CHI-SQUARE TESTS WITH ONE DEGREE OF FREEDOM; EXTENSIONS OF THE MANTEL-HAENSZEL PROCEDURE

NATHAN MANTEL
National Cancer Institute

A published method for analyzing multiple 2×2 contingency tables arising in retrospective studies of disease is extended in application and form. Extensions of application include comparisons of age-adjusted death rates, life-table analyses, comparisons of two sets of quantal dosage-response data, and miscellaneous laboratory applications as appropriate. Extensions in form involve considering multiple contingency tables with arbitrarily many rows and/or columns, where rows and columns are orderable, and may even be on a continuous scale. The assignment of some score for each row or column is essential to use of the method. With scores assigned, a deviation of the sum of cross products from expectation, and its variance conditioned on all marginal totals, are computed for each table and a chi square is determined corresponding to the grand total of the deviations. For various specific instances and for various scoring procedures, the procedure extends or is equivalent to the asymptotic form of many known nonparametric techniques.

1. INTRODUCTION

In 1959 Mantel and Haenszel [8], in a review of the statistical problems of retrospective studies of disease, provided some procedures for analyzing data and establishing associations from such studies. Noteworthily, they suggested the use of a summary chi square with one degree of freedom in testing the association of disease incidence with any particular factor when the effect of any other factor or group of factors is held constant. For any set of specified factors, e.g. age, race, sex, and occupation, there is a 2×2 contingency table, individuals being classified as with or without the disease, and with or without the study factor. In each such 2×2 table, conditional on the marginal totals, Mantel and Haenszel determine the expectation and variance of the number of diseased individuals positive for the study factor. Summation of the observed and expected number of such cases is made over all the 2×2 tables, and chi square is computed as the square of the cumulated discrepancy, corrected for continuity, divided by the sum of the conditional variances.

(Along with the significance testing procedure, the authors also provided various measures of relative risk as an index of the strength of the association. They pointed out that if, in some sense, e.g. constant logit or probit difference, the relative risk could be assumed uniform over the various contingency tables, it would be appropriate to make some best estimate of this uniform relative risk. Unwilling to make the assumption of uniformity, the authors suggested their procedure which, to some degree, weights the separate relative risks by importance, statistical power being increased by the reinforcement of relative risks prevailing in the same direction.)

The present report covers extensions of the Mantel-Haenszel (M-H) procedure in two directions. In the first it is recognized that the procedure is not

limited in application to the problem of retrospective studies, and its suitability for other kinds of problems is described. In the second extension, the case is considered in which the number of levels for the study factor is arbitrary, but orderable, rather than limited to two. Mantel and Haenszel did cover the case of study factors at several levels and gave a specific procedure for calculating chi square with two degrees of freedom for the three-level case. By taking ordering into account, single degree of freedom chi squares can be determined for the multiple level problem. A variety of standard nonparametric procedures, in asymptotic form, can be derived, and extended beyond their present range of application, by manipulating the score assigned to a particular study factor level.

2. EXTENSIONS IN APPLICATION

The author has found the Mantel-Haenszel chi-square procedure to be applicable to data from a variety of laboratory investigations. Suppose an investigator is comparing two treatment procedures where response to treatment can be put in the quantal, all-or-none, form. It will frequently be the case that the data obtained will include a number of comparisons of the very same treatments. There may have been one or more additional adjustable factors, so that comparison is made at several combinations of the adjustable factors. Allocation of litter mate animals to the two treatment procedures may permit a separate comparison for each litter. Finally, the investigator have may conducted a series of repeat experiments, comparison of the two procedures being possible in each of the experiments. One wishes then to employ a powerful statistical procedure which would be sensitive to any consistent difference in outcome for the two treatments.

By considering each of the several comparisons in a study as falling into the 2×2 contingency table form, the M-H procedure can be applied. Expectations and conditional variances can be determined for each such table, and a summary chi square calculated.

Application of the M-H procedure is suggested also for certain problems for which there are standard methodologies. Consider, for example, the comparison of two age-adjusted (or even age, sex, and race-adjusted) death or disease incidence rates [7]. Essentially, for each age interval one has a 2×2 contingency table, the age specific rates for each sample or population being based on the ratio of the number responding to the number exposed. An overall summary chi square for comparing the 2 sets of age specific rates can be calculated; effectively 2 age-adjusted rates are compared, where adjustment is by the indirect method and the set of standard age specific rates are defined by the rates for 2 populations combined.

A somewhat similar comparison arises in problems for which the life-table procedure [10] is ordinarily appropriate. For each particular interval following some zero time, one knows the population or sample size at the beginning of the interval and the number of deaths (responses) during the interval. Sample sizes may decrease between successive intervals because of death or loss to follow-up, or for reasons flowing from the arrangement of the study. Here it is suggested that the M-H procedure could be appropriate for comparing two

sets of such follow-up data, a 2×2 contingency table arising from each study interval. The usefulness of the procedure in this problem and the interesting leads to which it gives rise will be the subject of a separate communication.

As a final extension in application, it is suggested that the M-H procedure can sometimes be applied to data for which a complex methodology now exists, evaluation of quantal dosage response data by the probit method [3]. By standard methods, if one were interested in comparing two sets of data, one would determine the median effective dose (ED_{50}, or LD_{50} for median lethal dose) for each set and compare these for statistical significance. Under some situations it would be necessary to fit the data under the restriction of parallelism, and the correlation of the two ED_{50}'s would then have to be considered. Important to this procedure is the assumption of the validity of the probit model and also the assumption of parallelism. Presumably, the actual difference in the two ED_{50}'s is of interest.

But suppose that we are dealing with a single effective agent and wish to demonstrate that manipulation of another factor can increase (or decrease) the proportion responding. Situations may arise in which we may be unwilling to accept the reasonableness of the probit model for the altered situation, though in the original situation it could have been justified By the simple device of employing the same dose levels at both factor levels, the M-H procedure can be used to evaluate any consistent difference in the proportion responding with no need to assume the validity of the probit model at either level. There is no need either to calculate ED_{50}'s, and the M-H procedure can even be applied to data which do not permit the determination of ED_{50}'s. The M-H test is, in fact, a valid procedure for comparing two ED_{50}'s determined by the Kärber's method, where the data permit their calculation.

Application of the M-H procedure to data by Eagle and reported by Cornfield and Mantel [3] is illustrated in Table 1. The data relate to the effectiveness of immediately injected or 1½ hour delayed penicillin in protecting rabbits against lethal inoculation with β-hemolytic streptococci. (In Cornfield and Mantel [3], data are shown also for a 6-hour delay; the data are used to illus-

TABLE 1. APPLICATION OF MANTEL-HAENSZEL PROCEDURE TO DATA OF EAGLE

Penicillin level	*No delay* No. cured/ No. dying A_i/B_i	*1½ hour delay* No. cured/ No. dying C_i/D_i	Expectation* A_i	Variance** A_i
1/8	0/6	0/5	0	0
1/4	3/3	0/6	1.5	27/44
1/2	6/0	2/4	4.0	32/44
1	5/1	6/0	5.5	11/44
4	2/0	5/0	2.0	0

* $E(A_i)=N_{1_i}M_{1_i}/T_i$, $N_{1_i}=A_i+B_i$, $M_{1_i}=A_i+C_i$, $T_i=A_i+B_i+C_i+D_i$

** $V(A_i)=N_{1_i}M_{1_i}N_{2_i}M_{2_i}/T_i^2(T_i-1)$, $N_{2_i}=C_i+D_i$, $M_{2_i}=B_i+D_i$.

$\Sigma A_i=16$; $\Sigma E(A_i)=13$; $\Sigma V(A_i)=70/44$

$$\chi_c^2=(|\Sigma A_i-\Sigma E(A_i)|-0.5)^2/\Sigma V(A_i)=\frac{2.5^2\times 44}{70}=3.93.$$

trate the computation of three-way parallel dosage response curves by the Cornfield-Mantel method.) A 2×2 contingency table is constructed for each of the common levels of penicillin employed.

That there is significant evidence for a higher cure rate when treatment is immediate is demonstrated in Table 1 without the calculation of ED_{50}'s or the fitting of parallel dosage response curves.

3. EXTENSIONS TO THE CASE OF STUDY FACTORS AT SEVERAL LEVELS

Suppose the factor under study may assume k orderable levels. There will be a separate $2\times k$ contingency table for each specified set of control factors. Comparing disease cases with controls, for example, in the i'th contingency table there will be A_{j_i} disease cases at study factor level j, B_{j_i} control cases. We may arbitrarily assign an X value of 1 to disease cases, 0 to controls; to the j'th study factor level let us consider that there is attached a Y score, Y_j, the nature of which will we discuss below. The appearance of the i'th contingency table will then be as shown in Table 2.

TABLE 2. ILLUSTRATION OF A $2\times k$ CONTINGENCY TABLE

		Study factor level, j				
		0	1	2	$3 \cdots k-1$	Total
	$X\backslash Y$	Y_0	Y_1	Y_2	$Y_3 \cdots Y_{k-1}$	—
With disease	1	A_{0_i}	A_{1_i}	A_{2_i}	$A_{3_i} \cdots A_{k-1_i}$	N_{1_i}
Free of disease	0	B_0	B_{1_i}	B_{2_i}	$B_{3_i} \cdots B_{k-1_i}$	N_{2_i}
Total	—	M_{0_i}	M_{1_i}	M_{2_i}	$M_{3_i} \cdots M_{k-1_i}$	T_i

The statistic of interest with which we shall concern ourselves here, and which we justify below, is $\sum XY = \sum_j A_{j_i} Y_j$, or rather its deviation from expectation. Under the null hypothesis of no association, the expected value of $\sum XY$ subject to all marginal totals (the N's and M's) fixed is $T_i \overline{X}_i \overline{Y}_i$. But

$$\overline{X}_i = N_{1_i}/T_i; \quad \overline{Y}_i = \sum_j M_{j_i} Y_j / T_i \quad \text{and} \quad E(\sum XY) = \sum_i N_{1_i} \sum_j M_{j_i} Y_j / T_i.$$

The variance of $\sum XY$ conditional on the marginal totals can be determined once we recognize that, irrespective of the marginal totals, it is simply the variance of the total of a sample size N_{1_i} drawn without replacement from a finite population of size T_i. In the finite population the frequency of each Y_j value is M_{j_i}. The variance of the total is $N_{1_i}\times$variance of the population$\times$finite population correction factor. The population variance is given by

$$\sigma^2 = \frac{1}{T_i}\left[\sum_j M_{j_i} Y_j^2 - \frac{\left(\sum_j M_{j_i} Y_j\right)^2}{T_i}\right] \tag{1}$$

while the correction factor is $N_{2_i}/(T_i-1)$.

In summary, we get as the variance of $\sum XY = \sum_j A_{j_i} Y_j$

$$V\left(\sum_j A_{j_i} Y_j\right) = \frac{N_{1_i} N_{2_i}}{T_i^2(T_i - 1)}\left[T_i \sum M_{j_i} Y_j^2 - \left(\sum_j M_{j_i} Y_j\right)^2\right] \tag{2}$$

For a single contingency table chi square, with one degree of freedom, can be computed as

$$\chi^2 = \frac{\left(\sum_j A_{j_i} Y_i - \frac{N_{1_i}}{T_i}\sum_j M_{j_i} Y_j\right)^2}{V\left(\sum_j A_{j_i} Y_j\right)} \tag{3}$$

and a summary chi square, again with one degree of freedom, can be computed as

$$\chi^2 = \frac{\left[\sum_i \sum_j A_{j_i} Y_j - \sum_i E\left(\sum_j A_{j_i} Y_j\right)\right]^2}{\sum_i V\left(\sum_j A_{j_i} Y_j\right)} \tag{4}$$

(No general specification has been made here of how to correct for continuity. The correction would depend on the Y scores employed and some scoring procedures could make difficult the justification of any particular correction. A practical procedure might be to take the correction as one-half the smallest difference between any two successive Y scores—but the possible increments in $\sum A_{j_i} Y_j$ could, however, be much smaller than this smallest difference.)

We come now to justification for the use of the test statistics suggested. Note first that in a single contingency table the test statistic is of the form $\sum(X-\overline{X})(Y-\overline{Y})$. We might, in opposition, have considered as test statistics, the regression of Y on X, $\sum(X-\overline{X})(Y-\overline{Y})/\sum(X-\overline{X})^2$ or the regression of X on Y, $\sum(X-\overline{X})(Y-\overline{Y})/\sum(Y-\overline{Y})^2$. Whichever of the latter two we prefer, we find that weighting by precisions, which will be proportional to $\sum(X-\overline{X})^2$ or to $\sum(Y-\overline{Y})^2$, will leave the weighted regression coefficient dependent only on $\sum(X-\overline{X})(Y-\overline{Y})$. Assuming no important change in variability between contingency tables, the statistic selected becomes the one of choice. It can be seen also that, effectively, we are testing the sum total of Y scores of all diseased (or non-diseased) individuals, conditional on all marginal totals in all the contingency tables. By testing $\sum XY$ rather than the regression coefficients we also avoid problems of non-linearity. A linear dependence of X on Y could probably not be justified, but $\sum XY$ could still be powerful for detecting a progressive dependence.

The notion of regression comes into play when, in addition to testing overall significance, we wish to obtain some measure of the degree of association of disease and the study factor or the degree of difference between disease and

control subjects. The weighted average regression of Y on X would be

$$\sum\sum (X - \overline{X})(Y - \overline{Y}) / \sum\sum (X - \overline{X})^2;$$

for X on Y it would be

$$\sum\sum (X - \overline{X})(Y - \overline{Y}) / \sum\sum (Y - \overline{Y})^2.$$

In the problem considered, the first of the two weighted average regressions would seem the more appropriate. With X a 0, 1 variable, $\sum(X-\overline{X})^2$ within a contingency table reduces to $N_{1i}N_{2i}/T_i$ and the average regression or, more properly, the weighted average difference in score between disease cases and controls becomes

$$\left[\sum_i \sum_j A_{ji} Y_j - \sum_i \frac{N_{1i}}{T_i} \sum_j M_{ji} Y_j \right] \Big/ \sum_i \frac{N_{1i}N_{2i}}{T_i} . \tag{5}$$

This weighted average difference may be recognized to follow the standard formula, the difference in average score for disease and control cases in each contingency table being weighted by $N_{1i}N_{2i}/(N_{1i}+N_{2i})$.

Specification of scores

The results of analysis by the procedure just indicated may vary markedly according to how Y scores are assigned to the various study factor levels. It is important then that such scoring be made in an objective fashion, one independent of the differences occurring between disease and control subjects in their distribution by study factor level. It may be possible sometimes to make a reasonable assignment of scores in advance without reference to the study outcomes, or the assignment made may be such as to make the average scores meaningful, e.g. smoking category scores may be such that the average score approximates average cigarette use.

Some other objective scoring procedures which may be useful are now described.

1. Rather simply one might assign the score j to the j'th study factor level. This leads to simple computational procedures and does have increased statistical power for any progressive effects of the study factor. There is no necessary implication in using this scoring procedure that the various levels of the study factor are, in some sense, equally spaced.

2. Scoring may be by ranks, a tied ranking being assigned to individuals at the same study factor level. Rankings may be done separately for each contingency table or across all contingency tables combined. To standardize for the varying number of individuals in the different tables, the rank should be expressed relative to the total, T_i, for the table.

For separate table ranking procedures, and note that this results in a separate set of scores for each contingency table, the ranking scores become:

$$Y_{0i} = \frac{1}{T_i}\left(\frac{M_{0i}+1}{2}\right); \tag{6}$$

$$Y_{1_i} = \frac{1}{T_i}\left(M_{0_i} + \frac{M_{1_i}+1}{2}\right); \tag{7}$$

$$Y_{2_i} = \frac{1}{T_i}\left(M_{0_i} + M_{1_i} + \frac{M_{2_i}+1}{2}\right); \tag{8}$$

$$\vdots$$

$$Y_{r_i} = \frac{1}{T_i}\left(\sum_{j=0}^{r-1} M_{j_i} + \frac{M_{r_i}+1}{2}\right). \tag{9}$$

Where data from all tables are combined in assigning ranks, so that the same scoring procedure applies for all tables, the scoring formula is

$$Y_r = \frac{1}{\sum_i T_i}\left(\sum_i \sum_{j=0}^{r-1} M_{j_i} + \frac{\sum_i M_{r_i}+1}{2}\right). \tag{10}$$

(The scoring procedures just described based on tied ranks strongly resemble Bross' ridit scores [1]. The distinction is that Bross' ridits are based on only control data which presumably are so extensive that they define a *relatively identified distribution.* In practice [12, 13], however, such control data have been only about as extensive as the study data. Where statistical variation in both control and study group data must be taken into account, ridit procedures reduce to standard ranking procedures, though tied ranks may be frequent. The ridit procedure has been used primarily in a descriptive fashion, though the setting of confidence intervals on average ridits, assuming the uniform distribution variance of 1/12, has a significance testing aspect. If the identified distribution is obtained by combining control data across several contingency tables, it is likely that the resulting variance will be greater than that existing in the separate tables. The use of the standard variance of 1/12 would lead to excessively wide confidence intervals.

There is a disturbing feature about the use of ridits or ranks. In injury analysis, for example, fatal accidents may be relatively rare and will so be assigned a score only slightly higher than that for severe, or extremely severe, accidents. In contrast, with injuries of minor and moderate severity common, the difference in score between two such injuries will be large.)

Scoring of study factors with continuous scales

It may be noted that the variance and chi-square formulas given above would be appropriate even if data were not in contingency table form. This was deliberate, to permit their applicability to arbitrary situations including those in which there is a distinctive study factor value for each individual in the study. The relationship of the resulting analysis to standard analyses when this is done is interesting. Let us consider the results of applying various scoring procedures to the individual study factor values.

1. General scoring procedure —Y scores already known for each individual.

For a single contingency table, the chi-square test is a test of the difference in the two Y averages, conditional on the combined distribution of Y scores

and on the number of disease and control subjects in the table. As such it can be recognized to be the asymptotic form of the Pitman permutation test [9] for comparing two averages. The extension to the case of several contingency tables can then be viewed as a generalization of the asymptotic form of the Pitman test.

In particular, where each contingency table comprises but two individuals, one disease case and one control, the chi-square test corresponds to the asymptotic form of Fisher's randomization test for matched pairs [4]. That test is one in which the null distribution of the sum of the paired differences is determined assuming the sign of each particular difference to be positive or negative with equal probability.

2. Rank scoring procedures.

For a single contingency table in which the ranks would range from 1 to T (or from $1/T$ to 1), the chi-square test is a test of the difference in average ranks for the two groups. As such it is the asymptotic form of the rank sum test which has been reported by various investigators including Wilcoxon [11]. Extensions to the case of several contingency tables may be regarded as generalizations of the asymptotic form of the rank sum test. For several contingency tables, as indicated above, rankings may be done separately for each table, $1/T_i$ to 1, or in summary for all tables, $1/\sum T_i$ to 1.

Where ranking is done separately for each table, and there are but two individuals in each table, one disease case and one control, the chi-square test corresponds to the asymptotic form of the sign test [2]. The test here is on the departure from expected equality of the number of positive and negative signs.

With still other scoring procedures, e.g. rankits, the suggested chi-square methodology will correspond to still other standard analytic procedures.

The case of disease status at several levels

It may sometimes be the situation that there are more than the two disease categories, well or ill—disease may be in a more or less advanced stage or of a more or less serious nature. Suppose that the disease status can be ordered in some way so that disease scores can be assigned. For generality, so that we may cover both the continuous case and the case of a finite number of disease statuses, let the status score of the j'th individual be X_j, his study factor score, Y_j.

For a single set of data the departure of $\sum X_jY_j$ from its expectation is

$$\sum X_jY_j - \frac{\sum X_j \sum Y_j}{T}. \tag{11}$$

Using finite population concepts, under the hypothesis of independence the variance of $\sum X_jY_j$, subject to the observed set of X's and Y's in the data, becomes

$$V(\sum X_jY_j) = \frac{T^2\sigma^2_X\sigma^2_Y}{T-1} \tag{12}$$

where

$$\sigma_X^2 = \left(\sum X_j^2 - \frac{(\sum X_j)^2}{T}\right) \Big/ T$$

$$\sigma_Y^2 = \left(\sum Y_j^2 - \frac{(\sum Y_j)^2}{T}\right) \Big/ T.$$

For several sets of data chi square, with a single degree of freedom, can be computed on the basis of the cumulated departures and variances as

$$\chi^2 = \frac{\left[\sum_i \left(\sum_j X_{ji} Y_{ji} - \frac{\sum_j X_{ji} \sum_j Y_{ji}}{T_i}\right)\right]^2}{\sum_i \left[\frac{1}{T_i - 1}\left(\sum_j X_{ji}^2 - \frac{\left(\sum_j X_{ji}\right)^2}{T_i}\right)\left(\sum_j Y_{ji}^2 - \frac{\left(\sum_j Y_{ji}\right)^2}{T_i}\right)\right]} \cdot \quad (13)$$

Use of this chi-square procedure is equivalent to testing the simple regression coefficient for a single set of data or the pooled regression coefficient where there are several sets of data. That a linear regression coefficient is being tested does not mean that an assumption of linearity is being made. Rather it is that test of a linear component of regression provides power for detecting any progressive association which may exist between the variables studied.

According to how X and Y scores are assigned, chi square may have various interpretations, the basic one being that it provides a test with power for any progressive relation between X and Y. It may be noted that where the X's and Y's are ranks, for a single set of data, chi square tests the coefficient of rank correlation. For several sets of data, and using ranking scores from $1/T_i$ to 1 in each set, the procedure provides a generalization of the coefficient of rank correlation.

As desired, and corresponding to any particular value of chi square, summary regression coefficients, correlation coefficients, or weighted average differences between two groups can be calculated. Formulas for doing so are readily determined and are not given here.

Relation between proposed and standard methodology

The proposed chi-square test is, in general form, a test of the pooled covariance between X and Y, conditional on the set of observed X and Y values. By usual analysis of covariance techniques, one would have made an F test of the estimated common slope for the several sets of data—here the assumption of homogeneity of variances is made, and the data are considered as samples from infinite normal populations. The F value, and its significance, remains the same whether one considers the regression of Y on X or of X and Y, and so there is no loss in generality if we consider the F test to relate to the regression of Y on X. (Though the F value is not altered by incorrectly considering the regression of the independent on the dependent variable, the value and meaningfulness of the regression coefficient is affected. Going further, if one is

interested in computing meaningful correlation coefficients, one should be certain that the X, Y values can be considered as samples from a bivariate population, rather than as a pair of variate values, one independent, the other dependent.)

For a single set of data, the relationship between χ^2 and the F (or t^2) value obtained in testing the regression coefficient or the difference between two averages by standard methods can be expressed alternatively as

$$F = \chi^2\left(\frac{T-2}{T-1-\chi^2}\right) \tag{14}$$

or as

$$\chi^2 = F\left(\frac{T-1}{T-2+F}\right) \tag{15}$$

where F is with $(1, T-2)$ degrees of freedom, χ^2 with one degree of freedom. Where F or chi square is greater than unity, usually a minimum requirement for significance, F will exceed chi square. The excess of F over chi square will ordinarily be sufficient to correspond to more highly significant probability values; this is true even though chi square can be considered to be an F with infinite denominator degrees of freedom.

Perhaps the greater significance for the F value relates to the stronger assumption of normality implicit in its use. But there are indications that, in circumstances, use of the F statistic may yield probability values corresponding more closely to the permutational distribution—Greenhouse [5] suggests that the use of F errs slightly on the radical side, use of chi square on the conservative side. The range of possibly computed F values is from 0 to infinity, which is the range of the F distribution. The conservatism of chi square may be related to the fact that, as computed, it must range between 0 and $T-1$ although the chi square distribution has no upper bound. To the points above it may be noted that Kendall [6], in determining the asymptotic form of the Pitman permutation test obtains the t test, the use of which is equivalent to the use of F, rather than chi square. In his development Kendall takes into account the 3rd and 4th moments.

The simple relation between the proposed chi square and F no longer applies when several sets of data are considered. Although both are tests of the sum of cross-products of X's and Y's cumulated over the several data sets, there is implicit in the F test, but not in the chi square test, an assumption of homogeneity of variances. One may define a chi-square value, χ^{2*}, which is related to the common slope F value in an analogous manner to that obtaining between chi square and F for a single set of data. Given k sets of data which have been analyzed for a common slope or common difference between averages, the relation may be expressed as

$$F = \chi^{2*}\left(\frac{\sum T - k - 1}{\sum T - k - \chi^{2*}}\right) \tag{16}$$

or as

$$\chi^{2*} = F\left(\frac{\sum T - k}{\sum T - k - 1 + F}\right). \tag{17}$$

This alternative chi-square value will, of course, differ from the one proposed above.

REFERENCES

[1] Bross, I. D. J., "How to use ridit analysis," *Biometrics*, 14 (1958), 18–38.
[2] Cochran, W. G., "The efficiencies of the binomial series tests of significance or a mean and of a correlation coefficient," *Journal of the Royal Statistical Society*, 100 (1937), 69 ff.
[3] Cornfield, J., and Mantel, N., "Some new aspects of the application of maximum likelihood to the calculation of the dosage response curve," *Journal of the American Statistical Association*, 45 (1950), 181–210.
[4] Fisher, R. A. *The Design of Experiments*, Edinburgh: Oliver and Boyd, 1935. See pp. 43–47, 3rd edition, 1942.
[5] Greenhouse, S. W., Personal communication.
[6] Kendall, M. G., *The Advanced Theory of Statistics*, Volume 2. London: Charles Griffin and Company Limited, 1946. See pp. 127–31, 2nd edition, 1948.
[7] Linder, F. E., and Grove, R. D. "Technique of vital statistics," reprint of chapters 1–4 of *Vital Statistics Rates in the United States, 1900–1940*, National Office of Vital Statistics.
[8] Mantel, N., and Haenszel, W., "Statistical aspects of the analysis of data from retrospective studies of disease," *Journal of the National Cancer Institute*, 22 (1959), 719–48.
[9] Pitman, E. J. G., "Significance test which may be applied to samples from any population," *Supplement to the Journal of the Royal Statistical Society*, 4 (1937), 119 ff.
[10] Spiegelman, M., "The versatility of the life table," *American Journal of Public Health* 47 (1957), 297–304.
[11] Wilcoxon, F., "Individual comparisons by ranking methods." *Biometrics*, 1 (1945), 80–3.
[12] Wynder, E. L. and Bross, I. J., "A study of etiological factors in cancer of the esophagus," *Cancer* 14 (1961), 389–413.
[13] Wynder, E. L., Bross, I. J., and Hirayama, T., "A study of the epidemiology of cancer of the breast," *Cancer* 13 (1960), 559–601.