. In other words, the tabulated values of H_o and T_o from toss to toss are interdependent. Consequently, there is no way of obtaining an unbiased estimate of the theoretical frequency for any toss beyond the first (unless previous guesses are completely ignored). It is unreasonable to assume that a student's knowledge of his guess on one toss did not influence his guess on succeeding tosses. (It will be shown from the data that such an assumption is unsound.) The χ^2 test should never be based on an assumption that is already known to be false. On a single toss, the guess of each student was independent of the guess of the other students; but between tosses, the five guesses of each student were undoubtedly interrelated.

A lack of independence between separate events (measures) is the commonest flaw in the applications of χ^2 that are made by psychologists. Six of the fourteen papers (2, 10, 17, 28, 29, 30) referred to in the opening paragraphs contain applications having this shortcoming.

There are two ways in which the responses of subjects may be interdependent. They may be related from trial to trial, as they were in the coin-guessing illustration, or they may be internally linked within a single trial. Whenever individual subjects each make more than one response per trial, linkages among the measures within the trial must result unless there are no individual differences. Many investigators ignore this restriction and apply the χ^2 test even though the same subjects are used from trial to trial and make several responses on each trial.

The correct use of the additive principle with two-category data. The conditions under which separate values of χ^2 may be legitimately summated can be illustrated with two-category data. Five non-overlapping groups of subjects made a single guess on each of five successive tosses of a coin. The number of subjects per group was 86. Each subject wrote down his guesses in order on a sheet of paper, the sequence of his guesses remaining on the sheet before him. The results for the five groups are summarized in Tables IV (A to E, inclusive). Twenty-five separate values of χ^2 appear in the bottom rows of the five sections of the table. The question to be answered is: *Which of these 25 values may be legitimately summated and which may not be?* The five values in any one of the sections cannot be meaningfully added for reasons already given in the discussion of the data in Table III. However, the five values in the five sections for *any single toss* may be summated to yield a single composite value having 5 *df*. For example, a composite value may be obtained for the third toss and may be used to test the hypothesis that, on the third toss of the coin, the guesses of the members of the five groups were random occurrences with the probability of a heads guess equalling that of a tails guess. Neither the presence nor the absence of individual biases would nullify a meaningful test of this hypothesis. If the hypothesis could not be retained, it would be correct to conclude that the guesses were not chance occurrences and were perhaps influenced by what had gone before. The difference between this case and the one illustrated in Table III is clear-cut. The data in Table III are for five successive guesses by one group, and the successive theoretical probabilities cannot be established

TABLE IV

DATA FROM A THIRD COIN-GUESSING EXPERIMENT

	Tosses				
	1	2	3	4	5
			A. Group 1		
H_o	63	56	36	45	41
H_t	43	43	43	43	43
T_o	23	30	50	41	45
T_t	43	43	43	43	43
χ^2	18.60	7.86	2.28	0.19	0.19
			B. Group 2		
H_o	63	56	42	48	45
H_t	43	43	43	43	43
T_o	23	30	44	38	41
T_t	43	43	43	43	43
χ^2	18.60	7.86	0.05	1.16	0.19
			C. Group 3		
H_o	65	55	40	48	46
H_t	43	43	43	43	43
T_o	21	31	46	38	40
T_t	43	43	43	43	43
χ^2	22.51	6.70	0.42	1.16	0.42
			D. Group 4		
H_o	68	54	38	52	41
H_t	43	43	43	43	43
T_o	18	32	48	34	45
T_t	43	43	43	43	43
χ^2	29.07	5.62	1.16	3.77	0.19
			E. Group 5		
H_o	72	57	30	62	38
H_t	43	43	43	43	43
T_o	14	29	56	24	48
T_t	43	43	43	43	43
χ^2	39.12	9.12	7.86	16.79	1.16

without making specific assumptions concerning prior events. In contrast, when the five χ^2 values for a single toss are taken from the five sections of Table IV and summated, the only assumption that is made is that the guesses on that particular toss were in accordance with a theoretical probability for heads of .50.

Composite values of χ^2 for the five tosses are given in Part A, Table V.

TABLE V

VALUES OF CHI-SQUARE BASED UPON DATA FROM TABLE IV

Part A. Composite Values of χ^2		
Toss	χ^2	*df*
1	127.90	5
2	37.16	5
3	11.77	5
4	23.07	5
5	2.15	5
Part B. Values of χ^2 for Combined Frequencies		
Toss	χ^2	*df*
1	125.17	1
2	36.92	1
3	7.82	1
4	14.88	1
5	.149	1

The number of *df* in each case is 5. Except for toss 5, the composite values are all significant at far better than the 1% level of confidence. The value for toss 5 falls near the 80% level and lends support for the belief that on a fifth successive toss of a coin, the probability of a heads guess is .50. The values for tosses 1 and 2 show that, on these tosses, the guesses were strongly biased toward heads. Conclusions regarding tosses 3 and 4 should be made in the light of the large contributions made to the composite values for these tosses by the guesses of group 5 alone. As seen in Table IV-E, the χ^2 values for tosses 3 and 4 are 7.86 and 16.79, respectively. With 1 *df* in each case, both values are significant at better than the 1% level. The corresponding values in the other four parts of the table all fall below the 5% level. However, the deviations on tosses 3 and 4 are in the same direction for all five groups, and this fact indicates a definite departure from chance expectations.

It is legitimate to combine for each toss separately the empirical and theoretical frequencies listed in Table IV-A-B-C-D-E, and use the resulting sums to compute values of χ^2, each with 1 *df*. Values of χ^2 obtained in this way are given in Part B of Table V. As in the case of the composite values, all of the values except the one for toss 5 are significant at better than the 1% level of confidence. Nevertheless, the procedure of combining frequencies is not recommended except where the theoretical frequencies for each of several duplicated experiments are too small to yield satisfactory individual estimates of

χ^2. Other things equal, the greater the number of degrees of freedom is, the more stable is a value of χ^2 and the greater is the probability of rejecting a false hypothesis.

Five values of χ^2 may be selected from Table IV-A-B-C-D-E and used compositely to test the hypothesis that the guesses of persons on *five successive tosses* of a coin are chance occurrences, with the probability of a heads guess equalling that of a tails guess. The hypothesis is inclusive enough to cover the entire population from which the groups of subjects were randomly selected. To provide for independence between tosses, it is necessary to choose the five values of χ^2 so that there is one value for each toss and so that no two values are based on the guesses of a single group. To this end, numbers from 1 to 5 were assigned to the five sections of the table. The χ^2 value for toss 1 was taken from the section whose number first appeared in a table of random numbers; the χ^2 value for toss 2 from the section whose number next appeared, and so on. The values thus chosen are shown in Table VI. The composite

TABLE VI

COMPOSITE VALUE OF CHI-SQUARE, BASED ON DATA IN TABLE IV

	Tosses				
	1	2	3	4	5
H_o	63	54	30	48	46
H_t	43	43	43	43	43
T_o	23	32	56	38	40
T_t	43	43	43	43	43
χ^2	18.60	5.62	7.86	1.16	0.42

Composite $\chi^2 = 33.66$

value is 33.66, with 5 *df*. It is highly significant and leaves no grounds for believing that the hypothesis is true.

The discussion in the preceding pages on the necessity for independence between measures can perhaps be further clarified through a consideration of coin-tossing. Three different situations will be described to reveal unmistakable differences in hypotheses to be tested and methods of handling data. Suppose, first, that a single penny is selected at random from a large collection of pennies. This penny is tossed successively, say 100 times, and a record is kept of the way it turns. The probability of a head (or a tail) is .50. The χ^2 test may be applied to determine whether or not the empirical results conform to this theoretical probability. If they do conform, it may justifiably be concluded that the penny is "unbiased." The test is an unequivocal one. The extent to which the investigator should generalize to the collection of pennies from which the single penny was chosen is a matter for personal judgment. The χ^2 test, as made, reveals nothing concerning the probability that the selected penny either represents or misrepresents the collection of pennies.

Suppose next that *two* pennies are randomly selected from a collection of pennies and that each penny is tossed 50 times to give a total of 100 tosses.

The frequencies of occurrence of heads and of tails are combined (pooled) for the two coins. It may be assumed that the probability of a head is .50, but the χ^2 test cannot be *meaningfully* applied to test this theoretical probability.[12] The reason is that each penny makes its own unique contribution to the results. If one of them is biased while the other is unbiased, the obtained value of χ^2 could easily be significant and lead to a rejection of the hypothetical probability, even though it is correct for one of the pennies. Furthermore, one penny could be strongly biased for heads and the other equally strongly biased for tails, and the obtained value of χ^2 would turn out to be insignificant. The possible presence of individual idiosyncrasies precludes an unequivocal application of χ^2. The same thing would be true if five pennies were randomly selected, each one tossed, say 20 times, and the results pooled; or if 10 pennies were selected and each tossed 10 times, or 20 selected and each tossed five times.

Suppose, finally, that 100 pennies are randomly selected from a collection of pennies, that each penny is tossed a single time, and that the number of heads is recorded. It may be assumed that the probability of a turn of heads, in the population from which the pennies are selected, is .50. The fall of each coin is clearly independent of the fall of every other coin. The χ^2 test may be legitimately applied to determine whether or not the observed frequency of heads conforms to the hypothetical frequency. The results of the test can be generalized to the entire collection of pennies. This would hold even though less than 100 pennies were selected, so long as a sufficient number was chosen to provide theoretical frequencies of the occurrence of heads and the occurrence of tails of sufficient magnitude to warrant an application of the χ^2 test. No statements can be made, of course, regarding the tendencies of any individual penny.

The crucial point is that frequencies obtained from individuals, whether pennies or subjects in psychological experiments, should not be pooled if the χ^2 test is to be used, except when it can be shown that there is an absence of biases or idiosyncrasies among them[13] or when "interaction" effects are specifically under scrutiny. Results on individuals may be combined, but the combining should be done *after* the χ^2 test has been applied to the data on individuals separately. For example, if two pennies are each tossed 50 times, the χ^2 test may be applied to the results for each penny separately, and then the two values of χ^2 may be added to provide a composite value. Similarly, separate values of χ^2 may be obtained from the guesses made by two individuals. The separate values may then be combined to furnish a single composite value. As in the well-known analysis of variance techniques where each source of variability contributes to the total variability, each source of variability should be allowed to make its contribution to the value of χ^2. Unfortunately, χ^2 procedures provide no way, as analysis of variance techniques do, of introducing *statistical* controls over individual subjects as a source of variability.

[12] The pooling of two or more sets of frequencies to obtain a single value of χ^2 is warranted if the aim is to study the "heterogeneity" or "interaction" aspects of the data. In this connection, see Snedecor's discussion (31, pp. 191–192) of "pooled" and "total" chi-squares.

[13] If an investigator firmly intends to restrict all generalizations to the group of persons studied—the group considered *in toto*, as a sort of amorphous mass—then the pooling of individual frequencies may be logically defended. In such a situation, the group is analogous to a single individual and must be treated as such.

Therefore, in the use of χ^2, the control over individuals must be introduced as an intrinsic part of the sampling process.

Multi-category case (single dimension). Frequency data sometimes fall into several categories along a single dimension. If the frequencies from category to category are independent and if some hypothesis regarding their distribution can be meaningfully set up, then the chi-square test may be used to evaluate the hypothesis, provided that the theoretical frequencies for the various categories are of reasonable magnitude. The correct application of the test in such a situation will be illustrated with data obtained in die throwing. A single die was thrown 120 times. There was no reason for believing that any throw was influenced by any other. The results are given in Table VII, where the first

TABLE VII

DATA OBTAINED IN DIE THROWING

Face	F_o	F_t	$\frac{(F_o - F_t)^2}{F_t}$
1	23	20	.45
2	20	20	.00
3	22	20	.20
4	15	20	1.25
5	18	20	.20
6	22	20	.20
$\sum$	120	120	2.30

column lists the six faces of the die and the second column gives the number of times that each face turned up. The theoretical probability, on each throw, that any specified face of the die would turn up was 1/6. Consequently, the theoretical frequency of occurrence of each of the six faces was 20, as shown in the third column of the table. A single restriction, the sum of the observed frequencies, was placed in calculating the theoretical frequencies. Consequently, there are 5 *df*. With 5 *df*, the obtained value of χ^2 falls near the 80% level of confidence, and there is no basis for rejecting the hypothesis that the fall of the die was, on each throw, a strictly chance occurrence.[14]

[14] It would be possible to use the data in Table VII to test six separate hypotheses—that the appearance of *each* of the six faces was a chance occurrence, with a probability of 1/6. For example, to test the hypothesis that the appearance of the ace (one-spot) was a chance occurrence, a value of χ^2 would be computed as follows:

$$\chi^2 = \frac{(23 - 20)^2}{20} + \frac{(97 - 100)^2}{100} = .45 + .09 = .54.$$

The number of *df* is 1. Observe that in the calculation, the frequency of non-occurrence of the ace was taken into account. (The probability of occurrence of the ace was 1/6 [$=p$] while the probability of its non-occurrence was 5/6 [$=q$]. This is an example of an asymmetrical binomial.) It is clear that six separate values of χ^2 could be computed

Another illustration of a multi-category frequency distribution (along a single dimension) comes from results on the coin-guessing experiment. A total of 439 subjects each guessed heads or tails on each of five successive tosses of a coin. All but 9 of the subjects were the ones whose guesses were tabulated in Table IV. The χ^2 test will be applied to evaluate the hypothesis that chance factors operated in determining the frequencies of occurrence of the various possible "patterns" of guesses. With five successive tosses and five successive guesses, there were 32 possible patterns, as shown in the first column of Table VIII. If there was no biasing of the guesses (that is, if the guess of every subject on every trial was as likely to be heads as to be tails), then each of the 32 patterns was as probable as any other; and with $N = 439$, the theoretical frequency for each pattern was $1/32 \times 439 = 13.7$, rounded to the first decimal place. This is the value shown in the third column of the table, except where parentheses appear. Note that 32×13.7 does not equal 439 exactly, but equals 438.4. One of the requirements in the application of χ^2 is that $\sum F_o = \sum F_t$. Therefore, six of the theoretical frequencies are given as 13.8. These are the theoretical frequencies corresponding to the six largest observed frequencies. This insures that any slight error that may result from equalizing $\sum F_o$ and $\sum F_t$ operates to make the test more conservative.

In the calculation of χ^2 for Table VIII, a value of $(F_o - F_t)^2/F_t$ was secured for each row. These values are given in the fourth column, and their sum ($= 681.122$) is the desired estimate. The number of *df* is $(32-1) = 31$. The only restriction placed in figuring the theoretical frequencies was $\sum F_o$, and this meant the loss of a single degree of freedom. Even with 31 *df*, the value of χ^2 is so large as to leave no basis whatever for retaining the hypothesis that the guesses were chance occurrences.

Normal distribution function. The chi-square test is often used in evaluating the fit of a normal curve to a set of frequency data. Applications of this type are usually correct except for an occasional failure to equalize $\sum F_o$ and $\sum F_t$, a tendency to use some theoretical frequencies that are too small, and, most importantly, an incorrect specification of the number of degrees of freedom. The correct procedure will now be illustrated.

The distribution of the midterm scores of 486 students in a course in elementary psychology is shown in Table IX. The mid-points of class intervals of ten score units are given in the X-column, frequencies in the F_o-column. The mean M of the distribution of scores is 104.0, while the standard deviation is 16.1. There are two methods that can be used in fitting a normal curve to the data (that is, in calculating the theoretical frequencies that correspond to the observed frequencies). One method involves the estimation of areas under segments of the normal curve through the process of multiplying ordinate values by the class interval.[15] This is an approximation procedure. A more exact method (and the one used here) is to obtain the values for the areas from proportions taken from a table of the probability integral.

---

in the way just indicated. But these values could not then be legitimately summated to yield a composite value of χ^2 with 6 *df*. They could not be combined because they would lack independence; the frequency of non-occurrence in each calculation would include the frequency of occurrence of the other five faces.

[15] For an example of this method, see Guilford (**13**, p. 91).

The column in Table IX labeled X' gives the upper limits of the various score categories. Deviate scores and z scores based on the values of X' are

TABLE VIII

ANALYSIS OF COIN-GUESSING RESPONSES OF 439 SUBJECTS*

Patterns	F_o	F_t	$\frac{(F_o - F_t)^2}{F_t}$
H H H H H	25	(13.8)	9.089
H H H H T	12	13.7	0.211
H H H T H	22	13.7	5.028
H H H T T	18	13.7	1.350
H H T H H	29	(13.8)	16.742
H H T H T	96	(13.8)	489.626
H H T T H	22	(13.8)	4.872
H H T T T	14	13.7	0.007
H T H H H	5	13.7	5.525
H T H H T	15	13.7	0.123
H T H T H	12	13.7	0.211
H T H T T	4	13.7	6.868
H T T H H	33	(13.8)	26.713
H T T H T	17	13.7	0.795
H T T T H	12	13.7	0.211
H T T T T	3	13.7	8.357
T H H H H	3	13.7	8.357
T H H H T	7	13.7	3.277
T H H T H	10	13.7	0.999
T H H T T	14	13.7	0.007
T H T H H	5	13.7	5.525
T H T H T	1	13.7	11.773
T H T T H	6	13.7	4.328
T H T T T	0	13.7	13.700
T T H H H	4	13.7	6.868
T T H H T	6	13.7	4.328
T T H T H	25	(13.8)	9.089
T T H T T	7	13.7	3.277
T T T H H	2	13.7	9.992
T T T H T	5	13.7	5.525
T T T T H	2	13.7	9.992
T T T T T	3	13.7	8.357
$\sum$	439	439.0	$\chi^2 = 681.122$

* All but nine of the subjects are the same as those whose guesses are analyzed in Table IV.

shown in the fourth and fifth columns. Proportions of the total area under the normal curve from $-\infty$ to z are given in the P column. Proportions of the

area in the segments corresponding to the various score intervals are shown in column P' and were obtained by taking the differences between the successive values of P. The theoretical frequencies came from multiplying the values of P' by N which is 486 in this case.

TABLE IX

APPLICATION OF THE CHI-SQUARE TEST IN EVALUATING THE FIT OF A NORMAL CURVE TO A SET OF FREQUENCY DATA

X	F_o	X'	x $(=X'-M)$	z $(=x/\sigma)$	P	P'	F_t $(=P'N)$	$\frac{(F_o-F_t)^2}{F_t}$
				$(+\infty)$	(1.0000)			
144.5	1					.0141	6.9	
		139.5	35.5	2.195	.9859			.926
134.5	22					.0437	21.2	
		129.5	25.5	1.574	.9422			
124.5	56					.1125	54.7	.031
		119.5	15.5	0.953	.8297			
114.5	112					.1960	95.3	2.926
		109.5	5.5	0.342	.6337			
104.5	111					.2440	118.5	.475
		99.5	− 4.5	−0.280	.3897			
94.5	94					.2032	98.8	.233
		89.5	−14.5	−0.891	.1865			
84.5	54					.1211	58.8	.392
		79.5	−24.5	−1.512	.0654			
74.5	27					.0489	23.8	
		69.5	−34.5	−2.133	.0165			
64.5	7					.0135	6.6	.555
		59.5	−44.5	−2.754	.0030			
54.5	2					.0030	1.4	
				$(-\infty)$	(.0000)			
$\sum$	486					1.0000	486.0	5.538

Because the first and the last two values of F_t are less than 10, they were combined with the adjacent values, as were the corresponding values of F_o. The sum of the last column in the table ($=5.538$) is the value of χ^2. Seven differences between F_o and F_t entered into the calculations. The number of degrees of freedom is $7-3=4$. Three degrees were lost because three restrictions were placed in determining the theoretical frequencies.[16] The restrictions were

[16] Some statistics texts (6, 25, 26) perpetuate the view, erroneously attributed to Pearson (24), that the number of restrictions imposed in fitting a normal curve is *1* or *3*, depending upon the hypothesis that the investigator wishes to test. There is only one hypothesis open to test—that the frequency data arose from a normal population. If

the computed values of $\sum F_o$, M, and σ. The hypothesis being tested is that the frequency data arose from a normal population. With 4 *df*, the probability of obtaining, by chance, a value of χ^2 greater than 5.538 is around .25; so the hypothesis is tenable.

The Poisson distribution function. If the probability of the occurrence of an event is quite small, so that Np remains small even though N is relatively large, the distribution of observed frequencies in samples of size N may be of the Poisson type. The equation for the Poisson distribution may be written

$$P(n) \doteq \frac{m^n}{n!} e^{-m} \qquad [11]^{17}$$

where $m = Np$ and e has its conventional meaning. As in equations [3] and [4], the symbol $P(n)$ represents the probability of n occurrences out of N possible occurrences. The symbol $\doteq$ is used in place of the equal sign to indicate that [11] is an approximation formula. The errors introduced by the approximation

TABLE X

ANALYSIS OF FREQUENCIES OF OCCURRENCE OF THE CONSONANT "TH" IN SAMPLES OF AMERICAN SPEECH

n	F_o	$P(n)$	F_t	$(F_o - F_t)^2$	$\frac{(F_o - F_t)^2}{F_t}$
0	31	.1868	22.42	73.61	3.283
1	31	.3138	37.66	44.35	1.178
2	30	.2636	31.63	2.66	.084
3	11	.1476	17.71	45.02	2.542
4	11	.0620	7.44		
5	3	.0209	2.51		
6	3	.0050	.60	41.22	3.896
. . . . .					
>6		.0003	.03		
$\sum$	120	1.0000	120.00		$\chi^2 = 10.983$

are negligible, provided that N is quite large and provided also that N is very much larger than the largest value of n that may reasonably be expected in random sampling.

The χ^2 test may be applied in relation to the data of Table X. The column labeled F_o gives the number of samples, in a total of 120 short samples of

the mean and standard deviation of the fitted function are estimated from the data, three restrictions are imposed and 3 *df* are lost. The same texts give a similar misinterpretation of the number of restrictions imposed when χ^2 is applied in testing independence.

[17] A derivation of this formula is given by Lewis (19, pp. 168–169).

American speech, that contained n occurrences of the consonant "th" (as in thin). Each of the 120 samples was 400 sounds in length. As seen from the table, 31 of the samples did not contain any "th" sounds; 31 samples contained one "th" sound each; and so on. As a first step, the data in Table X will be compared with some results obtained by Voelker (33). In a study of over 600,000 sounds occurring in almost 6,000 announcements over the radio, Voelker found the proportion of "th" sounds to be .0065. Each of the 120 samples represented in Table X contained 400 sounds. This made a total of 48,000 sounds. The use of the proportion obtained by Voelker leads to 260 as the predicted, or theoretical, number of "th" sounds among the 48,000. The observed number was 201.[18] The χ^2 test may be applied in evaluating the hypothesis that the sounds in the present over-all sample, were drawn from a general population of American speech sounds which is assumed to be characterized exactly by the value of m obtained by Voelker. The value of χ^2 is computed as follows:

$$\chi^2 = \frac{(201 - 260)^2}{260} + \frac{(47{,}799 - 47{,}740)^2}{47{,}740} = 13.388 + 0.073 = 13.461.$$

Note that the observed and theoretical frequencies of non-occurrence of the "th" enter into the calculation. With 1 *df*, the obtained value of χ^2 is significant at better than the 0.1% level of confidence; so there is a firm basis for rejecting the hypothesis.

The observed proportion of .0042 ($=201/48{,}000$) may be used in testing the hypothesis that the 120 samples were all drawn from the same Poisson distribution. If $p_o = .0042$ and $N = 400$, then $m = Np = 1.68$, and the equation for the hypothetical distribution function may be written

$$P(n) = \frac{(1.68)^n}{n!} e^{-1.68}.$$

Values of $P(n)$ computed with this formula are given in the third column of Table X, and the corresponding values of F_t are given in the fourth column. As shown in the last column of the table, the value of χ^2 was computed in accordance with equation [10]. The last 4 values of F_t and the last 3 values of F_o were combined to avoid the use of theoretical frequencies of less than 10. The computed value of χ^2 is 10.982. Five differences between F_o and F_t were used in the computations. Two restrictions (N and m)[19] were imposed in calculating values of F_t. This leaves 3 *df*. With this number of *df*, the obtained value χ^2 falls at about the 3% level of confidence. The hypothesis may, therefore, be tentatively retained or may be rejected, depending upon the level of confidence that has been prescribed.

Applications: II. The Chi-Square Test of Independence

A common application of the χ^2 test enables an examination of the frequencies of a contingency table to determine whether or not the

[18] $\sum nF_o = 0(31) + 1(31) + 2(30) + 3(11) + 4(11) + 5(3) + 6(3) = 201.$

[19] If the value of p had not been estimated from the empirical data, but had been taken from the Voelker study or from some other completely independent source, the only restriction that would have been placed would have been N, and there would have been 4 *df* instead of 3 *df*.

two variables or attributes represented in the table are independent. The number of cells in the table may range from four (as in a 2×2 table) to an indefinitely large value. The χ^2 test is perhaps most commonly applied by psychologists in relation to 2×2 tables. The chief weaknesses in such applications are (a) a strong tendency to use excessively small theoretical frequencies and (b) an occasional failure to categorize adequately. The same two weaknesses are apt to occur when the number of categories in either, or both, of the "dimensions" of the table is greater than two.

Illustrations of the Chi-Square Test of Independence

Comparison of coin- and die-coin guessing. To obtain data for illustrating the χ^2 test of independence, 384 students of psychology were each asked to guess heads or tails on five successive tosses of a coin, where the tosses were interspersed among five throws of a die. The die was thrown; a guess was made as to the face that turned up. This guess was written on one edge of a sheet of paper. The edge of the paper was then folded under, to hide the guess. The

TABLE XI

RESULTS FROM COIN- AND DIE-COIN GUESSING EXPERIMENTS

Patterns	F_o			F_t	
	Coin Guessing	*Die-Coin Guessing*	*Totals*	*Coin Guessing*	*Die-Coin Guessing*
H H H H H	25	29	54	28.8	25.2
H H H H T	12	9	21	11.2	9.8
H H H T H	22	9	31	16.5	14.5
H H H T T	18	5	23	12.3	10.7
	77	52			
H H T H H	29	38	67	35.7	31.3
H H T H T	96	59	155	82.7	72.3
H H T T H	22	24	46	24.5	21.5
H H T T T	14	8	22	11.7	10.3
	161	129			
H T H H H	5	7	12	6.4	5.6
H T H H T	15	9	24	12.8	11.2
H T H T H	12	2	14	7.5	6.5
H T H T T	4	9	13	6.9	6.1
	36	27			

TABLE XI—*Continued*

Patterns	F_o			F_t	
	Coin Guessing	*Die-Coin Guessing*	*Totals*	*Coin Guessing*	*Die-Coin Guessing*
H T T H H	33	12	45	24.0	21.0
H T T H T	17	20	37	19.7	17.3
H T T T H	12	3	15	8.0	7.0
H T T T T	3	4	7	3.7	3.3
	65	39			
Sub-totals	(339)	(247)			
T H H H H	3	3	6	3.2	2.8
T H H H T	7	5	12	6.4	5.6
T H H T H	10	9	19	10.1	8.9
T H H T T	14	3	17	9.1	7.9
	34	20			
T H T H H	5	4	9	4.8	4.2
T H T H T	1	8	9	4.8	4.2
T H T T H	6	4	10	5.3	4.7
T H T T T	0	1	1	0.5	0.5
	12	17			
T T H H H	4	13	17	9.1	7.9
T T H H T	6	19	25	13.3	11.7
T T H T H	25	37	62	33.1	28.9
T T H T T	7	13	20	10.7	9.3
	42	82			
T T T H H	2	4	6	3.2	2.8
T T T H T	5	5	10	5.3	4.7
T T T T H	2	1	3	1.6	1.4
T T T T T	3	8	11	5.9	5.1
	12	18			
Sub-totals	(100)	(137)			
Totals	439	384	823	438.8	384.2

coin was then tossed, the guess being written down. Again the paper was folded under, to hide the guess. The die was thrown a second time, the guess made, the paper folded under. The coin was then thrown a second time, the guess made, the paper folded under. Each of the five guesses on the coin was preceded by a throw of the die and a guess on its fall. The paper was folded under after each guess on the die and each guess on the coin. Thus, the guesses on the coin were not only separated by guesses on the die, but the sequence of guesses was hidden from view. The subjects were never informed as to how the coin or the die actually fell.

The frequencies of occurrence of the 32 possible patterns of guesses on the five successive tosses of the coin (with guesses on the die ignored) are given in the third column of Table XI. The frequencies in the second column of this table were copied directly from Table VIII and are based on guesses on five successive tosses of a coin alone.[20] The subjects for the two conditions of guessing were completely different.

TABLE XII

2×2 CONTINGENCY TABLE BASED UPON THE COIN- AND DIE-COIN GUESSING EXPERIMENTS

	First Guess		
	H	*T*	
Coin Guessing	339 (312.6)	100 (126.4)	439
Die-Coin Guessing	247 (273.4)	137 (110.6)	384
	586	237	823

The first 16 patterns listed in Table XI begin with a guess of heads, the last 16 with a guess of tails. A 2×2 contingency table, shown in Table XII, was set up, the division along one "dimension" being between first-guess-heads and first-guess-tails and along the other "dimension" between the two conditions of guessing. The four sub-totals in Table XI constitute the observed frequencies appearing in the four cells of the 2×2 table.

The hypothesis to be tested in this case is that the occurrence of a first guess of heads was independent of the condition under which the guessing was done. If the observed frequencies in Table XII were independent of the conditions of

[20] As will be seen, the 32 patterns in Table XI have been divided into groups of four patterns each. The basis for the division will be discussed later. Short horizontal lines divide the corresponding observed frequencies. A numeral is placed to the right of each of these lines. Each numeral is the sum of the observed frequencies for the corresponding group of four patterns. Each is a sub-sum and will be used in Table XIII.

guessing, the probability of a guess of heads was 586/823. The probability of a subject's being in the coin guessing group was 439/823. The joint probability that a subject would be in the coin guessing group and would also guess heads was $439/823 \times 586/823$. The theoretical frequency for the upper left-hand cell of the table was obtained by multiplying this joint probability by 823. The other three theoretical frequencies were automatically determined by this single calculated frequency and by the restrictions of the border sums. The four values of F_t are shown in parentheses in the table. The value of χ^2 was computed as follows:

$$\chi^2 = \frac{(339-312.6)^2}{312.6} + \frac{(100-126.4)^2}{126.4} + \frac{(247-273.4)^2}{273.4} + \frac{(137-110.6)^2}{110.6}$$

$$= (26.4)^2\left(\frac{1}{312.6} + \frac{1}{126.4} + \frac{1}{273.4} + \frac{1}{110.6}\right)$$

$$= 696.96\ (.02381) = 165.946.$$

The number of *df* is 1. This follows because there are four cells and because three restrictions were placed in determining the theoretical frequencies. The three restrictions were the total number of subjects (=823) and two border sums, one for a row and one for a column.[21] The calculated value of χ^2 is highly significant and leads immediately to a rejection of the hypothesis that the conditions of guessing had no influence on the tendency to guess heads on the first guess.

Table XI may be regarded as a 32×2 contingency table, 32 "patterns" by 2 "conditions of guessing." In order to use the χ^2 test in evaluating the hypothesis that the patterns were independent of the conditions of guessing, the border sums would be used in calculating theoretical frequencies. For example, the two theoretical frequencies for the pattern *H H H H H* are given by the relations: $F_t = (54 \times 439)/823 = 28.8$; and $F_t = 54 - 28.8 = 25.2 = (54 \times 384)/823$. These two values of F_t are shown in the top row of the last two columns of Table XI. The theoretical frequencies for the other 31 patterns were obtained in a similar way and are listed in the table. These frequencies are included to emphasize the fact that the χ^2 test cannot be legitimately applied to this table as it stands. The reason is that 37 of the 64 theoretical frequencies are less than 10 (some of them very much less than 10) and cannot be depended upon to yield quantities distributed as χ^2. To make a legitimate application of χ^2 in this particular case, it would be necessary to increase the number of subjects to a point where the smallest of the theoretical frequencies was close to 10.

It must now be decided whether or not the observed frequencies in Table XI can be combined so as to permit the use of the χ^2 test in evaluating the hypothesis of non-relationship between the patterning of the guesses and the conditions of guessing. The frequencies have already been combined in a gross way to yield the observed frequencies in Table XII. This division was on the basis of the first guess. Divisions might be made on the basis of the first two guesses or the first three guesses or the first four guesses. It is not possible to use the patterning on all five guesses because the theoretical frequencies become too small, as already seen. It turns out that a division on the basis of the first four

[21] The sums 439 and 237 (or 384 and 237, or 384 and 586) could have been used instead of 586 and 439. The values of F_t for all but one cell may be obtained by subtraction.

guesses also leads to several theoretical frequencies that are less than 10. Consequently, a division based on the patterning of the first three guesses will be illustrated. In the division that was made, the following rule held: there would be a decreasing number of heads in the pattern and, contrariwise, there would be an increasing number of tails, from the first guess on. The resulting division of the patterns is shown in Table XI and also in Table XIII, the one to be used in applying the χ^2 test.

TABLE XIII

COMBINATION OF THE FREQUENCIES SHOWN IN TABLE XI, BASED ON THE FIRST THREE GUESSES

Pattern on 1st Three Guesses	*Coin Guessing*	*Die-Coin Guessing*	*Totals*
H H H	77 (68.8)	52 (60.2)	129
H H T	161 (154.7)	129 (135.3)	290
H T H	36 (33.6)	27 (29.4)	63
H T T	65 (55.5)	39 (48.5)	104
T H H	34 (28.8)	20 (25.2)	54
T H T	12 (15.5)	17 (13.5)	29
T T H	42 (66.1)	82 (57.9)	124
T T T	12 (16.0)	18 (14.0)	30
Totals	439	384	823

As seen, there are eight different patterns listed in Table XIII. The observed frequencies for the two conditions of guessing are shown (together with parenthesized theoretical frequencies) in the second and third columns. The hypothesis to be tested is that the patterns of guessing on the first three of five consecutive guesses were independent of the conditions of guessing. The theoretical frequencies were secured in the usual way by employing border sums and the value of $N(=823)$. The value of χ^2 was computed as follows:

$$\chi^2 = \frac{(77 - 68.8)^2}{68.8} + \frac{(161 - 154.7)^2}{154.7} + \cdots + \frac{(82 - 57.9)^2}{57.9} + \frac{(18 - 14.0)^2}{14.0} = 31.162.$$

The number of *df* is 7. The computed value of χ^2 is significant at far better than the 1% level of confidence and leads to a rejection of the hypothesis. It may be confidently concluded that the patterning of the guesses through the first three guesses was somehow influenced by the conditions under which the guessing was done.

TABLE XIV

5×4 CONTINGENCY TABLE BASED UPON LINDQUIST'S DATA (23)

Scores	*Enrollment Groups**				*Totals*
	A	*B*	*C*	*D*	
1	36	40	20	3	99
	(23.5)	(38.4)	(28.2)	(8.9)	
2	76	108	59	11	254
	(60.2)	(96.8)	(72.3)	(22.9)	
3	150	181	111	28	470
	(111.4)	(182.5)	(133.7)	(42.4)	
4	211	342	285	88	926
	(219.5)	(359.6)	(263.5)	(83.5)	
5	66	212	172	75	525
	(124.4)	(203.8)	(149.4)	(47.3)	
Totals	539	883	647	205	2274

* See footnote 22.

The number of *df* for a value of χ^2 obtained from a contingency table is always the number of cells in the table minus the number of restrictions imposed during the calculation of the theoretical frequencies. In Table XIII, for example, there are 16 cells. Nine restrictions must be imposed in obtaining values of F_t. These restrictions are: 7 of the row sums, 1 of the column sums, and the total number of cases. Thus, $df = 16 - 9 = 7$.

A convenient formula for determining the number of *df* for a contingency table when the χ^2 test is applied is

$$df = (n_c - 1)(n_r - 1), \qquad [12]$$

where n_c and n_r are the number of columns and the number of rows, respectively. There is only one hypothesis to be tested—that the variables are independent in the population from which the samples arise; so the number of *df* is always given by [12]. (See footnote 16.)

Contingency table with more than two categories in each direction. For the sake of completeness, a contingency table having five categories in one direction

and four in the other will be included. A total of 2,274 eighth-grade pupils, enrolled in 91 different schools, took an English Correctness Test. A summary of the scores obtained by these pupils has been taken from a report by Lindquist (23). The scores were divided into five categories and symbolized by numbers from 1 to 5, as seen in Table XIV. The schools were divided into four enrollment groups, labeled A to D in the table. The observed frequencies in the 20 cells of the table range from 3 to 342. These frequencies would obviously have been different if the enrollment groups had been differently established, and if the scores had been divided into different categories. The enrollment grouping of the schools was that commonly used in the Iowa Every-Pupil Testing Program.[22] The division of the scores was made by starting at the bottom and "stepping off" successive standard deviation "distances" (approximately). The distribution of scores was positively skewed; so it was necessary to combine the two upper score categories to provide satisfactorily large frequencies in the top row of cells.

The theoretical frequencies for Table XIV were computed in the usual way when the hypothesis of independence is under test. The computations required the use of three column sums and four row sums, as well as the value of $N(=2274)$. This made a total of 8 restrictions; so the number of $df = 20 - 8 = 12 = (5-1)(4-1)$. The value of χ^2 was computed as follows:

$$\chi^2 = \frac{(36-23.5)^2}{23.5} + \frac{(76-60.2)^2}{60.2} + \cdots + \frac{(88-83.5)^2}{83.5} + \frac{(75-47.3)^2}{47.3} = 99.038.$$

With $df = 12$, this value is highly significant and leads to a rejection of the hypothesis that the scores obtained on the test were independent of school size.

A comment should be made concerning the theoretical frequency in the upper right-hand cell of the table. Its value is 8.9. Ordinarily, a value this small should not be used in obtaining an estimate of χ^2. In this case, however, the other 19 values of F_t are satisfactorily large, and the inclusion of one theoretical frequency that is less than 10 is permissible since an error in a single category will have slight effect on the resulting value of χ^2. The obtained value of χ^2 is so large that it makes no difference whether or not the small theoretical frequency is included in the calculations. It is only in situations of this general kind that one or two small theoretical frequencies may be retained. When the number of df is less than 4 or 5, and especially when $df = 1$, the use of theoretical frequencies of less than 10 should be strictly avoided.

Use of the Chi-square Test with too Small Theoretical Frequencies

The studies of Lewis and Franklin (21) *and Lewis* (20). The commonest weakness in applications of the χ^2 test to contingency tables is the use of extremely small theoretical frequencies. This weakness is clearly present in most of the applications made in a paper by Lewis and Franklin (21). The paper is concerned with the Zeigarnik effect (that is, with the relative amounts of recall of interrupted and completed tasks). In one experiment, 12 subjects were each presented with 18 problems, 9 of which were interrupted by the experimenter, the other 9 being completed without interruption. The ratio (RI/RC) of the

[22] The enrollment categories were: A, greater than 400; B, 126–400; C, 66–125; D, less than 66.

number of interrupted tasks recalled to the number of completed tasks recalled is given for each subject in the top row of Table XV. In a previous study,

TABLE XV

DATA UPON THE "ZEIGARNIK EFFECT" PRESENTED BY LEWIS (20), AND LEWIS AND FRANKLIN (21)

Ratio of Recall Scores	*Subjects*													
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
RI/RC—Group I (Indiv. Work Condition)	1.00	.80	.80	.80	.75	.71	.63	.57	.57	.44	.40	.25		
RI/RC—Group II (Coop. Work Condition)	1.67	1.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.80	.75	.60	.14

Lewis (20) had employed a cooperative work situation in which a co-worker completed the tasks on which the subject was interrupted. The conditions were otherwise the same as in the later experiment by her and Franklin. Fourteen subjects were used. The *RI/RC* ratio for each subject is shown in the second row of Table XV. The median ratio for Group I was .67. The ratios for the two groups were divided (dichotomized) at this point. The 2×2 contingency table shown as Table XVI was the result. The hypothesis to be tested was that the conditions of the two experiments had no differential effects on the recall of interrupted and completed tasks. The theoretical frequencies are shown in parentheses in the table. The investigators calculated a value of χ^2 in the usual way. Finding it to be approximately 3.8 and to fall near the 5% level of confidence, they were inclined to reject the hypothesis.

Entirely aside from any of the conclusions reached by Lewis and Franklin, it must be firmly stated that all four of the theoretical frequencies in Table XVI, and especially the two that are less than 5, are too small to warrant an application of the χ^2 test. Furthermore, the other applications made in their paper, with one or two possible exceptions, involved theoretical frequencies that should be avoided (theoretical frequencies, for example, of 4.4, 6.0, 7.0, 7.6, etc.).

TABLE XVI

CONTINGENCY TABLE BASED UPON DATA IN TABLE XV

RI/RC	*Group*		
	I	II	
Greater than 0.67	6 (8.3)	12 (9.7)	18
Less than 0.67	6 (3.7)	2 (4.3)	8
	12	14	26

Kuenne's study of transposition behavior. An application, which is correct in principle but which must be regarded somewhat unfavorably because of the size of the F_t's, comes from a paper by Kuenne (18). Kuenne made a study of transposition behavior in four groups of young children, the ages for the groups being 3, 4, 5, and 6 years. Some of the children displayed size transposition, others of them did not. Kuenne realized that, because of the small number of cases in each age group, she could not apply the χ^2 test to the data for the four groups considered separately. Consequently, she combined the results for ages 3 and 4, and those for ages 5 and 6. The children were divided into two categories—those who did and those who did not meet the transposition criterion. The resulting 2×2 contingency table is shown as Table XVII. The hypothesis

TABLE XVII

CONTINGENCY TABLE BASED UPON KUENNE'S (18) TRANSPOSITION DATA

Age-Groups	*Transposition*	*Non-Transposition*	*Total*
3–4 years	3	15	18
	(9.4)	(8.6)	
5–6 years	20	6	26
	(13.6)	(12.4)	
	23	21	44

to be tested is that the occurrence of transposition behavior was independent of age. Theoretical frequencies were determined by using two of the border sums and the value of N. It will be seen that two of these frequencies are less than 10. Because they are fairly close to 10, many investigators would proceed as Kuenne did, and make a χ^2 test of the hypothesis of independence. In fact, there are textbooks in statistics which place 5 as the minimum value for theoretical frequencies. A value of 5 is believed to be too low. In any event, it is the smallest value that should be used even when there are several other theoretical frequencies that are far greater than 10.

The value of χ^2 computed from Table XVII is 15.434. With 1 *df*, this value falls close to the 0.1% level. It is only because the value is so large that confidence can be placed in the conclusion that transposition behavior was related to age. In view of the smallness of all four of the theoretical frequencies, very great doubt would have remained if the χ^2 value had fallen at a border-line level of confidence. Whenever small theoretical frequencies enter into calculations of χ^2, the experimenter has no sound basis either for accepting or rejecting a hypothesis except when the value is quite extreme.

Yates' correction for continuity. No mention has as yet been made of a correction proposed by Yates (30) which reduces the value of χ^2 to compensate for errors which may arise as a result of one of the approximations made in deriving the formula for the χ^2 distribution. It will be recalled that three approximations are made in this derivation. One of the three involves the substitution of an

integral for a summation of discrete quantities. This approximation introduces an error (an error of overestimation) that is of consequence when values of F_t are small. The correction is justified only when the number of *df* is 1.[23] It provides for the reduction of all differences between observed and theoretical frequencies by 0.5. For example, all of the differences between observed and theoretical frequencies in Table XVII are 6.4. These are reduced to 5.9 if Yates' correction is applied. The calculation, using the correction, would be as follows:

$$\chi^2 = (5.9)^2\left(\frac{1}{9.4}+\frac{1}{8.6}+\frac{1}{13.6}+\frac{1}{12.4}\right) = (34.81)(.3768) = 13.1.$$

With 1 *df*, this value is still highly significant and leads to a rejection of the hypothesis. But the use of Yates' correction does not remove the objection to theoretical frequencies that are less than 10.

Questionable or Incorrect Categorizing

Lewis' (20) *study of recall of interrupted and completed tasks.* Another weakness which is sometimes present in applications of the χ^2 test to contingency tables is that the categorizing is done on either a questionable or a clearly incorrect basis. An illustration of incorrect categorizing is found in the paper by Lewis (20) discussed above. The *RI/RC* ratios for 14 subjects were obtained in a "cooperative work experiment." The ratios are the ones given in the second row of Table XV. On the assumption that the recall of interrupted and completed tasks should have been the same, Lewis writes: " . . . we should have an equal number of ratios above 1.00 and below 1.00. . . . The expected distribution of ratios should, therefore, be 7 below 1.00 and 7 at 1.00 or above. The obtained distribution of ratios is 4 below 1.00 and 10 at 1.00 or above." The categorizing is plainly wrong; there is no more reason for placing an obtained ratio of 1.00 in the upper category than for placing it in the lower category. A better procedure would have been to divide the 8 ratios of 1.00 equally between the two categories. This flaw in Lewis' division of the ratios is remindful of the belief of many graduate students in psychology that it is quite permissible to set up several different sets of dichotomy lines, compute a value of χ^2 for each set, and finally select the dichotomies that yield a χ^2 value to support the experimenter's own point of view. In any investigation where the χ^2 test is to be applied, the categories must be established in a logically defensible and reliable manner—before the data are collected, if possible.

Anastasi and Foley's (1) *study of drawings of normal and abnormal subjects.* The whole problem of categorizing may be brought clearly before the reader by taking an illustrative case from a study by Anastasi and Foley (1). These two investigators had each of 340 normal subjects and 340 abnormal subjects draw a picture which depicted danger. The pictures were then divided into the 20 subject-matter categories listed in Table XVIII. The application of the χ^2 test yields a value of 99.603 which, with 19 *df*, falls far beyond the 0.1% level of confidence. This leads to a rejection of the hypothesis that the subject matter of the drawings was independent of the two "kinds" of subjects.

[23] The correction should not be made if several values of χ^2 are to be summated. The additive principle does not apply to corrected values.

TABLE XVIII

DATA FROM ANASTASI AND FOLEY'S (1) STUDY OF DRAWINGS OF NORMAL AND ABNORMAL SUBJECTS

Subject-Matter Categories	F_o (*Abnormal*)	F_o (*Normal*)	F_t	χ^2
1. Traffic	105	124	114.5	1.576
2. Conventional sign or signal	36	37	36.5	.014
3. Skating, ice	5	10	7.5	1.667
4. Falling	21	17	19.0	.421
5. Drowning, sinking, flood	8	10	9.0	.222
6. Falling objects, explosion	6	6	6.0	.000
7. Arms and explosives	10	22	16.0	4.500
8. War	3	7	5.0	1.600
9. Fire	26	37	31.5	1.921
10. Lightning, electricity	6	15	10.5	3.857
11. Animals	3	16	9.5	8.895
12. Abstract or conventionalized symbolism	3	8	5.5	2.273
13. Fantastic compositions	26	0	13.0	26.000
14. Several discrete objects	9	2	5.5	4.454
15. Scribbling or scrawl	6	0	3.0	6.000
16. Writing only	12	1	6.5	9.308
17. Miscellaneous	23	26	24.5	.184
18. Recognizable object not representing danger	8	1	4.5	5.444
19. Refusal to draw	14	1	7.5	11.267
20. No data for other reasons	10	0	5.0	10.000
	340	340		99.603

Let it again be emphasized that the criticism here, as elsewhere in the paper, is not directed at any of the conclusions reached by the investigators. But the illustration provides a very satisfactory basis for discussing the fundamental problem of categorizing. In the published article, the principles adopted in classifying the pictures are not explicitly stated. Furthermore, evidence is not presented regarding the reliability of the categories. Two generalizations may be offered. The first is that, *whenever possible, categories for frequency data should be established on the basis of completely external criteria* (for example, criteria that have been used or proposed by some other investigator) and should be set up independently of the data under study. Such a procedure frees a person from any charge of bias and guards against tendencies to juggle data. A second generalization is that *information on the reliability of categories should be offered,* and this is the case whether or not the categories have stemmed from an independent source.

A study of Table XVIII shows rather quickly that the value of χ^2 for the drawings depicting danger would have been quite different if the categorizing

had been different. For example, the amount 26.000 was contributed to the value of χ^2 by the frequencies of category 13 alone. As seen, this is the category "fantastic compositions." The decision which established this category and the judgments which placed 26 of the drawings of the abnormal subjects in the category and none of the drawings of the normal subjects in it, should have been explicitly justified and a precise statement concerning reliability should have been included. The discrepancy between the frequencies in category 13 (along with the discrepancies in such categories as "refusal to draw" and "no data for other reasons") required that there be discrepancies in one or more of the other categories. Unreliability at one point in a multicelled table automatically produces unreliability elsewhere.

It is well to emphasize, by reiteration, that when the χ^2 test is to be applied to a collection of data, the categories should be established independently of the data and, once established, should never be modified on the basis of the way the data happen to fall. Categories should usually, if not always, be established before the data have been scrutinized.

Applications: III. The Goodness of Fit of Functions in Which Frequency Is the Dependent Variable

The frequency (or relative frequency) of occurrence of a response is sometimes used as the dependent variable in psychological experiments. For example, in psychophysical investigations based on the method of constant stimuli, the number or proportion of judgments in a given direction serves as the dependent variable, while in studies of the conditioned response, the frequency of occurrence of the CR is often taken as the dependent variable. If a mathematical function is fitted to data of this type, it is sometimes possible to apply the χ^2 test in evaluating the goodness of the fit. However, care must be taken to insure a correct application.

Goodness of fit of the phi-gamma function. It is a fairly common practice among psychophysicists to apply the χ^2 test to the differences between observed and theoretical proportions in order to evaluate the phi-gamma hypothesis—the hypothesis that the observed proportions can be represented satisfactorily by the phi-gamma function.[24] The usual procedure in making the test will be illustrated with some weight-lifting data.

As part of a laboratory exercise, a graduate student in an advanced experimental course made 100 judgments on each of nine pairs of weights. The weights, in grams, are listed in the X column of Table XIX. The method of constant stimuli was employed. The 100-gram weight was the standard and was paired 100 times, not only with itself, but with *each* of the other eight ("variable") weights. The standard weight was always lifted first, the variable weight second. Each of the 900 judgments was made in terms of the question: Which weight is the heavier, the first or the second? The second column of the

[24] The use of the phi-gamma function in psychophysical research is explained by Guilford (13, Chap. VI).

TABLE XIX

TEST OF GOODNESS OF FIT OF THE PHI-GAMMA FUNCTION TO WEIGHT-LIFTING DATA

X	p_o	γ	$X\gamma$	X^2	γ_t	p_t	(A) $(p_o-p_t)^2$	(A) $\frac{(p_o-p_t)^2}{p_tq_t}$	(B) p_o	(B) p_t	(B) $n_k\frac{(p_o-p_t)^2}{p_tq_t}$
104	.97	1.3299	138.31	10,816	1.246	.960	.000100	.002600	.935	.934	.0032
102	.90	0.9062	92.43	10,404	.942	.908	.000064	.000768			
100	.79	0.5702	57.02	10,000	.638	.816	.000676	.004516	.79	.816	.4516
98	.65	0.2724	26.70	9,604	.334	.681	.000961	.004421	.65	.681	.4421
96	.53	0.0532	5.11	9,216	.030	.517	.000169	.000676	.53	.517	.0676
94	.37	−0.2347	− 22.06	8,836	− .274	.350	.000400	.001760	.37	.350	.1760
92	.20	−0.5951	− 54.75	8,464	− .578	.208	.000064	.000384	.20	.208	.0384
90	.13	−0.7965	− 71.68	8,100	− .882	.107	.000529	.005539	.085	.077	.1804
88	.04	−1.2379	−108.94	7,744	−1.186	.047	.000049	.001103			
$\sum$ 864		0.2677	62.14	83,184				.021767			1.3593

table, labelled p_o, gives the proportion of times (in 100) that each weight was judged heavier than the standard weight.[25]

A phi-gamma function may be fitted to the proportions in Table XIX, and the χ^2 test may then be applied (with a reservation to be specified later) to evaluate the goodness of fit. The phi-gamma function is basically the same as the equation for the normal ogive, which may be written

$$p = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}\, dx, \qquad [13]$$

with p standing for proportion of area under the normal curve from $-\infty$ to x. As usual, x represents a deviation of the variate X from the mean of X. The substitution in equation [13] of the symbol h for the quantity $1/\sigma\sqrt{2}$ yields

$$p = \int_{-\infty}^{x} \frac{h}{\sqrt{\pi}} e^{-h^2x^2}\, dx. \qquad [14]$$

In this expression, $x = X - L$, where L is the value of X corresponding to a proportion of .50. The constant h is an index of the steepness or "precision" of the ogive curve. The phi-gamma function, written directly from [14], takes the form

$$p = \int_{-\infty}^{\gamma} \frac{1}{\sqrt{\pi}} e^{-\gamma^2}\, d\gamma \qquad [15]$$

where γ has been substituted for hx.

In fitting the phi-gamma function to a set of observed proportions, it is necessary to obtain estimates of the constants h and L. Since $\gamma = hx$ and $x = X - L$, it follows that

$$\gamma = hX - hL. \qquad [16]$$

This equation is linear in γ and X. The principle of least-squares is invoked in deriving the following two equations with which estimates of h and L may be computed:

$$h = \frac{N\sum X\gamma - \sum\gamma\sum X}{N\sum X^2 - (\sum X)^2}; \qquad [17]$$

$$L = \frac{\sum Xh - \sum\gamma}{Nh}. \qquad 18$$

In the derivation of these equations, the quantity that is minimized is

$$\sum(\gamma_o - \gamma_t)^2 = \sum(\gamma - hX + hL)^2.$$

Before equations [17] and [18] can be solved, values of γ must be obtained. They may be found in a table of the phi-gamma function (that is, in a table based on solutions of equation [15] for values of p).[26] The table is entered with values of p_o.

[25] The reason that the value of p_o for the 100-gram weight is significantly greater than .50 is that a negative time error was operating. This means that there was a consistent tendency for the subject to over-estimate the weight of the second member of a pair. The tendency was present in all comparisons, but was most immediately apparent from the value of p_o when the standard weight was compared with itself.

[26] Tables of the phi-gamma function (which give paired values of p and γ) are not usually included in statistics books published in the United States. However, it is easy

Values of γ corresponding to the nine values of p_o in Table XIX are given in the third column of the table. The fourth column gives values of $X\gamma$, the fifth column values of X^2. The four sums needed for solving equations [17] and [18] appear at the bottom of the table. Estimates of constants h and L are computed as follows:

$$h = \frac{9(62.14) - (.2677)(864)}{9(83,184) - (864)^2} = 0.152;$$

$$L = \frac{(864)(.152) - (.2677)}{9(.152)} = 95.8.$$

The insertion of these estimates into equation [16] gives

$$\gamma_t = 0.152X - 14.562.$$

This equation was used to calculate the values of γ_t shown in Table XIX. The corresponding values of p_t were read from a table of the phi-gamma function.

The χ^2 test may be applied (with a reservation) to test the goodness of fit, that is, to test the hypothesis that the difference between the observed and theoretical proportions arose from sampling fluctuations and, consequently, that the observed proportions may be satisfactorily represented by the phi-gamma function. The procedure for calculating the value of χ^2 based on all nine theoretical proportions is illustrated in the two columns of Table XIX denoted by (A). Values of $(p_o - p_t)^2$ were first obtained and were then divided by the product $p_t q_t$. This single division by $p_t q_t$ took account of differences between p_o and p_t as well as differences between q_o and q_t (where, as usual, $q_o = 1 - p_o$ and $q_t = 1 - p_t$). It can easily be shown that

$$\frac{(p_o - p_t)^2}{p_t} + \frac{(q_o - q_t)^2}{q_t} = \frac{(p_o - p_t)^2}{p_t q_t}.$$

Emphasis is here given again to the fact that in the calculation of χ^2, account must be taken of the frequency of occurrence *and* the frequency of non-occurrence of an event.

The sum of the second column under (A) is .021767. This is not the value of χ^2. However, it becomes the value of χ^2 after it has been weighted by 100, the number of judgments upon which each value of p_o was based. In other words,

$$\chi^2 = n_k \sum \frac{(p_o - p_t)^2}{p_t q_t} = 100\ (.021767) = 2.1767$$

where n_k is the number of judgments per proportion. The number of df is $7 = 9 - 2$. It is two less than the number of rows in the table.[27] The computed

---

to obtain values of γ indirectly from an ordinary table of the probability integral. Since $z = x/\sigma$; $\gamma = hx$, and $h = 1/\sigma\sqrt{2}$, it follows that $\gamma = .7071z$ and $z = 1.4142\gamma$.

[27] The number of degrees of freedom for a table such as XIX is readily determined by starting with the number of rows. Each row contributes 2 df before any restrictions are imposed. (This is because each row actually covers two pairs of proportions. The pair which is omitted in the formula is q_o and q_t.) The restriction that $(p_t + q_t) = 1.00$ is imposed on *each* row; so 1 df is lost for each. Further, 1 df is lost for each constant estimated

value of χ^2 falls near the 95% level of confidence; so the phi-gamma hypothesis cannot be rejected.

The applicability of the last statement rests upon the validity of applying the χ^2 test in the manner illustrated. There are two weaknesses in the application, one relating to the size of two of the theoretical frequencies and the other relating to the restrictions imposed during curve fitting. Before the frequencies are considered, a few comments will be made on the nature of the restrictions that were imposed, although it must be said that a meaningful explanation of restrictions is mathematically rather complex and lies beyond the scope of this paper. Whenever a curve is fitted to a set of empirical data, there are usually several acceptable methods of determining values for the constants to be inserted in the equation. It is relatively easy to show that some of these methods lead to values of "χ^2," as computed by the familiar formula, which tend to be consistently larger or smaller than the values obtained when other curve fitting methods are employed. Therefore, when a value of χ^2 is computed to test the goodness of fit of such functions as the phi-gamma function, the method used to estimate the constants becomes a matter of importance. The computed quantity is distributed as χ^2 only when the restrictions imposed during curve fitting are both *linear* and *homogeneous*. (The meaning of these terms is explained more fully in a later section.) In many instances of curve fitting, and the present one is a case in point, the restrictions that are imposed do not have the requisite properties. When this is true, the investigator must be especially cautious in interpreting his results.

Some consideration must now be given to the theoretical frequencies represented by the theoretical proportions in Table XIX. The proportion .960 in the first row corresponds to two theoretical frequencies: 96.0 and 4.0. These values are estimates of the number of judgments "heavier" and the number of judgments "lighter" that the subject should have made when comparing the 104-gram weight with the standard weight. The proportion .047 in the bottom row also corresponds to theoretical frequencies of 4.7 and 95.3. These frequencies indicate that, according to the fitted function, the subject should have said "heavier" 4.7 times and "lighter" 95.3 times when comparing the 88-gram weight with the standard. The two frequencies of 4.0 and 4.7 are too small to yield quantities distributed as χ^2 and should not have been used in the calculations. They were included in the illustration as a means of contrasting the correct with the incorrect procedure. The two extreme proportions should have been combined with adjacent proportions. Table XIX shows the results after the combining has been done. New values of p_o, based on .97 and .90, and also on .13 and .04, are shown in the column labeled p_o under the general heading (B). Corresponding theoretical proportions are given in the column labeled p_t. The two combined values of p_o (.935 and .085) are averages based on 200 judgments each and the quantity $(p_o - p_t)^2/p_t q_t$ calculated for each must be weighted by 200 instead of 100. Because of this differential weighting, it is better to multiply by n_k, as shown in the last column of the table, before the

---

from the data during the curve fitting process. Since two constants (h and L) must be estimated, the number of df is 2 less than the number of rows (or the number of pairs of p_o and p_t) that enter into the calculations.

TABLE XX

DATA FROM THE PITCH DISCRIMINATION STUDY BY STEVENS, MORGAN AND VOLKMANN (32)

X	*Subjects*												
	JV		*GS*		*RR*		*JM*		*DM*			*MJ**	
	p_o	p_t	p_o	p_t	p_o	p_t	p_o	p_t	p_o	p_t		p_o	p_t
1.0							.00	—			(0.75)	.00	—
1.5	.00	—	.01	—			.00	—	.00	—	(1.00)	.00	—
2.0	.00	—	.00	.036	.01	—	.10	.088	.00	—	(1.25)	.13	.143
2.5	.03	.044	.18	.172	.01	—	.29	.273	.06	.089	(1.50)	.40	.365
3.0	.18	.136	.45	.464	.02	.093	.49	.560	.30	.233	(1.75)	.62	.647
3.5	.31	.310	.78	.778	.21	.234	.84	.816	.44	.457	(2.00)	.87	.864
4.0	.49	.542	.95	.947	.52	.450	.96	.950	.64	.695	(2.25)	.99	.966
4.5	.74	.761	1.00	—	.64	.682	.98	—	.90	.870	(2.50)	1.00	—
5.0	.94	.905	1.00	—	.84	.858	.99	—	1.00	—			
5.5	1.00	.972			.97	.953	1.00	—	1.00	—			
6.0	1.00	—			.99	—							
6.5					1.00	—							

* The frequency increments presented to subject MJ are given in parentheses at the left of the values of p_o in the last column. This subject was unusually acute and had to be tested with smaller increments than those used with the other five subjects.

sum of the column is taken. When this is done, the sum is the estimate of χ^2. In other words,

$$\chi^2 = \sum n_k \frac{(p_o - p_t)^2}{p_t q_t} = 1.3593.$$

The number of *df* is now 5, and the calculated value of χ^2 falls between the 98% and 99% levels. The phi-gamma hypothesis is still highly tenable.

This last calculation included two theoretical frequencies which are less than 10. They are 7.7 and 6.6. Further combining has not been carried out because such a step would, in effect, eliminate values of p which are critical in making a test of the phi-gamma function. There is perhaps some justification for retaining two theoretical frequencies as small as 7.7 and 6.6 when there are 12 other theoretical frequencies ranging from 18.4 to 93.4. An investigator may choose to be somewhat lenient in this regard, but leniency should never lead to the inclusion of frequencies of less than 5 or 6, and under no circumstances to a frequency as low as 0.5, one that is retained in an example in a widely used text (13, p. 181).

It is deemed advisable, except under unusual circumstances, to adhere to the policy of never using theoretical frequencies of less than 10. This means that if the χ^2 test is to be suitably applied in testing the phi-gamma hypothesis, the number of judgments at extreme values of the variable stimulus must be several hundred.

It is sometimes a temptation to pool the separate proportions obtained for several subjects in a psychophysical experiment. This practice should never be followed. The reason is clear; individual differences in judgments would yield interdependent proportions. It is permissible to apply the χ^2 test in relation to frequencies (or proportions) obtained from the judgments of a group of observers provided that each subject makes a single judgment. For example, a series of proportions for several different values of a variable stimulus, where no person makes more than a single judgment in the entire experiment, may be fitted with a phi-gamma function to obtain theoretical proportions; and the χ^2 test may be used to test the goodness of fit of the function. It is obvious, however that this situation is entirely different from one where the proportions are the averages of several sets of individual proportions.

Evaluation of an extended application of the χ^2 test to psychophysical data made by Stevens, Morgan and Volkmann. An interesting and instructive application of the χ^2 test of goodness of fit was made by Stevens, Morgan, and Volkmann (32) in their theoretical study of pitch and loudness discrimination. A representative segment of their data will be discussed. In one part of the investigation, they presented each of six subjects with an auditory stimulus which continued over a prolonged period but which was quickly changed in frequency, every three seconds, by a predetermined number of cycles. The duration of each altered segment was 0.3 sec. A subject was instructed to press a button every time a change in pitch was detected. The frequency increments that were used (in cycles per sec.) are listed in the first column of Table XX (except as explained in the note at the bottom of the table). Each subject made 100 judgments in relation to each of the increments which fell within his discrimination range. The observed proportions for the six subjects are shown in the table. (These proportions were furnished through the courtesy of Dr. Morgan.)

The study was designed to evaluate the theory of the neural quantum in auditory discrimination. The theory required, among other things, that the observed proportions for a subject could not be adequately represented by the phi-gamma function.[28] Consequently, phi-gamma functions were fitted to each of the six sets of proportions in Table XX. All proportions greater than .97 and less than .03 were omitted during the curve fitting process. A few of these proportions later played a part in the calculations of χ^2. This is one of the pitfalls to be carefully avoided. Another important aspect of the curve-fitting procedure was that the Müller-Urban weights were utilized. The nature of these weights and the detailed procedures for treating weighted data are adequately explained by Guilford (13) and need not be discussed here. It is enough to say that the weights are designed to diminish the influence of extreme proportions in the computations of the constants h and L.

TABLE XXI

VALUES OF h AND L SECURED BY STEVENS, MORGAN AND VOLKMANN (32)

Subject	h	L
JV	0.8520	3.9120
GS	1.2098	3.0525
RR	0.8476	4.1051
JM	1.0640	2.9001
DM	0.8754	3.5885
MJ	2.0432	1.6194

The values of h and L for the six fitted functions are shown in Table XXI. These constants were used in the manner illustrated in Table XIX to obtain theoretical proportions for each set of observed proportions. The theoretical proportions used by the investigators in calculating values of χ^2 are shown in Table XX. The six calculated values of χ^2 are given in the first row of Table

[28] The theory also required that the proportions for the several subjects could be fitted satisfactorily with straight-line functions. The results for the straight-line fits will be omitted here. It is obvious, however, that a straight line can be fitted to each set of proportions in Table XX as a means of obtaining theoretical (calculated) proportions. The χ^2 test could then be applied to the differences between the observed and theoretical proportions, in a manner exactly analogous to that employed when theoretical proportions are obtained with phi-gamma functions. Whenever a comparison is to be made between the goodness of fit of two different functions, the same quantity should be minimized in the process of obtaining the constants for the functions. In their study, Stevens, Morgan, and Volkmann minimized the sum of squared differences between observed and theoretical values of gamma in fitting phi-gamma functions but minimized the sum of squared differences between observed and theoretical proportions in fitting straight-line functions. According to a theorem given by Cramér (5, p. 426 ff.), the values of χ^2 obtained with the phi-gamma functions would be expected to be larger than those obtained with the straight lines.

XXII, along with the number of *df* for each. (The other two rows of values in this table will be explained later.) Only one of the values in the first row (the value for subject *RR*) is sufficiently large to justify the rejection of the phi-gamma hypothesis at better than the 5% level of confidence. However, the six separate values of χ^2 are completely independent estimates and, therefore, may be summated to provide a single composite value of χ^2, the number of *df* for which is the sum of the six separate *df*'s. The composite value appears in the

TABLE XXII

VALUES OF CHI-SQUARE BASED UPON THE DATA OF STEVENS, MORGAN AND VOLKMANN (32)

	Subjects						
	JV	*GS*	*RR*	*JM*	*DM*	*MJ*	*Composite*
χ^2	7.64	4.02	10.35	2.97	5.90	2.92	33.80
df	5	3	4	3	3	3	21
χ^2	4.94	0.13	4.06	2.97	5.90	1.02	19.02
df	4	2	3	3	3	2	17
χ^2	1.70	0.13	3.14	2.87	2.90	1.02	11.76
df	2	1	2	1	2	2	10

TABLE XXIII

SHOWING THE EFFECTS UPON CHI-SQUARE VALUES OF USING OBSERVED PROPORTIONS WHICH WERE EXCLUDED IN THE DERIVATION OF THEORETICAL PROPORTIONS (DATA FROM TABLE XX)

				(*A*)	(*B*)			(*C*)
X	p_o	p_t	$(p_o-p_t)^2$	$\frac{(p_o-p_t)^2}{p_tq_t}$	$\frac{(p_o-p_t)^2}{p_tq_t}$	p_o	p_t	$n_k\frac{(p_o-p_t)^2}{p_tq_t}$
3.0	(.02)	.093	.005329	.0629	—	—	—	—
3.5	.21	.234	.000576	.0032	.0032	.21	.234	.32
4.0	.52	.450	.004900	.0201	.0201	.52	.450	2.01
4.5	.64	.682	.001764	.0081	.0081	.64	.682	.81
5.0	.84	.858	.000324	.0027	.0027	.905	.905	.00
5.5	.97	.953	.000289	.0065	.0065			
				.1035	.0406			3.14

From Column A: $\chi^2 = 100\ (.1035) = 10.35$, with 4 *df*
From Column B: $\chi^2 = 100\ (.0406) = 4.06$, with 3 *df*
From Column C: $\chi^2 = 3.14$, with 2 *df*.

last column of the table and is 33.80, with 21 *df*. This value falls at about the 3% level of confidence and indicates that the fits of the phi-gamma curves were generally poor.

A serious flaw in this application was that four of the extreme observed proportions, no one of which played any part in determining constants h and L, were included during the calculations of the values of χ^2. A specific example will clarify the point. The observed and theoretical proportions for subject *RR* are listed in the second and third columns of Table XXIII. The observed proportion .02, shown in parentheses, was purposely excluded by the investigators when making least-squares solutions for h and L. Once obtained, these constants were crucial in determining the theoretical proportions. The investigators decided to include all observed proportions whose corresponding theoretical proportions were neither greater than .97 nor less than .03. This meant including the observed proportion .02 in the calculations for Table XXIII. It also meant the inclusion of observed proportions 1.00, .00, and .99 in the calculations for subjects JV, GS, and MJ, respectively.

As shown in Table XXIII, the estimated value of χ^2 was 10.35 when $p_o = .02$ and $p_t = .093$ were included in the calculation. If this pair of proportions is not included, the estimated value is 4.06. The number of *df* for the first estimate is 4, for the second, 3. There can be no doubt that the second estimate is the better of the two. In the calculation of χ^2, it is incorrect to use observed proportions which have not been allowed to influence the magnitude of the theoretical proportions.

The values of χ^2 shown in the second row of Table XXII were obtained by excluding the observed proportions which played no role in the curve-fitting process. There are still six independent values, each with its own number of *df*. It is entirely legitimate to add these values to obtain the composite value of 19.02 (with 17 *df*) given in the last column of the table. This new composite value falls at about the 30% level and provides no basis for rejecting the phi-gamma hypothesis for the six sets of proportions.

Still another flaw in the calculations of the values in Table XXII was that theoretical proportions representing theoretical frequencies of less than 10 were not first combined with adjacent theoretical proportions. For example, the theoretical proportion .953 (in Table XXIII) represents theoretical frequencies of 95.3 and 4.7. A theoretical frequency of 4.7 is too small to yield a quantity distributed as χ^2. Therefore, theoretical proportion .953 should have been combined, as shown in Table XXIII, with the theoretical proportion .858 to give $p_o = .905$. The modified value of χ^2 is 3.14, with 2 *df*. This value appears for subject RR in the third row of Table XXII. Similarly modified values for three other subjects are included in the same row. Again there are six independent estimates of χ^2, and these may be added to obtain the single composite value of 11.76, listed in the last column of the table. With 10 *df*, this estimate falls around the 30% level.

It will be realized that the elimination of the small and large proportions makes impossible a really critical evaluation of the phi-gamma hypothesis through the use of the χ^2 test. But this is not a fault of χ^2; it is a weakness in the experimental data. Whenever χ^2 is to be employed, the experimenter must take precautions to insure theoretical frequencies of adequate size.

The principal pitfalls in the use of χ^2 with proportions have been designated.

Two others should be mentioned—the tendency to divide $(p_o - p_t)^2$ by p_t alone instead of by the product $p_t q_t$ without including frequencies of non-occurrence, and the tendency to neglect to weight by the number of judgments upon which the proportions are based.

An application of χ^2 to percentages (Grant and Norris, **10**). A recent application of the χ^2 test to percentages (or relative frequencies) is taken from a paper by Grant and Norris (**10**). These investigators were concerned with the influence of different amounts of dark-adaptation on the sensitization of the beta-response of the human eyelid to light.[29] One of their measures of degree of sensitization was the frequency of occurrence of the response. A subject looked straight ahead into a small box-like enclosure which was painted flat black inside and out. The stimulus was a small circle of light emitted from a circular milk-glass plate, 10 cm. in diameter, located at the back of the enclosure. When illuminated, the plate had a surface brightness of 241 millilamberts. The duration of the stimulus was about 750 milliseconds.

Four experimental conditions were employed, all subjects serving in each condition. The conditions differed in the amount of dark-adaptation present in the subjects. The amount of dark-adaptation depended upon the total length of time the subjects spent in darkness. The measure of amount of adaptation was the product It, where I was the surface brightness of the stimulus plate and t was the number of seconds spent in darkness. The It products for the four conditions were 28,920; 187,980; 347,040; and 506,100. The corresponding values of t were 120, 780, 1414, and 2100 seconds.

TABLE XXIV

DATA FROM THE EXPERIMENT OF GRANT AND NORRIS (10)

	Conditions			
	1	2	3	4
P_o	13.2	30.0	45.6	51.5
P_t	11.05	36.10	44.30	49.35

Thirty-three subjects participated in the experiment. A subject was first dark-adapted for 120 seconds. The stimulus light was then presented four times, with a "control" trial between the first two and last two presentations. The control trial served as a check on possible conditioning. The four stimulus trials were separated by dark intervals of 35 sec. on the average. The first four stimulus trials (coming after 120 secs. in darkness) were the trials for condition 1. Condition 2, in which the stimulus light was again presented four times along with a control trial, came after the subject had spent a total of 780 secs. in darkness. Conditions 3 and 4, with the stimuli presented in the same general fashion, came after 1414 and 2100 secs. in darkness.

It should be noted that each of the thirty-three subjects was given four stimulus trials in each of the four dark-adapted conditions. A count was made

[29] The beta-response is one of two reflexes displayed by the eyelid when the eye is light-stimulated.

of the number of beta-responses occurring in each subject in each condition. The results were then combined to provide frequencies of occurrence of the response for each condition. These frequencies were used in computing percentages. The four observed percentages are shown in the first row of Table XXIV. They indicate the *relative* frequency of occurrence of the beta-response in each of the experimental conditions. In condition 1, for example, 13.2% of the 132 possible responses of the eyelid displayed the beta-response. The other three percentages may be similarly interpreted.

A logarithmic function was fitted to the data. The two variables were It, the amount of dark-adaptation, and P_o, the relative frequency of occurrence of the beta-response. The fitted function was: $P_t = 13.38 \log_e It - 126.39$.[30] Its solution for the empirical values of It yielded the values of P_t given in the bottom row of Table XXIV.

A value of χ^2 was calculated as follows:

$$\chi^2 = \frac{(13.2 - 11.05)^2}{11.05} + \frac{(30.0 - 36.10)^2}{36.10} + \cdots + \frac{(51.5 - 49.35)^2}{49.35} = 1.579.$$

Two constants were estimated from the data, leaving 2 df. With this number of df, the obtained value of χ^2 falls near the 40% level of confidence. This led the investigators to conclude that the data could be satisfactorily represented by a logarithmic function.

There are four mistakes in this application of the χ^2 test: two computational mistakes and two "theoretical" mistakes. The computational mistakes will be discussed first. The calculated value of χ^2 was not corrected to take account of the use of percentages instead of frequencies. It should have been multiplied by the ratio 132/100. The second computational mistake was a failure to take account of the frequency of non-occurrence of the beta-response in each of the four conditions. In condition 1, for example, the beta-response occurred 13.2% of the time and failed to occur 86.8% of the time. This latter percentage played no part in the calculation of χ^2. All four percentages of non-occurrence should have been employed. The value of χ^2 for Table XXIV, when correctly computed, is 3.081. The number of df is still 2.

The other two mistakes were more basic. The use of χ^2 to test goodness of fit was not warranted, for two reasons. In the first place, there were linkages within conditions. Each of the subjects was given four trials in each of the four conditions of dark-adaptation. The results for the subjects were pooled. There were undoubtedly individual differences in capacity to display the beta-response; so there must have been linkages within conditions. In applying the χ^2 test, the investigators assumed, in effect, that the 132 trials per condition were given to 132 instead of 33 subjects. The test was inapplicable because the assumption could not justifiably be made.

The second reason that the χ^2 test should not have been used arises from the lack of independence from condition to condition. As already stated, a fundamental requirement for the use of the χ^2 test is independence between individual measures. The differences between individual subjects that manifested themselves in any condition were certain to be maintained in the other conditions; so it cannot be assumed that the values of P_o in Table XXIV are unrelated.

[30] With common logarithms, the function becomes:
$$P_t = 30.81 \log_{10} It - 126.39.$$

Applications of the χ^2 Test to Non-Frequency Data

Some investigators are prone to use a χ^2 test of goodness of fit whenever a set of observed and theoretical values of any kind is available for comparison. This mistake apparently grows from a misinterpretation of the well-known formula for computing χ^2:

$$\chi^2 = \sum \frac{(F_o - F_t)^2}{F_t} . \qquad [10]$$

Because this equation involves differences between observed and theoretical values, the conclusion is reached that the summation of the weighted squares of differences between observed and theoretical quantities yields a meaningful estimate of χ^2. Formulas superficially resembling equation [10] have been applied to non-frequency data, where theoretical values have been obtained from fitted curves.

Suppose that a study has been made of the amount of activity displayed by several groups of white rats deprived of food for differing lengths of time. Enough rats have been included in each deprivation group to provide means that are quite stable. These means, when plotted against time of food deprivation, show a systematic trend; so a curve is fitted to them. The equation for the curve permits the calculation of theoretical means which may be compared with the observed means. To test goodness of fit, differences between the observed and theoretical means are obtained. Each difference is squared and then divided by the appropriate theoretical mean. The sum of these weighted squared differences is taken as a meaningful estimate of χ^2.

Two of the fourteen papers (4, 16) referred to in the opening paragraphs contain applications of this type. The fallaciousness of such applications becomes obvious when it is realized that values of χ^2 computed from non-frequency data vary in magnitude with the size of the units employed in measurement. Assume that two investigators have a common aim: To determine how the height of human males varies with age. They make measurements of the same individuals in various age groups, fit equations of the same form to their data, use these equations to compute theoretical values of height, and then obtain estimates of χ^2 in a manner similar to the one described above. One of the investigators has measured height in centimeters, the other in inches. Except for incidental discrepancies, the value of χ^2 calculated from the centimeter data will turn out to be 2.54 times the value calculated from the inch data.

The χ^2 test of linearity of regression. A χ^2 test is recommended in certain textbooks (26, p. 319; 12, p. 237) as suitable for use, with non-frequency data,

in evaluating linearity of regression. The formula, as usually presented, is as follows:

$$\chi^2 = \frac{\eta^2 - r^2}{1 - \eta^2}(N - k). \qquad [19]$$

In this formula, η stands for correlation ratio; r for product-moment coefficient of correlation; N for the total number of measures; and k for the number of columns (groups) into which the measures have been divided. The number of df is $k-2$. The formula yields a variable, the distribution of which approximates the χ^2 distribution, under certain conditions. It can be used with some degree of confidence provided that N is quite large and k is quite small, and provided also that the measures from column to column are independent and homoscedatic as well as normally distributed.

An exact test of linearity of regression, which is applicable whenever equation [19] is, can be made by computing an F-ratio. The formula is

$$F = \frac{(\eta^2 - r^2)(N - k)}{(1 - \eta^2)(k - 2)}. \qquad [20]$$

The df's for this F are $(k-2)$ and $(N-k)$. The reason that the χ^2 test, as defined by equation [19], can be substituted under any circumstances for the F test, as defined by equation [20], is that the distribution of F approximates the distribution of χ^2 when one of the df's for F is very small and the other is very large. The particular χ^2 distribution that is approximated is the one for the smaller df (that is, for $df = k-2$). To state the point in another way: The sampling distribution of estimates of F, obtained with equation [20], approximates the sampling distribution of estimates of χ^2 obtained with equation [19], provided that k is very small relative to N.

Inasmuch as the χ^2 test, as represented by equation [19], is inexact, while the F test, as represented by equation [20], is exact, nothing is gained by using the χ^2 test.[31]

Special Problems: I. Indeterminate Theoretical Frequencies

It sometimes happens, despite superficial indications to the contrary, that meaningful theoretical frequencies cannot be determined for a set of observed frequencies. This situation commonly arises from a lack of independence between measures. Two illustrations will be presented to reveal some of the chief sources of difficulty.

First illustration. The first illustration is concerned with coin-guessing data. Two hundred and forty university students each made a guess of heads or tails on four successive tosses of a coin. They recorded their guesses on individual record sheets. They were not told in advance the number of tosses that would be made, nor were they told how the coin turned up on the four tosses until the record sheets had been collected. The turns, in order of occurrence, were

[31] A fuller explanation of equation [20], together with a detailed discussion of other F tests of goodness of fit, is given by Lewis (**19**).

H T T H. The succession of guesses of each student could easily be compared with this succession of turns and the number of correct guesses for each could be tabulated. As expected, the number of correct guesses ranged from none to 4. The results are summarized in the first two columns of Table XXV. As shown,

TABLE XXV

ANALYSIS OF COIN-GUESSING DATA

Number Correct	F_o	$P(n)$	F_t	$\frac{(F_o-F_t)^2}{F_t}$	$P(n)'$	F_t'	$\frac{(F_o-F_t')^2}{F_t'}$
0	15	.0625	15	.000	.0490	11.76	.893
1	47	.2500	60	2.817	.1680	40.32	1.107
2	60	.3750	90	10.000	.3129	75.10	3.036
3	86	.2500	60	11.267	.3289	78.93	.633
4	32	.0625	15	19.267	.1412	33.89	.105
$\sum$	240	1.0000	240	43.351	1.0000	240.00	5.774

15 students made no correct guesses; 47 made one correct guess; 60 made two correct guesses, etc. With frequency distributions of this type, it is common practice to apply the chi-square test after a binomial distribution function has been employed to calculate the required theoretical frequencies. The function for the present case, modeled after equation [3], would have the form

$$P(n) = \frac{4!}{n!(4-n)!}\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^{4-n} = \frac{4!}{n!(4-n)!}\left(\frac{1}{2}\right)^4, \qquad [3']$$

and would yield theoretical probabilities of obtaining n correct guesses in 4.

Values of $P(n)$ calculated with equation [3'] are given in the third column of Table XXV with corresponding theoretical frequencies appearing in the fourth column. If these values of F_t were legitimate estimates, a value of χ^2 could be computed as illustrated in the fifth column of the table and used to test the hypothesis that the observed frequencies are distributed in accordance with equation [3']. In effect, this would be testing the hypothesis that each guess of every student was a purely chance occurrence, completely independent of every other guess, the probability of a guess of heads always being .50. The number of *df* is 4. If the hypothesis were true, the obtained value of χ^2, 43.351, would not be expected to arise in random sampling once in a million times.

The use of a binomial distribution function in calculating theoretical probabilities cannot be justified in this case. There is good reason for believing that the guess of a student on any of the last three tosses was not independent of previous guesses. The binomial distribution function is applicable only when there is a sound basis for assuming that every event under consideration is completely independent of every other event.

Theoretical frequencies for the empirical data in Table XXV are indeterminate—except as they might be estimated from probabilities yielded by extraneous empirical data. When hypothetical probabilities do stem from

other empirical results, the hypothesis that can be tested may be quite different from the one that would be tested if a binomial distribution function yielded the theoretical values. For example, the data of Table XXV may be compared with somewhat similar findings published several years ago by Goodfellow (9), who analyzed the patterning of the guesses on five successive tosses of a coin by a large number of radio listeners. These listeners participated in the "telepathic experiments" conducted in 1937–38 by the Zenith Foundation. The coin was tossed in the broadcasting studio in Chicago. A total of 5,687 members of the radio audience wrote down their guesses in order and mailed in their answer sheets. They were told in advance that they were to make five guesses, but were not told until two or three weeks after the broadcast how the coin had actually turned up each time. The conditions of guessing were not identical with those holding when the data in Table XXV were secured, but were similar enough to permit a comparison of the results.

Goodfellow (9, Table II) tabulated the results on the radio listeners in a way closely approximating that used in Table VIII. In fact, the only difference of any consequence is that he tabulated percentages instead of frequencies. The "correct" pattern of *H T T H* was checked against the 32 patterns in Goodfellow's table to obtain the proportions of individuals in the total group of 5,687 that hypothetically made 0, 1, 2, 3, and 4 correct guesses. These proportions are given in Table XXV, in the column headed F_t'. The resulting value of χ^2, as shown in the last column of the table, is 5.774 with $df = 4$. The hypothesis that may now be tested is that the patterning of the guesses by the 240 students was the same as the patterning of the first four of five guesses by the large group of Zenith Foundation listeners. The obtained value of χ^2 falls near the 20% level of confidence and provides no basis for rejecting this hypothesis. If the value of χ^2 *had been* large enough to justify a rejection of the hypothesis, it would not be possible to decide whether the patterning tendencies were basically different in the two groups or whether the differences in the conditions of guessing produced an apparent difference in patterning. Nevertheless, the test of the hypothesis, as stated, is an exact one. This is in sharp contrast to the indefiniteness which was present when a value of χ^2 was based on theoretical frequencies obtained with a binomial distribution function. Because of the strong likelihood of interdependence between the guesses of individual guessers, the highly significant value of χ^2 in the fifth column of Table XXV could be interpreted to mean that all guessers were biased, that some of the guessers were strongly biased while others were unbiased, that the probability of a guess of heads was not .50, that the probability of a guess of heads was .50 on some tosses but not on others, etc. The absence of independence and the consequent inability to obtain unequivocal theoretical (chance) frequencies made this application of the chi-square test a meaningless procedure.

Second illustration. Another illustration of the indeterminateness of theoretical frequencies comes from a paper by Seward, Dill and Holland (28). These investigators were concerned with an aspect of learning theory which need not be explained in order to describe their application of χ^2. The experimental procedure was relatively simple. A subject sat at a table facing a panel. On the table in front of him was a row of twelve push buttons. Ten colored cards were used as stimuli. They were exposed, one at a time, in a small rectangular window in the panel.

In the "learning" series, the colors (except blue-green) were exposed once each in a predetermined order and the subject pressed the buttons until one was found which turned on a light. Nine of the ten colors were each paired with one of the buttons. The blue-green color was associated with two buttons, the fifth and eighth. The ninth button was a blank. The blue-green color was presented twice in the "learning" series. When first presented, the light was connected with the fifth button for half of the subjects and with the eighth button for the other half. On the second presentation, the connections were reversed. Thus each subject had an opportunity to develop associative connections between blue-green and button 5, and also between blue-green and button 8.

In the "test" series, the ten colors were again presented in a predetermined order and each subject was given six chances to push the correct button for each color. In this series, blue-green was exposed last and the connection was such that the pressing of botton 5 *or* button 8 would turn on the light. The testing on any color was terminated as soon as the subject pushed the button that turned on the light or had pushed six buttons.

The investigators applied the chi-square test to the results on blue-green in an attempt to determine whether or not learning had occurred. Values of F_o and F_t, taken directly from their Table I (**28**, p. 231), are shown in the first two rows of Table XXVI. As seen, 24 of the 110 subjects pressed either button 5 or button 8 on the first trial. Twenty-five subjects pressed one of these two buttons on Trial 2, 22 on Trial 3, etc. Ten of the subjects failed to find either of the correct buttons in six trials.

TABLE XXVI

DATA FROM THE LEARNING EXPERIMENT OF SEWARD, DILL AND HOLLAND (**28**)

	Trials Required to Find Correct Button							*Total*
	1	*2*	*3*	*4*	*5*	*6*	*>6*	
F_o	24	25	22	21	4	4	10	110
F_t	18.3	16.7	15.0	13.3	11.7	10.0	25.0	110
p	2/12	2/11	2/10	2/9	2/8	2/7	?	
n	110	91.7	75.0	60.0	46.7	35.0	25.0	

The values of F_t in the second row of Table XXVI were obtained by assuming that the probabilities of success on the six trials were those given in the third row. In other words, p was assumed to be 2/12 on trial 1, 2/11 on trial 2, 2/10 on trial 3, etc. A value of χ^2 was computed as follows:

$$\chi^2 = \frac{(24-18.3)^2}{18.3} + \frac{(25-16.7)^2}{16.7} + \cdots + \frac{(10-25.0)^2}{25.0} = 31.3.$$

The only restriction imposed during the calculations was that $\sum F_o = \sum F_t = 110$. This left 6 degrees of freedom. With this number of df, the computed value of χ^2 is highly significant. It led the investigators to reject the hypothesis that the button pushing occurred by chance and to conclude that the subjects had, in fact, learned to associate blue-green with button 5 and/or button 8.

This application of the chi-square test was unwarranted because unequivocal theoretical frequencies for all trials, except possibly for trial 1, were indeterminate. The reason is that the events on trials 2 to 6 were related to previous events. The investigators seemed to be partially aware of this fact, as is shown by their choice of values of p. For example, consider the probability of succces on trial 4; it was put at 2/9. This probability arose from the assumption that each subject remembered and avoided the three buttons that had already been pushed, completely forgetting, meanwhile, all choices that had previously been made when other colors were presented. Just prior to the blue-green test series, the subjects had each pushed up to six buttons in relation to each of the other nine colors. It is inconceivable that the subjects could have remembered perfectly the buttons that had been "used up" either during the blue-green series or before on other series, unless they adopted "systems" of pressing—a system, for example, of pressing from either end toward the center, or from the center toward either end, or from the third button toward the right, or from the tenth button toward the left, etc. The adoption of any system or systems of pressing by any subject would mean that any hypothesis concerning chance proportions was open to rejection prior to the calculation of a value of χ^2.

When a decision concerning the legitimacy of theoretical probabilities is difficult to reach, it is sometimes helpful to see what happens if different probabilities are secured on the basis of new and not-unreasonable assumptions. It could be assumed, for example, that the 110 subjects used by Seward, Dill and Holland had, during the learning and early test series, discovered (on the average) that three of the twelve buttons (the two at the ends perhaps, and one in the middle) "belonged" to three colors other than blue-green, and thus by inference *did not belong* to blue-green. When blue-green was finally exposed,

TABLE XXVII

ALTERNATIVE ANALYSIS OF THE DATA OF SEWARD, DILL AND HOLLAND (28)

	Trials Taken to Find Correct Button							*Total*
	1	*2*	*3*	*4*	*5*	*6*	*6*	
F_o	24	25	22	21	4	4	10	110
F_t	24.4	21.4	18.3	15.3	12.2	9.2	9.2	110
p	2/9	2/8	2/7	2/6	2/5	2/4	?	
n	110	85.6	64.2	45.9	30.6	18.4		

three of the buttons were already "used up" and were not likely to be pressed. The theoretical frequencies might be the ones given in the second row of Table XXVII. A value of χ^2 for this table is calculated as follows:

$$\chi^2 = \frac{(24 - 24.4)^2}{24.4} + \frac{(25 - 21.4)^2}{21.4} + \cdots + \frac{(10 - 9.2)^2}{9.2} = 12.005.$$

With 6 *df*, this value falls slightly below the 5% level of confidence. It rests on dubious assumptions and, regardless of its confidence level, provides no grounds for meaningful conclusions; but the fact that it differs so greatly from the χ^2

value of 31.3 computed for Table XXVI strongly suggests that unequivocal theoretical frequencies are indeterminate in both cases.

It is not possible to provide rules-of-thumb for deciding whether theoretical frequencies are calculable or incalculable in particular situations. Decisions must ordinarily be based on careful logical analysis. However, *it is usually true that theoretical frequencies are incalculable if the observed frequencies are in any way related, and also if mutually contradictory assumptions can be made, with about equal justification, concerning the likelihood of occurrence or non-occurrence of the events (responses) that yielded the observed frequencies.*

Special Problems: II. The Nature of Imposed Restrictions

All restrictions that are imposed during the determination of theoretical frequencies should be both linear and homogeneous. This limitation is seldom mentioned in either theoretical or practical treatments of the chi-square test, and even when mentioned is usually left unexplained. The main reason for the omission is that anything beyond a very superficial explanation cannot be given in other than mathematical terms.[32] It follows that only a few very general ideas can profitably be included here.

It is probably quite obvious to most readers why restrictions must be imposed—why the sums of the observed and theoretical frequencies, for example, must always be equalized. The value of ΣF_o is fixed for any set of empirical data. The value of ΣF_t cannot "wander around any place" without at times yielding entirely impossible cell frequencies. Any hypothesis must be tied down at some point to the sample data. There should be as much freedom as possible (for example, as much freedom as possible for fluctuations in the individual cell frequencies) and yet there must be enough restrictions to bring the over-all values within the same general area.

The one restriction that must always hold may be symbolized as follows:

$$\sum F_o = \sum F_t, \qquad [21]$$

or

$$\sum (F_o - F_t) = 0. \qquad [21a]$$

[32] The excellent attempt at elementary explanation offered by Greenhood (11, Chap. 3) is about as non-mathematical as it could conceivably be made, and yet requires a considerable amount of mathematical sophistication.

This restriction is clearly linear.

Some of the other restrictions that must be imposed in familiar applications of the chi-square test may be shown to be linear—and to be homogeneous also. Suppose, for example, that a normal curve is to be fitted to an array of observed frequencies. It is not enough to impose the single restriction specified by equation [21]. The reason is that a multitude of different combinations of values of F_t, all arising from normal distributions, will each, when summated, equal ΣF_o. A second restriction must obviously be placed. It may be written

$$\sum F_o X = \sum F_t X, \qquad [22]$$

or

$$\sum X(F_o - F_t) = 0, \qquad [22a]$$

where X is the measure associated with the cell. Stated in more familiar verbal terms, the restriction is that the means of the observed and hypothetical arrays of measures shall be equal.

There is still too much freedom for values of F_t, if a normal curve is being fitted. The "scatter" of the hypothetical measures must also be restricted. In symbols,

$$\sum F_o X^2 = \sum F_t X^2, \qquad [23]$$

or

$$\sum X^2(F_o - F_t) = 0. \qquad [23a]$$

These equations, in effect, state that the variances of the observed and hypothetical arrays shall be the same.

The three restrictions represented by equations [21], [22], and [23] are imposed when a normal curve is fitted to a set of observed frequencies.[33] The equations are all linear *and* homogeneous in $(F_o - F_t)$. This is best seen from a comparison of equations [21a], [22a], and [23a].

Examples of non-linear restrictions. As a general rule, the restrictions imposed in applications of the chi-square test meet the linearity requirement. However, there are a few situations where non-linear restrictions are made. One of these has already been mentioned in the discussion of the phi-gamma function. In the phi-gamma case, theoretical frequencies are commonly obtained through a process that involves the minimizing of the sum of the squares of the differences between empirical and theoretical values of γ, where γ and p are non-linearly related. This non-linear relationship is clearly indicated by

[33] One exception to this generalization should be mentioned. If values of the hypothetical mean and hypothetical variance are not estimated from the empirical data but come from some extraneous source, the only restriction is that $\sum F_o = \sum F_t$.

equation [15]. Despite the lack of linearity, the chi-square test is often applied. It must be realized that this test is not a rigid one; any estimate of χ^2 obtained from the differences between observed and theoretical values of p is not necessarily distributed as χ^2. Consequently, whenever χ^2 is used in evaluating the goodness of fit of the phi-gamma function, conclusions regarding the fit must be made with caution.

Another situation where a non-linear restriction is imposed in calculating theoretical proportions is in the use of Thurstone's Case V in treating data obtained by the method of paired comparisons (**13**, Chap. VII). The observed proportions are based on comparative judgments, the various stimuli being compared with each other. The proportions are translated into normal deviates (that is, into z-scores). Several z-scores typically enter into the calculation of each scale separation (distance). A scale separation based on several z-scores may be reconverted into a single "theoretical" z-score, the z-score then being used to obtain a theoretical proportion. A chi-square value computed from the differences between observed and theoretical proportions would not necessarily be distributed as χ^2. Ordinarily, of course, the chi-square test is not applicable to paired comparisons data because of a lack of independence between the observed proportions.

Non-linear restrictions are always imposed in obtaining theoretical frequencies or proportions if the reduction (transformation) process is used in curve fitting. (Almost any method of estimating parameters for complex functions may involve non-linear restrictions.) To illustrate: Grant and Norris (**10**) fitted a logarithmic function to the percentages in Table XXIV. In doing this, they probably transformed values of the independent variable (that is, values of *It*) into their logarithmic equivalents and used these in estimating the parameters for the logarithmic function. If they followed such a procedure, the calculated theoretical percentages were dependent upon a non-linear restriction.

Unfortunately, there is no way of determining the exact influence of non-linear restrictions on estimated values of χ^2. Therefore, if the chi-square test is applied, despite the imposition of non-linear restrictions, the investigator must be extremely cautious in interpreting the results,[34] bearing in mind that, as Cramér shows (**5**, Ch. 30), the calculated value of χ^2 is probably somewhat larger than it would be if the restrictions were linear.

CONCLUSIONS

Most readers will by now have correctly concluded that the chi-square test has a restricted usefulness. However, it usually cannot be replaced in those situations where it is applicable and it thus stands as a valuable research tool. Perhaps the chief trouble is that the test is too often applied without adequate prior planning; it is frequently "hit upon" and adopted after data have been collected and sometimes after other techniques of statistical analysis have been found unproductive.

[34] These statements hold also when the F statistic, based as it is on two independent estimates of χ^2, is applied in testing goodness of fit. For example, several of the F-tests proposed by Lindquist (**22**) depend for their exactness on the type of restrictions placed in estimating the parameters of the fitted functions.

The aim of every investigator should be to plan, in advance, not only every detail of every experiment but every step in the analysis of the anticipated data. All contingencies cannot be foreseen; but if the chi-square test is to be employed, there is no good reason for failing to provide for independence among the measures and for frequencies of adequate size.

There should seldom if ever be any compromising on the requirement of independence.[35] There should usually be no compromising on the size of frequencies. There are occasions, of course, when it is very time-consuming and perhaps very expensive to add more cases to a mere handful. The best procedure under such circumstances is to try for an experimental design which utilizes each subject to the limit and leads to an analysis of data on an individual rather than a group basis. If it turns out that only a few subjects can be studied and the data on each one cannot be analyzed separately, it may still be possible to find a method of analysis which is more exact than the chi-square test. For example, if the data can be arranged into a 2×2 table and the individual cell frequencies are less than 10, the *exact treatment* proposed by Fisher (7, pp. 96–97) is to be preferred to the chi-square test. The treatment is rather tedious to apply, but in view of its exactness there is no adequate excuse for avoiding it.

Many users and would-be users of the chi-square test gain erroneous impressions from what they read about limitations on the size of theoretical frequencies. A textbook says that frequencies of less than 10 are to be avoided. This statement is often interpreted to mean not that 10 is a limiting value to be exceeded whenever possible, but that 10 is a value around which the various theoretical frequencies may fall; and if an occasional frequency happens to be as low as 4 or 5, that is all right because other frequencies will be larger than 10 and everything will average out in the end. A textbook that gives 5 as the suggested minimum tends to encourage the retention of impossibly small theoretical frequencies. And so does a text which states, in effect, that Yates' correction for continuity should be applied if the cell frequencies are 5 or less and precision is desired. This implies not only that frequencies of less than 5 are quite acceptable, but also that Yates' correction is an antidote for small frequencies. Both implications are fallacious.

The following excerpts from Yule and Kendall (**35**, p. 422) may help

[35] Any investigator who applies the chi-square test to interdependent frequency data should always feel obligated to include, in published accounts of his findings, a full explanation of the procedures employed and a justification of them.

to dispel false notions concerning the size of theoretical frequencies and also concerning the size of N:

> In the first place, N must be reasonably large. . . . It is difficult to say exactly what constitutes largeness, but as an arbitrary figure we may say that N should be at least 50, however few the number of cells.
>
> No theoretical cell frequency should be small. Here again it is hard to say what constitutes smallness, but 5 should be regarded as the very minimum, and 10 is better.

Hoel (14, p. 191), while giving 5 instead of 10 as the recommended minimal value of F_t, nevertheless emphasizes the importance of having a fairly large value of N by stating that if the number of cells or categories is less than 5, the individual theoretical frequencies should be larger than 5. Cramér (5, p. 420 f.) firmly recommends a minimal value of 10 and says that if the number of observations is so few that the theoretical frequencies, even after grouping, are not greater than 10, the chi-square test should not be applied. In all but one of the illustrations used by Cramér, the theoretical frequencies are considerably larger than 10, and in the exceptional case, he admits (p. 440) that the frequencies are smaller "than is usually advisable." An investigator handicaps himself whenever he applies the chi-square test in relation to small theoretical frequencies.

There are a few applications of the χ^2 test which have not been described and illustrated in the present survey, either because they are quite specialized in character or because they provide only approximate solutions. One of the specialized applications which may be of interest to some readers is Bartlett's test of the homogeneity of variance (3). Those who do not have access to Bartlett's original discussion will find a description of the test in Snedecor (31, pp. 249–251). Another specialized application of interest is the use of χ^2 in setting the confidence limits for a population variance from a known sample variance. The procedure is explained by Hoel (14, pp. 138–140) and need not be included here.

In general, any suggested applications of χ^2 which deviate from the well-established tests should be avoided except by those qualified to evaluate their full import or upon the advice of an expert.

BIBLIOGRAPHY

1. ANASTASI, ANNE, & FOLEY, J. P., JR. An experimental study of the drawing behavior of adult psychotics in comparison with that of a normal control group. *J. exp. Psychol.*, 1944, **34**, 169–194.
2. ARNOLD, MAGDA B. Emotional factors in experimental neuroses. *J. exp. Psychol.*, 1944, **34**, 257–281.
3. BARTLETT, M. S. Properties of sufficiency and statistical tests. *Proceed. Royal Soc.*, 1937, *A* **160**, 268–282.
4. CHEN, H. P., & IRWIN, O. C. Development of speech during infancy: Curve of differential percentage indices. *J. exp. Psychol.*, 1946, **36**, 522–525.
5. CRAMÉR, H. *Mathematical methods of statistics*. Princeton: Princeton Univ. Press, 1946.
6. EDWARDS, A. L. *Statistical analysis for students in psychology and education*. New York: Rinehart, 1946.
7. FISHER, R. A. *Statistical methods for research workers* (10th Ed.). Edinburgh: Oliver and Boyd, 1946.
8. FRY, T. C. The chi-square test of significance. *J. Amer. stat. Ass.*, 1938, **33**, 513–525.
9. GOODFELLOW, L. D. A psychological interpretation of the results of the Zenith Radio experiments in telepathy. *J. exp. Psychol.*, 1938, **23**, 601–632.
10. GRANT, D. A., & NORRIS, EUGENIA B. Dark adaptation as a factor in the sensitization of the beta response of the eyelid to light. *J. exp. Psychol.*, 1946, **36**, 390–397.
11. GREENHOOD, E. R., JR. *A detailed proof of the chi-square test of goodness of fit*. Cambridge: Harvard Univ. Press, 1940.
12. GUILFORD, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1942.
13. GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936.
14. HOEL, P. G. *Introduction to mathematical statistics*. New York: Wiley, 1947.
15. HOLT, R. B. Level of aspiration: Ambition or defense? *J. exp. Psychol.*, 1946, **36**, 398–416.
16. IRWIN, O. C., & CHEN, H. P. Development of speech during infancy: Curve of phonemic types. *J. exp. Psychol.*, 1946, **36**, 431–436.
17. KING, H. E., LANDIS, C., & ZUBIN, J. Visual subliminal perception where a figure is obscured by the illumination of the ground. *J. exp. Psychol.*, 1944, **34**, 60–69.
18. KUENNE, MARGARET R. Experimental investigation of the relation of language to transposition behavior in young children. *J. exp. Psychol.*, 1946, **36**, 471–490.
19. LEWIS, D. *Quantitative methods in psychology*. Ann Arbor: Edwards Bros., 1948.
20. LEWIS, HELEN B. An experimental study of the role of the ego in work. I. The role of the ego in cooperative work. *J. exp. Psychol.*, 1944, **34**, 113–126.
21. LEWIS, HELEN B., & FRANKLIN, MURIEL. An experimental study of the role of the ego in work. II. The significance of task-orientation in work. *J. exp. Psychol.*, 1944, **34**, 195–215.
22. LINDQUIST, E. F. Goodness of fit of trend curves and significance of trend differences. *Psychometrika*, 1947, **12**, 65–78.
23. LINDQUIST, E. F. *Summary report of results*. Iowa City: Sixth Annual Every-Pupil Testing Program, 1934.
24. PEARSON, K. Experimental discussion of the (χ^2, p) test for goodness of fit. *Biometrika*, 1932, **24**, 351–381.
25. PEATMAN, J. G. *Descriptive and sampling statistics*. New York: Harpers, 1947.
26. PETERS, C. C., & VAN VOORHIS, W. R.

Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940.

27. PRONKO, N. H. An exploratory investigation of language by means of oscillographic and reaction time techniques. *J. exp. Psychol.*, 1945, **35**, 433–458.
28. SEWARD, J. P., DILL, JANE B., & HOLLAND, MILDRED A. Guthrie's theory of learning: A second experiment. *J. exp. Psychol.*, 1944, **34**, 227–238.
29. SHAW, F. J., & SPOONER, ALICE. Selective forgetting when the subject is not "ego-involved." *J. exp. Psychol.*, 1945, **35**, 242–247.
30. SMITH, S. The essential stimuli in stereoscopic depth perception. *J. exp. Psychol.*, 1946, **36**, 518–521.
31. SNEDECOR, G. W. *Statistical methods* (4th Ed.). Ames, Ia.: Iowa State College Press, 1946.
32. STEVENS, S. S., MORGAN, C. T., & VOLKMANN, J. Theory of the neural quantum in the discrimination of loudness and pitch. *Amer. J. Psychol.*, 1941, **54**, 315–335.
33. VOELKER, C. H. Phonetic distribution in formal American pronunciation. *J. acous. Soc. Amer.*, 1934, **5**, 242–246.
34. YATES, F. Contingency tables involving small numbers and the χ^2 test. *Suppl. J. Royal stat. Soc.*, 1934, **1**, 217–235.
35. YULE, G. U., & KENDALL, M. G. *An introduction to the theory of statistics* (12th Ed.). London: Griffin, 1940.

Received June 19, 1949.