Chapter 4
Nominal-Level Variables, II

Chapter 3 of *The Measurement of Association* applied permutation statistical methods to measures of association based on Pearson's chi-squared test statistic for two nominal-level (categorical) variables, e.g., Pearson's ϕ^2, Tschuprov's T^2, Cramér's V^2, and Pearson's C. This fourth chapter of *The Measurement of Association* continues the examination of measures of association designed for nominal-level variables, but concentrates on exact and Monte Carlo permutation statistical methods for measures of nominal association that are based on criteria other than Pearson's chi-squared test statistic. First, two asymmetric measures of nominal-level association proposed by Goodman and Kruskal in 1954, λ and t, are described [37]. Next, Cohen's unweighted kappa coefficient, κ, provides an introduction to the measurement of agreement, in contrast to measures of association [23]. Also included in Chap. 4 are McNemar's [63] and Cochran's [22] Q tests that measure the degree to which response measurements change over time, Leik and Gove's [52] d_N^c measure of nominal association, and a solution to the matrix occupancy problem proposed by Mielke and Siddiqui [68]. Fisher's [32] exact probability test is the iconic permutation test for contingency tables. While Fisher's exact test is typically limited to 2×2 contingency tables, for which it was originally intended, in this chapter Fisher's exact test is extended to 2×c, 3×3, 2×2×2, and other larger contingency tables.

Some measures designed for ordinal-level variables also serve as measures of association for nominal-level variables when $r = 2$ rows and $c = 2$ columns, i.e., a 2×2 contingency table. Other measures were originally designed for 2×2 contingency tables with nominal-level variables. Included in measures of association for 2×2 contingency tables are percentage differences, Yule's Q and Y measures [90], the odds ratio, and Somers' asymmetric measures, d_{yx} and d_{xy} [78]. These measures are more appropriately described and discussed in Chaps. 9 and 10, which are devoted to measures of association for analyzing 2×2 contingency tables, where the level of measurement is often irrelevant.

K. J. Berry et al., *The Measurement of Association*,
https://doi.org/10.1007/978-3-319-98926-6_4

Table 4.1 Notation for a 2×2 contingency table

	A_1	A_2	Total
B_1	n_{11}	n_{12}	R_1
B_2	n_{21}	n_{22}	R_2
Total	C_1	C_2	N

4.1 Hypergeometric Probability Values

Exact permutation statistical methods, especially when applied to contingency tables, are heavily dependent on hypergeometric probability values.[1] In this section, a brief introduction to hypergeometric probability values illustrates their calculation and interpretation. For 2×2 contingency tables, the calculation of hypergeometric probability values is easily demonstrated. Consider the 2×2 contingency table in Table 4.1 where $n_{11}, \ldots, n_{22}$ denote the four cell frequencies, R_1 and R_2 denote the two row marginal frequency totals, C_1 and C_2 denote the two column marginal frequency totals, and

$$N = \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij} \ .$$

Because the contingency table given in Table 4.1 is a 2×2 table and, consequently, has only one degree of freedom, the probability of any one cell frequency constitutes the probability of the entire contingency table. Thus, the hypergeometric point probability value for the cell containing n_{11} is given by:

$$p(n_{11}|R_1, C_1, N) = \binom{C_1}{n_{11}}\binom{C_2}{n_{12}}\binom{N}{R_1}^{-1} = \binom{R_1}{n_{11}}\binom{R_2}{n_{21}}\binom{N}{C_1}^{-1} = \frac{R_1!\ R_2!\ C_1!\ C_2!}{N!\ n_{11}!\ n_{12}!\ n_{21}!\ n_{22}!} \ . \tag{4.1}$$

To illustrate the calculation of a hypergeometric point probability value for a 2×2 contingency table, consider the frequency data given in Table 4.2 with $N = 20$ observations. Following Eq. (4.1)

$$p(n_{11}|R_1, C_1, N) = \frac{R_1!\ R_2!\ C_1!\ C_2!}{N!\ n_{11}!\ n_{12}!\ n_{21}!\ n_{22}!} = \frac{11!\ 9!\ 12!\ 8!}{20!\ 9!\ 2!\ 3!\ 6!} = 0.0367 \ .$$

The calculation of hypergeometric probability values for $r \times c$ contingency tables is more complex than for simple 2×2 contingency tables. Consider the

[1] While exact permutation statistical methods for $r \times c$ contingency tables depend on hypergeometric probability values for each of the M possible arrangements of cell frequencies, Monte Carlo resampling permutation statistical methods do not rely on hypergeometric probability values.

Table 4.2 Example 2×2 contingency table

	A_1	A_2	Total
B_1	9	2	11
B_2	3	6	9
Total	12	8	20

Table 4.3 Notation for a 4×3 contingency table

	A_1	A_2	A_3	Total
B_1	n_{11}	n_{12}	n_{13}	R_1
B_2	n_{21}	n_{22}	n_{23}	R_2
B_3	n_{31}	n_{32}	n_{33}	R_3
B_4	n_{41}	n_{42}	n_{43}	R_4
Total	C_1	C_2	C_3	N

4×3 contingency table given in Table 4.3 where $n_{11}, \ldots, n_{43}$ denote the 12 cell frequencies, $R_1, \ldots, R_4$ denote the four row marginal frequency totals, C_1, C_2, and C_3 denote the three column marginal frequency totals, and

$$N = \sum_{i=1}^{4} \sum_{j=1}^{3} n_{ij} \; .$$

When there are only two rows, as in the previous 2×2 example, each column probability value is binomial, but with four rows each column probability value is multinomial. It is well known that a multinomial probability value can be obtained from an inter-connected series of binomial expressions. For example, for column A_1 in Table 4.3,

$$\binom{C_1}{n_{11}}\binom{C_1 - n_{11}}{n_{21}}\binom{C_1 - n_{11} - n_{21}}{n_{31}} = \frac{C_1!}{n_{11}!\,(C_1 - n_{11})!} \times \frac{(C_1 - n_{11})!}{n_{21}!\,(C_1 - n_{11} - n_{21})!} \times \frac{(C_1 - n_{11} - n_{21})!}{n_{31}!\,(C_1 - n_{11} - n_{21} - n_{31})!} = \frac{C_1!}{n_{11}!\,n_{21}!\,n_{31}!\,n_{41}!} \; ,$$

for column A_2 in Table 4.3,

$$\binom{C_2}{n_{12}}\binom{C_2 - n_{12}}{n_{22}}\binom{C_2 - n_{12} - n_{22}}{n_{32}} = \frac{C_2!}{n_{12}!\,(C_2 - n_{12})!} \times \frac{(C_2 - n_{12})!}{n_{22}!\,(C_2 - n_{12} - n_{22})!} \times \frac{(C_2 - n_{12} - n_{22})!}{n_{32}!\,(C_2 - n_{12} - n_{22} - n_{32})!} = \frac{C_2!}{n_{12}!\,n_{22}!\,n_{32}!\,n_{42}!} \; ,$$

for column A_3 in Table 4.3,

$$\binom{C_3}{n_{13}}\binom{C_3-n_{13}}{n_{23}}\binom{C_3-n_{13}-n_{23}}{n_{33}} = \frac{C_3!}{n_{13}!\,(C_3-n_{13})!}$$
$$\times\frac{(C_3-n_{13})!}{n_{23}!\,(C_3-n_{13}-n_{23})!}\times\frac{(C_3-n_{13}-n_{23})!}{n_{33}!\,(C_3-n_{13}-n_{23}-n_{33})!}$$
$$=\frac{C_3!}{n_{13}!\,n_{23}!\,n_{33}!\,n_{43}!}\ ,$$

and for the row marginal frequency distribution in Table 4.3,

$$\binom{N}{R_1}\binom{N-R_1}{R_2}\binom{N-R_1-R_2}{R_3} = \frac{N!}{R_1!\,(N-R_1)!}$$
$$\times\frac{(N-R_1)!}{R_2!\,(N-R_1-R_2)!}\times\frac{(N-R_1-R_2)!}{R_3!\,(N-R_1-R_2-R_3)!}$$
$$=\frac{N!}{R_1!\,R_2!\,R_3!\,R_4!}\ .$$

Thus, for an $r\times c$ contingency table,

$$p(n_{ij}|R_i,C_j,N)=\frac{\left(\prod_{i=1}^{r}R_i!\right)\left(\prod_{j=1}^{c}C_j!\right)}{N!\prod_{i=1}^{r}\prod_{j=1}^{c}n_{ij}!}\ . \tag{4.2}$$

In this form, Eq. (4.2) can easily be generalized to more complex multi-way contingency tables [64].

To illustrate the calculation of a hypergeometric point probability value for an $r\times c$ contingency table, consider the sparse frequency data given in Table 4.4 with $N = 14$ observations. Following Eq. (4.2)

$$p(n_{ij}|R_i,C_j,N)=\frac{\left(\prod_{i=1}^{r}R_i!\right)\left(\prod_{j=1}^{c}C_j!\right)}{N!\prod_{i=1}^{r}\prod_{j=1}^{c}n_{ij}!}$$
$$=\frac{3!\;4!\;3!\;4!\;5!\;5!\;5!}{14!\;2!\;1!\;0!\;0!\;1!\;3!\;0!\;3!\;0!\;3!\;0!\;1!}=0.1903\times10^{-3}\ .$$

Table 4.4 Example 4×3 contingency table

	A_1	A_2	A_3	Total
B_1	2	1	0	3
B_2	0	1	3	4
B_3	0	3	0	3
B_4	3	0	1	4
Total	5	5	4	14

While this section illustrates the calculation of a hypergeometric point probability value, for an exact permutation test of an $r \times c$ contingency table it is necessary to calculate the selected measure of association for the observed cell frequencies and, then, exhaustively enumerate all possible, equally-likely arrangements of the N objects in the rc cells, given the observed marginal frequency distributions.

For each arrangement in the reference set of all permutations of cell frequencies, a measure of association, say, T, is calculated and the exact hypergeometric point probability value, $p(n_{ij}|R_i, C_j, N)$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$, is calculated. If T_{o} denotes the value of the observed test statistic, i.e., measure of association, the exact two-sided probability value of T_{o} is the sum of the hypergeometric point probability values associated with the values of T computed on all possible arrangements of cell frequencies that are equal to or greater than T_{o}.

When the number of possible arrangements of cell frequencies is very large, exact tests are impractical and Monte Carlo permutation statistical methods become necessary. Monte Carlo permutation statistical methods generate a random sample of all possible arrangements of cell frequencies, drawn with replacement, given the observed marginal frequency distributions. The resampling two-sided probability value is simply the proportion of the T values computed on the randomly selected arrangements that are equal to or greater than T_{o}. In the case of Monte Carlo resampling, hypergeometric probability values are not involved—simply the proportion of the values of the measures of association (T values) equal to or greater than the value of the observed measure of association (T_{o}).

4.2 Goodman and Kruskal's λ_a and λ_b Measures

A common problem that many researchers confront is the analysis of a cross-classification table where both variables are categorical, as categorical variables usually do not contain as much information as ordinal- or interval-level variables [54]. As noted in Chap. 3, the usual measures of association based on chi-squared, such as Pearson's ϕ^2, Tschuprov's T^2, Cramér's V^2, and Pearson's C, have proven to be less than satisfactory due to difficulties in interpretation; see, for example, discussions by Agresti and Finlay [2, p. 284], Berry, Martin, and

Table 4.5 Notation for the cross-classification of two categorical variables, A_j for $j = 1, \ldots, c$ and B_i for $i = 1, \ldots, r$

	A				
B	a_1	a_2	$\cdots$	a_c	Total
b_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
b_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
b_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	N

Olson [11], Berry, Johnston, and Mielke [8, 9], Blalock [18, p. 306], Costner [27], Ferguson [30, p. 422], Guilford [42, p. 342], and Wickens [86, p. 226].

In 1954, Leo Goodman and William Kruskal proposed several new measures of association [37].[2] Among the measures were two asymmetric proportional-reduction-in-error (PRE) prediction measures for the analyses of a random sample of two categorical variables: λ_a, for when A was considered to be the dependent variable, and λ_b, for when B was considered to be the dependent variable [37].[3]

Consider an $r \times c$ contingency table such as depicted in Table 4.5, where a_j for $j = 1, \ldots, c$ denotes the c categories for dependent variable A, b_i for $i = 1, \ldots, r$ denotes the r categories for independent variable B, n_{ij} denotes a cell frequency for $i = 1, \ldots, r$ and $j = 1, \ldots, c$, and N denotes the total of cell frequencies in the table. Denote by a dot $(\cdot)$ the partial sum of all rows or all columns, depending on the position of the $(\cdot)$ in the subscript list. If the $(\cdot)$ is in the first subscript position, the sum is over all rows and if the $(\cdot)$ is in the second subscript position, the sum is over all columns. Thus, $n_{i.}$ denotes the marginal frequency total of the ith row, $i = 1, \ldots, r$, summed over all columns, and $n_{.j}$ denotes the marginal frequency total of the jth column, $j = 1, \ldots, c$ summed over all rows.

Given the notation in Table 4.5, let

$$W = \sum_{i=1}^{r} \max(n_{i1}, n_{i2}, \ldots, n_{ic})$$

and

$$X = \max(n_{.1}, n_{.2}, \ldots, n_{.c}) .$$

Then, λ_a, with variable A the dependent variable, is given by:

$$\lambda_a = \frac{W - X}{N - X} .$$

[2] This formative 1954 article by Goodman and Kruskal [37] was followed by three subsequent articles on measures of association for cross-classifications in 1959, 1963, and 1972 [38, 39, 40]

[3] These same statistics, λ_a and λ_b, were independently developed by Louis (Eliyahu) Guttman in 1941 [43].

In like manner, let

$$Y = \sum_{j=1}^{c} \max(n_{1j}, n_{2j}, \ldots, n_{rj})$$

and

$$Z = \max(n_{1.}, n_{2.}, \ldots, n_{r.}) .$$

Then, λ_b, with variable B the dependent variable, is given by:

$$\lambda_b = \frac{Y - Z}{N - Z} .$$

Both λ_a and λ_b are proportional-reduction-in-error (PRE) measures. Consider λ_a and two possible scenarios:

Case 1: Knowledge of only the disjoint categories of dependent variable A.
Case 2: Knowledge of the disjoint categories of variable A, and also knowledge of the disjoint categories of independent variable B.

For Case 1, it is expedient for a researcher to guess the category of dependent variable A that has the largest marginal frequency total (mode), which in this case is $X = \max(n_{.1}, \ldots, n_{.c})$. Then, the probability of error is $N - X$; label these "errors of the first kind" or E_1. For Case 2, it is expedient for a researcher to guess the category of dependent variable A that has the largest cell frequency (mode) in each category of the independent variable B, which in this case is

$$W = \sum_{i=1}^{r} \max(n_{i1}, n_{i2}, \ldots, n_{ic}) .$$

The probability of error is then $N - W$; label these "errors of the second kind" or E_2. Then, λ_a may be expressed as:

$$\lambda_a = \frac{E_1 - E_2}{E_1} = \frac{N - X - (N - W)}{N - X} = \frac{W - X}{N - X} .$$

As noted by Goodman and Kruskal in 1954, a problem was immediately observed with the interpretations of both λ_a and λ_b. Since both measures were based on the modal values of the categories of the independent variable, when the modal values all occurred in the same category of the dependent variable λ_a and λ_b returned results of zero [37, p. 742]. Thus, while λ_a and λ_b were equal to zero under independence, λ_a and λ_b could also be equal to zero for cases other than independence. This made both λ_a and λ_b difficult to interpret; consequently, λ_a and λ_b are seldom found in the contemporary literature. The problem is easy to illustrate

Table 4.6 Example 2×2 contingency table with variables A and B independent

	A_1	A_2	Total
B_1	36	24	60
B_2	24	16	40
Total	60	40	100

Table 4.7 Example 2×2 contingency table with variables A and B not independent

	A_1	A_2	Total
B_1	32	28	60
B_2	28	12	40
Total	60	40	100

with simple 2×2 contingency tables. Consider first the 2×2 contingency table given in Table 4.6 where the cell frequencies indicate independence between variables A and B. For the frequency data given in Table 4.6,

$$W = \sum_{i=1}^{r} \max(n_{i1}, \ldots, n_{ic}) = \max(36, 24) + \max(24, 16) = 36 + 24 = 60 \,,$$

$$X = \max(n_{.1}, \ldots, n_{.c}) = \max(60, 40) = 60 \,,$$

and the observed value of λ_a is

$$\lambda_a = \frac{W - X}{N - X} = \frac{60 - 60}{100 - 60} = 0.00 \,.$$

Now, consider the 2×2 contingency table given in Table 4.7 where the cell frequencies do not indicate independence between variables A and B. For the frequency data given in Table 4.7,

$$W = \sum_{i=1}^{r} \max(n_{i1}, \ldots, n_{ic}) = \max(32, 28) + \max(28, 12) = 32 + 28 = 60 \,,$$

$$X = \max(n_{.1}, \ldots, n_{.c}) = \max(60, 40) = 60 \,,$$

and the observed value of λ_a is

$$\lambda_a = \frac{W - X}{N - X} = \frac{60 - 60}{100 - 60} = 0.00 \,.$$

Finally, consider the 2×2 contingency table given in Table 4.8, where the cell frequencies indicate perfect association between variables A and B. For the

Table 4.8 Example 2×2 contingency table with variables A and B in perfect association

	A_1	A_2	Total
B_1	60	0	60
B_2	0	40	40
Total	60	40	100

frequency data given in Table 4.8,

$$W = \sum_{i=1}^{r} \max(n_{i1}, \ldots, n_{ic}) = \max(60, 0) + \max(0, 40) = 60 + 40 = 100 ,$$

$$X = \max(n_{.1}, \ldots, n_{.c}) = \max(60, 40) = 60 ,$$

and the observed value of λ_a is

$$\lambda_a = \frac{W - X}{N - X} = \frac{100 - 60}{100 - 60} = 1.00 .$$

Thus, as Goodman and Kruskal explained in 1954 [37, p. 742]:

1. λ_a is indeterminate if and only if the population lies in one column; that is, it appears in one category of variable A.
2. Otherwise, the value of λ_a lies between the limits 0 and 1.
3. λ_a is 0 if and only if knowledge of the B classification is of no help in predicting the A classification.
4. λ_a is 1 if and only if knowledge of an object's B category completely specifies its A category, i.e., if each row of the cross-classification table contains at most one non-zero value.
5. In the case of statistical independence, λ_a, when determinate, is zero. The converse need not hold: λ_a may be zero without statistical independence holding.
6. λ_a is unchanged by any permutation of rows or columns.

4.2.1 Example λ_a and λ_b Analyses

For a more realistic application of Goodman and Kruskal's λ_a and λ_b measures of nominal association, consider the 3×4 contingency table given in Table 4.9, where for λ_a

$$W = \sum_{i=1}^{r} \max(n_{i1}, \ldots, n_{ic}) = \max(5, 0, 15, 0) + \max(5, 5, 15, 5)$$
$$+ \max(5, 20, 5, 10) = 15 + 15 + 20 = 50 ,$$

$$X = \max(n_{.1}, \ldots, n_{.c}) = \max(15, 25, 35, 15) = 35 ,$$

Table 4.9 Example 3×4 contingency table for Goodman and Kruskal's λ_a and λ_b

	A_1	A_2	A_3	A_4	Total
B_1	5	0	15	0	20
B_2	5	5	15	5	30
B_3	5	20	5	10	40
Total	15	25	35	15	90

and the observed value of λ_a is

$$\lambda_a = \frac{W - X}{N - X} = \frac{50 - 35}{90 - 35} = 0.2727 \ .$$

The exact probability value of an observed value of λ_a under the null hypothesis is given by the sum of the hypergeometric point probability values associated with values of λ_a equal to or greater than the observed λ_a value. For the frequency data given in Table 4.9, there are only $M = 3{,}453{,}501$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, $\{20, 30, 40\}$ and $\{15, 25, 35, 15\}$, respectively, making an exact permutation analysis possible. The exact upper-tail probability value of the observed λ_a value is $P = 0.2715$, i.e., the sum of the hypergeometric point probability values associated with values of $\lambda_a = 0.2727$ or greater.

The frequency data given in Table 4.9 can also be considered with variable B as the dependent variable. Thus, for λ_b

$$Y = \sum_{j=1}^{c} \max(n_{1j},\ \ldots,\ n_{rj}) = \max(5, 5, 5) + \max(0, 5, 20)$$
$$+ \max(15, 15, 5) + \max(0, 5, 10) = 5 + 20 + 15 + 10 = 50 \ ,$$

$$Z = \max(n_{1.}, \ldots, n_{r.}) = \max(20, 30, 40) = 40 \ ,$$

and the observed value of λ_b is

$$\lambda_b = \frac{Y - Z}{N - Z} = \frac{50 - 40}{90 - 40} = 0.20 \ .$$

For the frequency data given in Table 4.9, there are only $M = 3{,}453{,}501$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, $\{20, 30, 40\}$ and $\{15, 25, 35, 15\}$, respectively, making an exact permutation analysis feasible. The exact upper-tail probability value of the observed λ_b value is $P = 0.7669$, i.e., the sum of the hypergeometric point probability values associated with values of $\lambda_b = 0.20$ or greater.

Table 4.10 Notation for the cross-classification of two categorical variables, A_j for $j = 1, \ldots, c$ and B_i for $i = 1, \ldots, r$

	A				
B	a_1	a_2	$\cdots$	a_c	Total
b_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
b_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
b_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	N

4.3 Goodman and Kruskal's t_a and t_b Measures

As noted, *vide supra*, in 1954 Leo Goodman and William Kruskal proposed several new measures of association. Among the measures was an asymmetric proportional-reduction-in-error (PRE) prediction measure, t_a, for the analysis of a random sample of two categorical variables [37]. Consider two cross-classified unordered polytomies, A and B, with variable A the dependent variable and variable B the independent variable. Table 4.5 on p. 144, replicated in Table 4.10 for convenience, provides notation for the cross-classification, where a_j for $j = 1, \ldots, c$ denotes the c categories for dependent variable A, b_i for $i = 1, \ldots, r$ denotes the r categories for independent variable B, N denotes the total of cell frequencies in the table, $n_{i.}$ denotes a marginal frequency total for the ith row, $i = 1, \ldots, r$, summed over all columns, $n_{.j}$ denotes a marginal frequency total for the jth column, $j = 1, \ldots, c$, summed over all rows, and n_{ij} denotes a cell frequency for $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

Goodman and Kruskal's t_a statistic is a measure of the relative reduction in prediction error where two types of errors are defined. The first type is the error in prediction based solely on knowledge of the distribution of the dependent variable, termed "errors of the first kind" (E_1) and consisting of the expected number of errors when predicting the c dependent variable categories ($a_1, \ldots, a_c$) from the observed distribution of the marginals of the dependent variable ($n_{.1}, \ldots, n_{.c}$). The second type is the error in prediction based on knowledge of the distributions of both the independent and dependent variables, termed "errors of the second kind" (E_2) and consisting of the expected number or errors when predicting the c dependent variable categories ($a_1, \ldots, a_c$) from knowledge of the r independent variable categories ($b_1, \ldots, b_r$).

To illustrate the two error types, consider predicting category a_1 only from knowledge of its marginal distribution, $n_{.1}, \ldots, n_{.c}$. Clearly, $n_{.1}$ out of the N total cases are in category a_1, but exactly which $n_{.1}$ of the N cases is unknown. The probability of incorrectly identifying one of the N cases in category a_1 by chance alone is given by:

$$\frac{N - n_{.1}}{N} \; .$$

Since there are $n_{.1}$ such classifications required, the number of expected incorrect classifications is

$$\frac{n_{.1}(N - n_{.1})}{N}$$

and, for all c categories of variable A, the number of expected errors of the first kind is given by:

$$E_1 = \sum_{j=1}^{c} \frac{n_{.j}(N - n_{.j})}{N} \ .$$

Likewise, to predict $n_{11}, \ldots, n_{1c}$ from the independent category b_1, the probability of incorrectly classifying one of the $n_{1.}$ cases in cell n_{11} by chance alone is

$$\frac{n_{1.} - n_{11}}{n_{1.}} \ .$$

Since there are n_{11} such classifications required, the number of incorrect classifications is

$$\frac{n_{11}(n_{1.} - n_{11})}{n_{1.}}$$

and, for all cr cells, the number of expected errors of the second kind is given by:

$$E_2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{n_{ij}(n_{i.} - n_{ij})}{n_{i.}} \ .$$

Goodman and Kruskal's t_a statistic is then defined as:

$$t_a = \frac{E_1 - E_2}{E_1} \ .$$

An efficient computation form for Goodman and Kruskal's t_a is given by:

$$t_a = \frac{N \displaystyle\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i.}} - \sum_{j=1}^{c} n_{.j}^2}{N^2 - \displaystyle\sum_{j=1}^{c} n_{.j}^2} \ . \tag{4.3}$$

A computed value of t_a indicates the proportional reduction in prediction error given knowledge of the distribution of independent variable B over and above knowledge of only the distribution of dependent variable A. As defined, t_a is a point estimator of Goodman and Kruskal's population parameter τ_a for the population from which the sample of N cases was obtained. If variable B is considered the dependent variable and variable A the independent variable, then Goodman and Kruskal's test statistic t_b and associated population parameter τ_b are analogously defined.

While parameter τ_a norms properly from 0 to 1, possesses a clear and meaningful proportional-reduction-in-error interpretation [27], and is characterized by high intuitive and factorial validity [45], test statistic t_a poses difficulties whenever the null hypothesis posits that $H_0\colon \tau_a = 0$ [61]. The problem is that the sampling distribution of t_a is not asymptotically normal under the null hypothesis $H_0\colon \tau_a = 0$. Consequently, the applicability of Goodman and Kruskal's t_a to typical tests of null hypotheses has been severely circumscribed.

Although t_a was developed by Goodman and Kruskal in 1954, it was not until 1963 that the asymptotic normality for t_a was established and an asymptotic variance was given for t_a, but only for $0 < \tau_a < 1$ [39]. Unfortunately, the asymptotic variance for t_a given in 1963 was later found to be incorrect, and it was not until 1972 that the correct asymptotic variance for t_a was obtained, but again, only for $0 < \tau_a < 1$.

In 1971, Richard Light and Barry Margolin developed R^2, an analysis-of-variance technique for categorical response variables, called CATANOVA for CATegorical ANalysis Of VAriance [55]. Light and Margolin apparently were unaware that R^2 was identical to Goodman and Kruskal's t_a and that they had asymptotically solved the longstanding problem of testing $H_0\colon \tau_a = 0$. The identity between R^2 and t_a was first recognized by Särndal in 1974 [75] and later discussed by Margolin and Light [61], where they showed that $t_a(N-1)(r-1)$ was distributed as chi-squared with $(r-1)(c-1)$ degrees of freedom under $H_0\colon \tau_a = 0$ as $N \to \infty$ [13].

4.3.1 Example Analysis for t_a

Consider the same 3×4 contingency table analyzed with Goodman and Kruskal's λ_a, replicated in Table 4.11 for convenience. Following Eq. (4.3), the observed value of Goodman and Kruskal's t_a is

$$t_a = \frac{N \displaystyle\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i.}} - \sum_{j=1}^{c} n_{.j}^2}{N^2 - \displaystyle\sum_{j=1}^{c} n_{.j}^2}$$

$$= \frac{90\left(\dfrac{5^2}{20} + \dfrac{0^2}{20} + \cdots + \dfrac{10^2}{40}\right) - (15^2 + 25^2 + 35^2 + 15^2)}{90^2 - (15^2 + 25^2 + 35^2 + 15^2)} = 0.1659\ .$$

Table 4.11 Example 3×4 contingency table

	A_1	A_2	A_3	A_4	Total
B_1	5	0	15	0	20
B_2	5	5	15	5	30
B_3	5	20	5	10	40
Total	15	25	35	15	90

The exact probability value of an observed t_a under the null hypothesis is given by the sum of the hypergeometric point probability values associated with values of t_a equal to or greater than the observed value of t_a. For the frequency data given in Table 4.11, there are only $M = 3{,}453{,}501$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, {20, 30, 40} and {15, 25, 35, 15}, respectively, making an exact permutation analysis possible. The exact upper-tail probability value of the observed t_a value is $P = 0.3828$, i.e., the sum of the hypergeometric point probability values associated with values of $t_a = 0.1659$ or greater.

4.3.2 Example Analysis for t_b

Now, consider variable B as the dependent variable. A convenient computing formula for t_b is

$$t_b = \frac{N \displaystyle\sum_{j=1}^{c} \sum_{i=1}^{r} \frac{n_{ij}^2}{n_{.j}} - \sum_{i=1}^{r} n_{i.}^2}{N^2 - \displaystyle\sum_{i=1}^{r} n_{i.}^2} \; .$$

Thus, for the frequency data given in Table 4.11 the observed value of t_b is

$$t_b = \frac{90\left(\dfrac{5^2}{15} + \dfrac{0^2}{25} + \cdots + \dfrac{10^2}{40}\right) - (20^2 + 30^2 + 40^2)}{90^2 - (20^2 + 30^2 + 40^2)} = 0.2022 \; .$$

For the frequency data given in Table 4.11, there are only $M = 3{,}453{,}501$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, {20, 30, 40} and {15, 25, 35, 15}, respectively, making an exact permutation analysis feasible. The exact upper-tail probability value of the observed t_b value is $P =$

0.5187, i.e., the sum of the hypergeometric point probability values associated with values of $t_b = 0.2022$ or greater.

4.4 An Asymmetric Test of Homogeneity

Oftentimes a research question involves determining if the proportions of items in a set of mutually exclusive categories are the same for two or more groups. When independent random samples are drawn from each of $g \geq 2$ groups and then classified into $r \geq 2$ mutually exclusive categories, the appropriate test is a test of homogeneity of the g distributions. In a test of homogeneity, one of the marginal distributions is known prior to collecting the data, i.e., the row or column marginal frequency totals indicating the numbers in each of the g groups. This is termed *product* multinomial sampling, since the sampling distribution is the product of g multinomial distributions and the null hypothesis is that the g multinomial distributions are identical [19, 49, 61].

A test of homogeneity is quite different from a test of independence, where a single sample is drawn and then classified on both variables. In a test of independence, both sets of marginal frequency totals are known only after the data have been collected [62]. This is termed *simple* multinomial sampling, since the sampling distribution is a multinomial distribution [19, 49]. The most widely used test of homogeneity is the Pearson [69] chi-squared test of homogeneity with degrees of freedom given by $df = (r - 1)(g - 1)$. The Pearson chi-squared test of homogeneity tests the null hypothesis that there is no difference in the proportions of subjects in a set of mutually exclusive categories between two or more populations [60].

Pearson's chi-squared test of homogeneity is a symmetrical test, yielding only a single value for an $r \times g$ contingency table. In contrast, an asymmetrical test yields two values depending on which variable is considered to be the dependent variable. As noted by Berkson, if the differences are all in one direction, a symmetrical test such as chi-squared is insensitive to this fact [6, p. 536].

A symmetrical test of homogeneity, by its nature, excludes known information about the data—which variable is the independent variable and which variable is the dependent variable. While it is sometimes necessary to reduce the level of measurement when distributional requirements cannot be met, in general it is not advisable to use a statistical test that discounts important information [29, p. 911]. For example, a researcher should not discard the magnitude of a set of scores and use a signed-ranks test instead of a Fisher–Pitman test, nor should a researcher subsequently ignore the ranks and reduce the analysis to a simple sign test. In the same fashion, given the problem of examining the contingency of two ordered polytomies, the use of a chi-squared-based measure of association does not take into consideration the inherent ordering of the categories [7].

Consider two cross-classified unordered polytomies, A and B, with B the dependent variable. Let $b_1, \ldots, b_r$ represent the $r \geq 2$ categories of the dependent

Table 4.12 Notation for the cross-classification of two categorical variables, A_j for $j = 1, \ldots, g$ and B_i for $i = 1, \ldots, r$

B	A				Total
	a_1	a_2	$\cdots$	a_g	
b_1	n_{11}	n_{12}	$\cdots$	n_{1g}	$n_{1.}$
b_2	n_{21}	n_{22}	$\cdots$	n_{2g}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
b_r	n_{r1}	n_{r2}	$\cdots$	n_{rg}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.g}$	N

variable, $a_1, \ldots, a_g$ represent the $g \geq 2$ categories of the independent variable, n_{ij} indicate the cell frequency in the ith row and jth column, $i = 1, \ldots, r$ and $j = 1, \ldots, g$, and N denote the total sample size. Denote by a dot $(\cdot)$ the partial sum of all rows or all columns, depending on the position of the $(\cdot)$ in the subscript list. If the $(\cdot)$ is in the first subscript position, the sum is over all rows and if the $(\cdot)$ is in the second subscript position, the sum is over all columns. Thus, $n_{1.}, \ldots, n_{r.}$ denotes the marginal frequency totals of row variable B summed over all columns and $n_{.1}, \ldots, n_{.g}$ denotes the marginal frequency totals of column variable A summed over all rows. The cross-classification of variables A and B is displayed in Table 4.12.

Although never advanced as a test of homogeneity, the asymmetrical test t_b, first introduced by Goodman and Kruskal in 1954 [37], is an attractive alternative to the symmetrical chi-squared test of homogeneity. The test statistic is given by:

$$t_b = \frac{N \displaystyle\sum_{j=1}^{g} \sum_{i=1}^{r} \frac{n_{ij}^2}{n_{.j}} - \sum_{i=1}^{r} n_{i.}^2}{N^2 - \displaystyle\sum_{i=1}^{r} n_{i.}^2} ,$$

where B is the dependent variable and the associated population parameter is denoted as τ_b. If variable A is considered the dependent variable, the test statistic is given by:

$$t_a = \frac{N \displaystyle\sum_{i=1}^{r} \sum_{j=1}^{g} \frac{n_{ij}^2}{n_{i.}} - \sum_{j=1}^{g} n_{.j}^2}{N^2 - \displaystyle\sum_{j=1}^{g} n_{.j}^2}$$

and the associated population parameter is τ_a.

Test statistic t_b takes on values between 0 and 1; t_b is 0 if and only if there is homogeneity over the r categories of the dependent variable (B) for all g groups, and t_b is 1 if and only if knowledge of variable A_j for $j = 1, \ldots, g$ completely determines knowledge of variable B_i for $i = 1, \ldots, r$. In like fashion, test statistic t_a is 0 if and only if there is homogeneity over the g categories of the dependent variable (A) for all r groups, and t_a is 1 if and only if knowledge of variable B_i for $i = 1, \ldots, r$ completely determines knowledge of variable A_j for $j = 1, \ldots, g$.

While no general equivalence exists for test statistics t_b, t_a, and χ^2, certain relationships hold among t_b, t_a, and χ^2 under special conditions. If $g = 2$, $\chi^2 = Nt_b$, and if $g > 2$ and $n_{.j} = N/g$ for $j = 1, \ldots, g$, $\chi^2 = N(g-1)t_b$. Similarly, if $r = 2$, $\chi^2 = Nt_a$, and if $r > 2$ and $n_{i.} = N/r$ for $i = 1, \ldots, r$, $\chi^2 = N(r-1)t_a$. It follows that if $r = g = 2$, $t_b = t_a = \chi^2/N$, which is the Pearson mean-squared contingency coefficient, ϕ^2. Finally, as $N \to \infty$, $t_b(N-1)(r-1)$ and $t_a(N-1)(g-1)$ are distributed as chi-squared with $(r-1)(g-1)$ degrees of freedom.

There are three methods to determine the probability value of a computed t_b or t_a test statistic: exact, Monte Carlo resampling, and asymptotic procedures. The following discussions consider only t_b, but the methods are analogous for t_a.

Exact Probability Values Under the null hypothesis, H_0: $\tau_b = 0$, each of the M possible arrangements of the N cases over the rg categories of the contingency table is equally probable with fixed marginal frequency distributions. For each arrangement of the observed data in the reference set of all possible arrangements, the desired test statistic is calculated. The exact probability value of an observed t_b test statistic is the sum of the hypergeometric point probability values associated with values of t_b or greater.

Resampling Probability Values An exact test is computationally not practical except for fairly small samples. An alternative method that avoids the computational demands of an exact test is a resampling permutation approximation. Under the null hypothesis, H_0: $\tau_b = 0$, resampling permutation tests generate and examine a Monte Carlo random subset of all possible, equally-likely arrangements of the observed data. For each randomly selected arrangement of the observed data, the desired test statistic is calculated. The Monte Carlo resampling probability value of an observed t_b test statistic is simply the proportion of the randomly selected values of t_b equal to or greater than the observed value of t_b.

Asymptotic Probability Values Under the null hypothesis, H_0: $\tau_b = 0$, as $N \to \infty$, $t_b(N-1)(g-1)$ is distributed as chi-squared with $(r-1)(g-1)$ degrees of freedom [61]. The asymptotic probability value is the proportion of the appropriate chi-squared distribution equal to or greater than the observed value of $t_b(N-1)(g-1)$.

4.4.1 Example 1

Consider a sample of $N = 80$ seventh grade female students, all from complete families with three children, stratified by Resident Type (Rural, Suburban, or Urban). Each subject is categorized into one of four Personality Characteristics

Table 4.13 Example data set of residence type (A) and personality type (B)

	Residence (A)			
Personality (B)	Rural	Suburb	Urban	Total
Domineering	15	15	15	45
Assertive	15	0	0	15
Submissive	0	15	0	15
Passive	0	0	5	5
Total	30	30	20	80

(Domineering, Assertive, Submissive, or Passive) in a classroom setting by a panel of trained observers. The data are given in Table 4.13. The null hypothesis posits that the proportions of the $r = 4$ observed Personality Types are the same for each of the $g = 3$ Residence Types. Thus, Residence Type (A) is the independent variable and Personality Type (B) is the dependent variable.

For the frequency data given in Table 4.13,

$$t_b = \frac{N \sum_{j=1}^{g} \sum_{i=1}^{r} \frac{n_{ij}^2}{n_{.j}} - \sum_{i=1}^{r} n_{i.}^2}{N^2 - \sum_{i=1}^{r} n_{i.}^2}$$

$$= \frac{80\left(\frac{15^2}{30} + \frac{15^2}{30} + \cdots + \frac{5^2}{20}\right) - (45^2 + 15^2 + 15^2 + 5^2)}{80^2 - (45^2 + 15^2 + 15^2 + 5^2)} = 0.2308\ .$$

There are only $M = 359{,}961$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, $\{45, 15, 15, 5\}$ and $\{30, 30, 20\}$, respectively, making an exact permutation analysis reasonable. The exact upper-tail probability value for the observed value of t_b is $P = 0.1728$, i.e., the sum of the hypergeometric point probability values associated with values of $t_b = 0.2308$ or greater.

In dramatic contrast, the Pearson chi-squared test of homogeneity yields a computed value of $\chi^2 = 66.6667$ for the frequency data given in Table 4.13 and the exact Pearson χ^2 probability value is $P = 0.1699 \times 10^{-12}$. For comparison, the asymptotic Pearson χ^2 probability value based on $(r-1)(g-1) = (4-1)(3-1) = 6$ degrees of freedom is $P = 0.1969 \times 10^{-11}$.

The Pearson χ^2 test of homogeneity is a symmetrical test and does not distinguish between independent and dependent variables, thus excluding important information. Because the Pearson χ^2 test of homogeneity considers both variables A and B, some insight can be gained by calculating a value for t_a. For the frequency data given in Table 4.13,

$$t_a = \frac{N \displaystyle\sum_{i=1}^{r} \sum_{j=1}^{g} \frac{n_{ij}^2}{n_{i.}} - \sum_{j=1}^{g} n_{.j}^2}{N^2 - \displaystyle\sum_{j=1}^{g} n_{.j}^2}$$

$$= \frac{80\left(\dfrac{15^2}{45} + \dfrac{15^2}{45} + \cdots + \dfrac{5^2}{5}\right) - (30^2 + 30^2 + 20^2)}{80^2 - (30^2 + 30^2 + 20^2)} = 0.4286\ ,$$

which is considerably larger than the value for t_b of 0.2308. There are only $M = 359{,}961$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, {45, 15, 15, 5} and {30, 30, 20}, respectively, making an exact permutation analysis feasible. The exact upper-tail probability value for the observed value of t_a is $P = 0.0073$, i.e., the sum of the hypergeometric point probability values associated with values of $t_a = 0.4286$ or greater.

Clearly, the Pearson χ^2 test of homogeneity is detecting the substantial departure from homogeneity of the row proportions. This is reflected in the relatively low probability value for t_a ($P = 0.0073$) where the column variable (A) is considered to be the dependent variable. As the dependent variable of interest is variable B, the Pearson χ^2 test of homogeneity yields a misleading result with an asymptotic probability value of $P = 0.1969 \times 10^{-11}$ compared with the exact probability value for t_b of $P = 0.1728$.

Table 4.14 displays the conditional column proportions obtained from the sample cell frequencies of Table 4.13. In Table 4.14, variable B is the dependent variable and the conditional column proportions are given by $p_{i|j} = n_{ij}/n_{.j}$, e.g., $p_{1|1} = 15/30 = 0.5000$. Table 4.15 displays the conditional row proportions obtained from the sample cell frequencies of Table 4.13. In Table 4.15, variable A is the dependent variable and the conditional row proportions are given by $p_{j|i} = n_{ij}/n_{i.}$, e.g., $p_{1|1} = 15/45 = 0.3333$.

Table 4.14 Conditional column proportions for residence type (A) and personality type (B)

	Residence (A)		
Personality (B)	Rural	Suburb	Urban
Domineering	0.5000	0.5000	0.7500
Assertive	0.5000	0.0000	0.0000
Submissive	0.0000	0.5000	0.0000
Passive	0.0000	0.0000	0.2500
Total	1.0000	1.0000	1.0000

Table 4.15 Conditional row proportions for residence type (A) and personality type (B)

	Residence (A)			
Personality (B)	Rural	Suburb	Urban	Total
Domineering	0.3333	0.3333	0.3333	1.0000
Assertive	1.0000	0.0000	0.0000	1.0000
Submissive	0.0000	1.0000	0.0000	1.0000
Passive	0.0000	0.0000	1.0000	1.0000

Even the most casual inspection of Tables 4.14 and 4.15 reveals the relative homogeneity extant among the proportions in the columns of Table 4.14, compared with the lack of homogeneity among the proportions in the rows of Table 4.15. Compare, for example, the Domineering (0.3333, 0.3333, 0.3333) and Assertive (1.0000, 0.0000, 0.0000) row proportions in Table 4.15. It is this departure from homogeneity in the row proportions that contributes to the low probability value, i.e., $P = 0.1969 \times 10^{-11}$, associated with the Pearson χ^2 test of homogeneity.

4.4.2 Example 2

To clarify the utility of a test of homogeneity based on Goodman and Kruskal's t_b test statistic, consider a simplified example. Suppose that a researcher wishes to conduct a test of homogeneity with respect to Voting Behavior on three categories of Marital Status. The null hypothesis posits that the proportions of the $r = 3$ observed categories of Marital Status (independent variable) are the same for each of the $g = 3$ categories of Voting Behavior (dependent variable). The researcher obtains three independent simple random samples of 80 individuals from each of the three categories of Marital Status—Single, Married, and Divorced—in a local election. Table 4.16 contains the raw frequency data and conditional row proportions where independent variable Marital Status (Single, Married, Divorced) is cross-classified with dependent variable Voting Behavior (Republican, Democrat, Independent).

Table 4.16 Example data set of marital status (A) and voting behavior (B) with row proportions in parentheses

	Voting Behavior (B)			
Marital Status (A)	Republican	Democrat	Independent	Total
Single	50 (0.625)	20 (0.250)	10 (0.125)	80 (1.000)
Married	50 (0.625)	20 (0.250)	10 (0.125)	80 (1.000)
Divorced	50 (0.625)	20 (0.250)	10 (0.125)	80 (1.000)
Total	150	60	30	240

Because the frequency data given in Table 4.16 correspond to the expected values for each of the nine cells, Pearson's chi-squared test of homogeneity is $\chi^2 = 0.00$ with a probability value under the null hypothesis of $P = 1.00$. In contrast, Goodman and Kruskal's test statistic, with variable B (Voting Behavior) the dependent variable is $t_b = 1.00$ with a probability value under the null hypothesis of $P = 0.00$.

4.5 The Measurement of Agreement

The measurement of agreement is a special case of measuring association between two or more variables. A number of statistical research problems require the measurement of agreement, rather than association or correlation. Agreement indices measure the extent to which a set of response measurements are identical to another set, i.e., agree, rather than the extent to which one set of response measurements is a linear function of another set of response measurements, i.e., correlated.

The usual research situation involving a measure of agreement arises when several judges or raters assign objects to a set of disjoint, unordered categories. In 1957, W.S. Robinson published an article in *American Sociological Review* on "The statistical measurement of agreement" [73]. In this formative article, Robinson developed the idea of agreement, as contrasted with correlation, and showed that a simple modification of the intraclass correlation coefficient was an appropriate measure of statistical agreement, which he called A, presumably for agreement [73, p. 20]. Robinson explained that statistical agreement requires that paired values be identical, while correlation requires only that the paired values be linked by some mathematical function [73, p. 19]. Thus, agreement is a more restrictive measure than is correlation. Robinson argued that the distinction between agreement and correlation leads to the conclusion that a logically correct estimate of the reliability of a test is given by the intraclass correlation coefficient rather than the Pearsonian (interclass) correlation coefficient and that the concept of agreement, rather than correlation, is the proper basis of reliability theory [73, p. 18]. The 1957 Robinson article, which was quite mathematical, was followed by a more interpretive article in 1959 in the same journal on "The geometric interpretation of agreement" [74].

A measure of inter-rater agreement should, as a minimum, embody seven basic attributes [16]. First, it is generally agreed that a measure of agreement should be chance corrected, i.e., any agreement coefficient should reflect the amount of agreement in excess of what would be expected by chance. Several researchers have advocated chance-corrected measures of agreement, including Brennan and Prediger [20], Cicchetti, Showalter, and Tyrer [21], Cohen [23], Conger [26], and Krippendorff [50]. Although some investigators have argued against chance-corrected measures of agreement, e.g., Armitage, Blendis, and Smyllie [3] and Goodman and Kruskal [37], supporters of chance-corrected measures of agreement far outweigh detractors.

Second, as noted by Bartko [4, 5], Bartko and Carpenter [5], Krippendorff [50], and Robinson [72], a measure of inter-rater agreement possesses an added advantage if it is directly applicable to the assessment of reliability. Robinson, in particular, was emphatic that reliability could not simply be measured by some function of Pearsonian product-moment correlation, such as in the split-half or test–retest methods, and argued that the concept of agreement should be the basis of reliability theory, not correlation [73, p. 18].

Third, a number of researchers have commented on the simplicity of Euclidean distance for measures of inter-rater agreement, noting that the squaring of differences between scale values is questionable at best, while acknowledging that squared differences allow for familiar interpretations of coefficients [34, 50]. Moreover, Graham and Jackson noted that squaring of differences between values, i.e., quadratic weighting, results in a measure of association, not agreement [41]. Thus, Euclidean distance is a desired property for measures of inter-rater agreement.

Fourth, every measure of agreement should have a statistical base [5]. A measure of agreement without a proper test of significance is severely limited in application to practical research situations. Asymptotic analyses are interesting and useful, under large sample conditions, but often limited in their practical utility when sample sizes are small.

Fifth, a measure of agreement that analyzes multivariate data has a decided advantage over univariate measures of agreement. Thus, if one observer locates a set of objects in an r-dimensional space, a multivariate measure of agreement can ascertain the degree to which a second observer locates the same set of objects in the defined r-dimensional space.

Sixth, a measure of agreement should be able to analyze data at any level of measurement. Cohen's kappa measure of inter-rater agreement is, at the present time, the most widely used measure of agreement. Extensions of Cohen's kappa to incompletely ranked data by Iachan [46] and to continuous categorical data by Conger [26] have been established. An extension of Cohen's kappa measure of agreement to fully ranked ordinal data and to interval data was provided by Berry and Mielke in 1988 [16].

Seventh, a measure of agreement should be able to evaluate information from more than two raters or judges. Fleiss proposed a measure of agreement for multiple raters on a nominal scale [33]. Williams presented a measure that was limited to comparisons of the joint agreement of several raters with another rater singled out as being of special interest [88]. Landis and Koch considered agreement among several raters in terms of a majority opinion [51]. Light focused on an extension of Cohen's [23] kappa measure of inter-rater agreement to multiple raters that was based on the average of all pairwise kappa values [54].

Unfortunately, the measure proposed by Fleiss was dependent on the average proportion of raters who agree on the classification of each observation. The limitation in the measure proposed by Williams appears to be overly restrictive, and the formulation by Landis and Koch becomes computationally prohibitive if either the number of observers or the number of response categories is large. Moreover, the extension of kappa proposed by Fleiss did not reduce to Cohen's kappa when the number of raters was two. Finally, Hubert [44] and Conger [25] provided critical summaries of the problem of extending Cohen's kappa measure of inter-rater agreement to multiple raters for categorical data.

4.5.1 Robinson's Measure of Agreement

An early measure of maximum-corrected agreement was developed by W.S. Robinson in 1957 [73, 74]. Assume that $k = 2$ judges independently rate N objects. Robinson argued that the Pearson product-moment (interclass) correlation calculated between the ratings of two judges was an inadequate measure of agreement because it measures the degree to which the paired values of the two variables are proportional, when expressed as deviations from their means, rather than identical [73, p. 19]. Robinson proposed a new measure of agreement based on the intraclass correlation coefficient that he called A. Consider two sets of ratings such as given in Table 4.17, where there are $N = 3$ pairs of values. Robinson defined A as:

$$A = 1 - \frac{D}{D_{\text{max}}} \,,$$

where D (for Disagreement) is given by:

$$D = \sum_{i=1}^{N} \left(X_{1i} - \bar{X}_i\right)^2 + \sum_{i=1}^{N} \left(X_{2i} - \bar{X}_i\right)^2$$

and

$$\begin{aligned} X_{1i} &= \text{ the value of } X_1 \text{ for the } i\text{th pair of ratings} \,, \\ X_{2i} &= \text{the value of } X_2 \text{ for the } i\text{th pair of ratings} \,, \\ \bar{X}_i &= \text{the mean of } X_1 \text{ and } X_2 \text{ for the } i\text{th pair of ratings} \,. \end{aligned}$$

Robinson noted that, by itself, D is not a very useful measure because it involves the units of X_1 and X_2. To find a relative, rather than an absolute, measure of agreement, Robinson standardized D by its range of possible variation, given by:

$$D_{\text{max}} = \sum_{i=1}^{N} \left(X_{1i} - \bar{X}\right)^2 + \sum_{i=1}^{N} \left(X_{2i} - \bar{X}\right)^2 \,,$$

Table 4.17 Example data for Robinson's *A* coefficient of agreement

X_1	X_2
1	2
3	7
8	12

Table 4.18 Illustration of the calculation of Robinson's *D* coefficient of agreement

X_{1i}	X_{2i}	$\bar{X}_i$	$(X_{1i} - \bar{X}_i)^2$	$(X_{2i} - \bar{X}_i)^2$
1	2	1.50	0.25	0.25
3	7	5.00	4.00	4.00
8	12	10.00	4.00	4.00
12	21		8.25	8.25

where the common mean is given by:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_{1i} + \sum_{i=1}^{N} X_{2i}}{2N} \; .$$

Example

Consider the data listed in Table 4.17 on p. 162 with $N = 3$ paired observations and $k = 2$ sets of ratings, replicated in Table 4.18 for convenience. Then,

$$D = \sum_{i=1}^{N} \left(X_{1i} - \bar{X}_i\right)^2 + \sum_{i=1}^{N} \left(X_{2i} - \bar{X}_i\right)^2 = 8.25 + 8.25 = 16.50 \; .$$

Define the common mean as:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_{1i} + \sum_{i=1}^{N} X_{2i}}{2N} = \frac{12 + 21}{(2)(3)} = 5.50 \; ,$$

then the maximum value of D is illustrated in Table 4.19. The maximum value of D is then

$$D_{\max} = \sum_{i=1}^{N} \left(X_{1i} - \bar{X}\right)^2 + \sum_{i=1}^{N} \left(X_{2i} - \bar{X}\right)^2 = 32.75 + 56.75 = 89.50$$

Table 4.19 Illustration of calculation of Robinson's maximum value of D

X_{1i}	X_{2i}	$\bar{X}_i$	$(X_{1i} - \bar{X}_i)^2$	$(X_{2i} - \bar{X}_i)^2$
1	2	5.50	20.25	12.25
3	7	5.50	6.25	2.25
8	12	5.50	6.25	42.25
12	21		32.75	56.75

Table 4.20 The $M = 6$ possible arrangements of the X_{1i} values, $i = 1, 2, 3$, with associated values of Robinson's D and A

Arrangement	X_1	D	A
1*	1, 3, 8	16.50	0.8156
2	3, 1, 8	26.50	0.7039
3	1, 8, 3	41.50	0.5363
4	3, 8, 1	61.50	0.3128
5	8, 1, 3	76.50	0.1453
6	8, 3, 1	86.50	0.0335

and Robinson's A is

$$A = 1 - \frac{D}{D_{\max}} = 1 - \frac{16.50}{89.50} = 0.8156 \,.$$

The sums,

$$\sum_{i=1}^{N} X_{1i} = 12 \quad \text{and} \quad \sum_{i=1}^{N} X_{2i} = 21,$$

are invariant under permutation. Therefore, $\bar{X} = 5.50$ and $D_{\max} = 89.50$ are also invariant under permutation. Moreover,

$$\sum_{i=1}^{N} \left(X_{1i} - \bar{X}_i\right)^2 = \sum_{i=1}^{N} \left(X_{2i} - \bar{X}_i\right)^2$$

for all arrangements of the observed data. Thus, for an exact permutation analysis, it is only required to calculate either

$$\sum_{i=1}^{N} \left(X_{1i} - \bar{X}_i\right)^2 \quad \text{or} \quad \sum_{i=1}^{N} \left(X_{2i} - \bar{X}_i\right)^2 .$$

In addition, it is only necessary to shuffle either the X_{1i} values or the X_{2i} values, $i = 1, 2, 3$, while holding the X_{2i} or X_{1i} values, respectively, constant.

For the data listed in Table 4.18, there are only $M = 6$ possible, equally-likely arrangements of the observed data. Since $M = 6$ is a very small number, it will be illustrative to list the shuffled X_{1i} values and the associated D and A values in Table 4.20, where the arrangement with the observed values in Table 4.18 is indicated with an asterisk. The exact upper-tail probability of the observed value of

Table 4.21 Example data for the intraclass correlation coefficient

X_{1i}	X_{2i}	X_{1i}^2	X_{2i}^2	$X_{1i}X_{2i}$
1	2	1	4	2
3	7	9	49	21
8	12	64	144	96
2	1	4	1	2
7	3	49	9	21
12	8	144	64	96
33	33	271	271	238

$A = 0.8156$ under the null hypothesis is given by:

$$P(A \geq A_o | H_0) = \frac{\text{number of } A \text{ values } \geq A_o}{M} = \frac{1}{6} = 0.1667\,,$$

where A_o denotes the observed value of Robinson's A. Alternatively,

$$P(D \leq D_o | H_0) = \frac{\text{number of } D \text{ values} \leq D_o}{M} = \frac{1}{6} = 0.1667\,,$$

where D_o denotes the observed value of Robinson's D.

The Intraclass Correlation Coefficient

It is well known that the intraclass correlation coefficient (r_I) between N pairs of observations on two variables is by definition the ordinary Pearson product-moment (interclass) correlation between $2N$ pairs of observations, the first N of which are the original observations, and the second N the original observations with X_{1i} replacing X_{2i} and vice versa for $i = 1, \ldots, N$ [31, Sect. 38]. Thus, the intraclass correlation between the values of X_{1i} and X_{2i} for $i = 1, \ldots, N$ given in Table 4.18 on p. 162 is the Pearson product-moment correlation between the six pairs of values, as illustrated in Table 4.21.

For the data given in Table 4.21 with $N = 6$ pairs of observations, the intraclass correlation coefficient is

$$r_I = r_{12} = \frac{N\sum_{i=1}^{N} X_{1j}X_{2i} - \sum_{i=1}^{N} X_{1i}\sum_{i=1}^{N} X_{2i}}{\sqrt{\left[N\sum_{i=1}^{N} X_{1i}^2 - \left(\sum_{i=1}^{N} X_{1i}\right)^2\right]\left[N\sum_{i=1}^{N} X_{2i}^2 - \left(\sum_{i=1}^{N} X_{2i}\right)^2\right]}}$$

$$= \frac{(6)(238) - (33)(33)}{\sqrt{\left[(6)(271) - (33)^2\right]\left[(6)(271) - (33)^2\right]}} = +0.6313\,. \qquad (4.4)$$

It is obvious from Eq. (4.4) that certain computational simplifications follow from the reversal of the variable values, i.e., the row and column marginal frequency distributions for the new variables are identical and, therefore, the means and variances of the new variables are identical [73, p. 20].

For the case of two variables, the relationships between Robinson's coefficient of agreement and the coefficient of intraclass correlation are given by:

$$r_{\mathrm{I}} = 2A - 1 \quad \text{and} \quad A = \frac{r_{\mathrm{I}} + 1}{2} \ .$$

Thus, in the case of two variables the intraclass correlation is a simple linear function of the coefficient of agreement. For the example data given in Table 4.18 on p. 162,

$$r_{\mathrm{I}} = 2(0.8156) - 1 = 0.6313 \quad \text{and} \quad A = \frac{0.6313 + 1}{2} = 0.8156 \ .$$

For $k > 2$ sets of ratings, the relationships between the intraclass correlation coefficient and Robinson's A are not so simple and are given by:

$$r_{\mathrm{I}} = \frac{kA - 1}{k - 1} \quad \text{and} \quad A = \frac{r_{\mathrm{I}}(k-1) + 1}{k} \ . \tag{4.5}$$

It is apparent from the expressions in Eq. (4.5) that the value of the intraclass coefficient depends not only upon A but also upon k, the number of observations per case. The range of Robinson's A is always from zero to unity regardless of the number of observations. Therefore, comparisons between agreement coefficients based upon different numbers of variables are commensurable [73, p. 22]. The upper limit of the intraclass correlation coefficient is always unity, but its lower limit is $-1/(k-1)$ [31, Sect. 38]. For $k = 2$ variables, the lower limit of r_{I} is -1, but for $k = 3$ variables the lower limit is $-1/2$, for $k = 4$ the lower limit is $-1/3$, for $k = 5$ the lower limit is $-1/4$, and so on.

4.5.2 Scott's π Measure of Agreement

An early measure of chance-corrected agreement was introduced by William Scott in 1955 [76]. Assume that two judges or raters independently classify each of N observations into one of c categories. The resulting classifications can be displayed in a $c \times c$ contingency table, such as the 3×3 table in Table 4.22, with frequencies for cell entries. Denote by a dot $(\cdot)$ the partial sum of all rows or all columns, depending on the position of the $(\cdot)$ in the subscript list. If the $(\cdot)$ is in the first subscript position, the sum is over all rows and if the $(\cdot)$ is in the second subscript position, the sum is over all columns. Thus, $n_{i.}$ denotes the marginal frequency total of the ith row, $i = 1, \ldots, r$, summed over all columns; $n_{.j}$ denotes the marginal frequency total

Table 4.22 Example 3×3 cross-classification (agreement) table with frequencies for cell entries

	Column			
Row	1	2	3	Total
1	n_{11}	n_{12}	n_{13}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

of the jth column, $j = 1, \ldots, c$, summed over all rows; and

$$N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

denotes the table frequency total. In the notation of Table 4.22, Scott's coefficient of agreement for nominal-level data is given by:

$$\pi = \frac{p_\text{o} - p_\text{e}}{1 - p_\text{e}} , \tag{4.6}$$

where

$$p_\text{o} = \frac{1}{N} \sum_{i=1}^{c} n_{ii} \quad \text{and} \quad p_\text{e} = \frac{1}{4N^2} \sum_{k=1}^{c} \left(n_{.k} + n_{k.}\right)^2 .$$

In this configuration, p_o is the observed proportion of observations on which the judges agree, p_e is the proportion of observations for which agreement is expected by chance, $p_\text{o} - p_\text{e}$ is the proportion of agreement beyond that expected by chance, $1 - p_\text{e}$ is the maximum possible proportion of agreement beyond that expected by chance, and Scott's π is the proportion of agreement between the two judges, after chance agreement has been removed.

Example

For an example of Scott's π measure of inter-rater agreement, consider the frequency data given in Table 4.23, where two judges have independently classified $N = 40$ objects into four disjoint categories: A, B, C, and D. For the agreement data given in Table 4.23,

$$p_\text{o} = \frac{1}{N} \sum_{1=1}^{c} n_{ii} = \frac{4+4+4+4}{40} = 0.40 ,$$

$$p_\text{e} = \frac{1}{4N^2} \sum_{k=1}^{c} (n_{.k} + n_{k.})^2 = \frac{1}{(4)(40^2)} \big[(10+10)^2 + (10+10)^2 + (10+10)^2 + (10+10)^2\big] = 0.25 ,$$

Table 4.23 Example 4×4 cross-classification (agreement) table

	Judge 2				
Judge 1	A	B	C	D	Total
A	4	3	2	1	10
B	3	4	1	3	10
C	2	1	4	2	10
D	1	2	3	4	10
Total	10	10	10	10	40

and the observed value of Scott's π is

$$\pi = \frac{p_\mathrm{o} - p_\mathrm{e}}{1 - p_\mathrm{e}} = \frac{0.40 - 0.25}{1 - 0.25} = +0.20 \, , \tag{4.7}$$

indicating 20% agreement above that expected by chance.

The exact probability value of an observed value of Scott's π under the null hypothesis is given by the sum of the hypergeometric point probability values associated with the π values equal to or greater than the observed π value. For the frequency data given in Table 4.23, there are only $M = 5{,}045{,}326$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, {10, 10, 10, 10} and {10, 10, 10, 10}, respectively, making an exact permutation analysis possible. The exact upper-tail probability value of the observed π value is $P = 0.2047$, i.e., the sum of the hypergeometric point probability values associated with values of $\pi = +0.20$ or greater.

While Scott's π is interesting from a historical perspective, π has fallen into desuetude and is no longer found in the current literature. Based as it is on joint proportions, Scott's π makes the assumption that the two judges have the same distribution of responses, as in the example data in Table 4.18 on p. 162 with identical marginal distributions, {10, 10, 10, 10} and {10, 10, 10, 10}. Cohen's κ measure does not make this assumption and, consequently, has emerged as the preferred chance-corrected measure of inter-rater agreement for two judges/raters.

4.5.3 Cohen's κ Measure of Agreement

Currently, the most popular measure of agreement between two judges or raters is the chance-corrected measure of inter-rater agreement first proposed by Jacob Cohen in 1960 and termed kappa [23]. Cohen's kappa measures the magnitude of agreement between $b = 2$ observers on the assignment of N objects to a set of c disjoint, unordered categories. In 1968, Cohen proposed a version of kappa that allowed for weighting of the c categories [24]. Whereas the original (unweighted)

Table 4.24 Example 3×3 cross-classification table with proportions for cell entries

	Column			
Row	1	2	3	Total
1	p_{11}	p_{12}	p_{13}	$p_{1.}$
2	p_{21}	p_{22}	p_{23}	$p_{2.}$
3	p_{31}	p_{32}	p_{33}	$p_{3.}$
Total	$p_{.1}$	$p_{.2}$	$p_{.3}$	$p_{..}$

kappa did not distinguish among magnitudes of disagreement, weighted kappa incorporated the magnitude of each disagreement and provided partial credit for disagreements when agreement was not complete [57]. The usual approach is to assign weights to each disagreement pair with larger weights indicating greater disagreement.[4]

In both the unweighted and weighted cases, kappa is equal to +1 when perfect agreement among two or more judges occurs, 0 when agreement is equal to that expected under independence, and negative when agreement is less than expected by chance. Because weighted kappa applies to ordered categories, it is discussed in Chap. 6. Unweighted kappa is discussed here as it is typically used for unordered categorical data.

Assume that two judges or raters independently classify each of N observations into one of c mutually exclusive, exhaustive, unordered categories. The resulting classifications can be displayed in a $c \times c$ cross-classification, such as the 3×3 contingency table in Table 4.24, with proportions for cell entries. Denote by a dot $(\cdot)$ the partial sum of all rows or all columns, depending on the position of the $(\cdot)$ in the subscript list. If the $(\cdot)$ is in the first subscript position, the sum is over all rows and if the $(\cdot)$ is in the second subscript position, the sum is over all columns. Thus, $p_{i.}$ denotes the marginal proportion total of the ith row, $i = 1, \ldots, c$, summed over all columns; $p_{.j}$ denotes the marginal proportion total of the jth column, $j = 1, \ldots, c$, summed over all rows; and $p_{..} = 1.00$. In the notation of Table 4.24, Cohen's unweighted kappa coefficient for nominal-level data is given by:

$$\kappa = \frac{p_{\text{o}} - p_{\text{e}}}{1 - p_{\text{e}}} \,, \tag{4.8}$$

where

$$p_{\text{o}} = \sum_{i=1}^{c} p_{ii} \quad \text{and} \quad p_{\text{e}} = \sum_{i=1}^{c} p_{i.}p_{.i} \;.$$

[4] Some authors prefer to define kappa in terms of agreement weights, instead of disagreement weights, e.g., Fleiss [33] and Vanbelle and Albert [83].

Cohen's kappa can also be defined in terms of raw frequency values, making calculations somewhat more straightforward. Thus,

$$\kappa = \frac{\displaystyle\sum_{i=1}^{c} O_{ii} - \sum_{i=1}^{c} E_{ii}}{\displaystyle N - \sum_{i=1}^{c} E_{ii}} \,,$$

where O_{ii} denotes an observed cell frequency value on the principal diagonal of a $c \times c$ agreement table, E_{ii} denotes an expected cell frequency value on the principal diagonal, and

$$E_{ii} = \frac{n_{i.} n_{.i}}{N} \qquad \text{for } i = 1, \ldots, c \,.$$

In the configuration of Table 4.24, p_{o} is the observed proportion of observations on which the judges agree, p_{e} is the proportion of observations for which agreement is expected by chance, $p_{\mathrm{o}} - p_{\mathrm{e}}$ is the proportion of agreement beyond that expected by chance, $1 - p_{\mathrm{e}}$ is the maximum possible proportion of agreement beyond that expected by chance, and Cohen's kappa test statistic is the proportion of agreement between the two judges, after chance agreement has been removed.

Example 1

To illustrate Cohen's kappa measure of chance-corrected inter-rater agreement, consider the frequency data given in Table 4.25 where two judges have independently classified $N = 5$ objects into $c = 3$ disjoint, unordered categories: A, B, and C. For the agreement data given in Table 4.25,

$$p_{\mathrm{o}} = \sum_{i=1}^{c} p_{ii} = \frac{0}{5} + \frac{2}{5} + \frac{1}{5} = 0.60 \,,$$

$$p_{\mathrm{e}} = \sum_{i=1}^{c} p_{i.} p_{.i} = \left(\frac{1}{5}\right)\left(\frac{1}{5}\right) + \left(\frac{2}{5}\right)\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right)\left(\frac{1}{5}\right) = 0.36 \,,$$

and following Eq. (4.8), the observed value of Cohen's κ is

$$\kappa = \frac{p_{\mathrm{o}} - p_{\mathrm{e}}}{1 - p_{\mathrm{e}}} = \frac{0.60 - 0.36}{1 - 0.36} = +0.3750 \,,$$

indicating approximately 37% agreement above that expected by chance.

Table 4.25 Example 3×3 cross-classification table for Cohen's unweighted kappa

	Judge 2			
Judge 1	A	B	C	Total
A	0	1	0	1
B	0	2	0	2
C	1	0	1	2
Total	1	3	1	5

Table 4.26 Listing of the eight sets of 3×3 cell frequencies with row marginal distribution {1, 2, 2} and column marginal distribution {1, 3, 1}

Table 1			Table 2			Table 3			Table 4		
0	0	1	0	1	0	0	1	0	0	0	1
0	2	0	0	1	1	0	2	0	1	1	0
1	1	0	1	1	0	1	0	1	0	2	0
Table 5			Table 6			Table 7			Table 8		
0	1	0	0	1	0	1	0	0	1	0	0
1	0	1	0	1	1	0	1	1	0	2	0
0	2	0	0	1	1	0	2	0	0	1	1

Table 4.27 Kappa and hypergeometric probability values for the eight 3×3 contingency tables listed in Table 4.26

Table	κ	Probability
8∗	+0.6875	0.2000
3∗	+0.3750	0.1000
1	+0.0625	0.1000
6	+0.0625	0.1000
7	+0.0625	0.1000
2	−0.2500	0.1000
4	−0.2500	0.1000
5	−0.5625	0.2000

The exact probability value of an observed κ value under the null hypothesis is given by the sum of the hypergeometric point probability values associated with the κ values equal to or greater than the observed κ value. For the frequency data given in Table 4.25, there are only $M = 8$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, $\{1, 2, 2\}$ and $\{1, 3, 1\}$, respectively, making an exact permutation analysis possible. The eight possible arrangements of cell frequencies, given the observed marginal frequency totals, are listed in Table 4.26, where Table 3 of Table 4.26 contains the $N = 5$ observed cell frequencies.

Table 4.27 lists the computed κ values and associated hypergeometric point probability values for the $M = 8$ tables given in Table 4.26, ordered from high to low by the κ values. Only two κ values are equal to or greater than the observed value of $\kappa = +0.3750$, those belonging to Tables 8 and 3 (indicated with asterisks). Thus, the exact upper-tail probability value of the observed κ value is $P = 0.2000 + 0.1000 = 0.3000$, the sum of the hypergeometric point probability

Table 4.28 Example 4×4 cross-classification table

	Judge 2				
Judge 1	A	B	C	D	Total
A	8	4	2	1	15
B	1	7	6	3	17
C	2	4	9	5	20
D	0	1	7	8	16
Total	11	16	24	17	68

values associated with values of $\kappa = +0.3750$ or greater, i.e., $\kappa_8 = +0.6875$ and $\kappa_3 = +0.3750$.

Example 2

For a second, more realistic, example of Cohen's unweighted kappa measure of chance-corrected inter-rater agreement, consider the frequency data given in Table 4.28, where two judges have independently classified $N = 68$ objects into four disjoint, unordered categories: A, B, C, and D. For the agreement data given in Table 4.28,

$$p_{\text{o}} = \sum_{i=1}^{c} p_{ii} = \frac{8}{68} + \frac{7}{68} + \frac{9}{68} + \frac{8}{68} = 0.4706 \ ,$$

$$\begin{aligned} p_{\text{e}} &= \sum_{i=1}^{c} p_{i.}p_{.i} \\ &= \left(\frac{15}{68}\right)\left(\frac{11}{68}\right) + \left(\frac{17}{68}\right)\left(\frac{16}{68}\right) + \left(\frac{20}{68}\right)\left(\frac{24}{68}\right) + \left(\frac{16}{68}\right)\left(\frac{17}{68}\right) \\ &= 0.2571 \ , \end{aligned}$$

and following Eq. (4.8), the observed value of Cohen's κ is

$$\kappa = \frac{p_{\text{o}} - p_{\text{e}}}{1 - p_{\text{e}}} = \frac{0.4706 - 0.2571}{1 - 0.2571} = +0.2873 \ ,$$

indicating approximately 29% agreement above that expected by chance.

The exact probability value of an observed κ value under the null hypothesis is given by the sum of the hypergeometric point probability values associated with κ values equal to or greater than the observed κ value. For the frequency data given in Table 4.28, there are $M = 181{,}260{,}684$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies, given the observed row and column marginal frequency distributions, $\{15, 17, 20, 16\}$ and $\{11, 16, 24, 17\}$, respectively, making an exact permutation analysis feasible. The exact upper-tail probability value of the observed κ value is $P = 0.1098\times10^{-3}$, i.e., the sum of the hypergeometric point probability values associated with values of $\kappa = +0.2873$ or greater.

4.5.4 Application with Multiple Judges

Cohen's κ measure of chance-corrected inter-rater agreement was originally designed for, and limited to, only $b = 2$ judges. In this section, a procedure is introduced for computing unweighted kappa with multiple judges. Although the procedure is appropriate for any number of $c \geq 2$ disjoint, unordered categories and $b \geq 2$ judges, the description of the procedure is confined to $b = 3$ independent judges and the example is limited to $b = 3$ independent judges and $c = 3$ disjoint, unordered categories to simplify presentation.

Consider $b = 3$ judges who independently classify N objects into c disjoint, unordered categories. The classification may be conceptualized as a $c\times c\times c$ contingency table with c rows, c columns, and c slices. Let n_{ijk}, R_i, C_j, and S_k denote the observed cell frequencies and the row, column, and slice marginal frequency totals for $i, j, k = 1, \ldots, c$ and let the frequency total be given by:

$$N = \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} n_{ijk} \; .$$

Cohen's unweighted kappa test statistic for a three-way contingency table is given by:

$$\kappa = 1 - \frac{N^2 \displaystyle\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} n_{ijk}}{\displaystyle\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} R_i C_j S_k} \; , \tag{4.9}$$

where w_{ijk} are disagreement "weights" assigned to each cell for $i, j, k = 1, \ldots, c$. For unweighted kappa, the disagreement weights are given by:

$$w_{ijk} = \begin{cases} 0 & \text{if } i = j = k \; , \\ 1 & \text{otherwise} \; . \end{cases}$$

Given a $c\times c\times c$ contingency table with N objects cross-classified by $b = 3$ independent judges, an exact permutation test involves generating all possible, equally-likely arrangements of the N objects to the c^3 cells, while preserving the marginal frequency distributions. For each arrangement of cell frequencies, the unweighted kappa statistic, κ, and the exact hypergeometric point probability value under the null hypothesis, $p(n_{ijk}|R_i, C_j, S_k, N)$, are calculated, where

$$p(n_{ijk}|R_i, C_j, S_k, N) = \frac{\left(\prod_{i=1}^{c} R_i!\right)\left(\prod_{j=1}^{c} C_j!\right)\left(\prod_{k=1}^{c} S_k!\right)}{(N!)^{b-1}\prod_{i=1}^{c}\prod_{j=1}^{c}\prod_{k=1}^{c} n_{ijk}!} . \tag{4.10}$$

If κ_{o} denotes the value of the observed unweighted kappa test statistic, the exact probability value of κ_{o} under the null hypothesis is given by:

$$P(\kappa_{\text{o}}) = \sum_{l=1}^{M} \Psi_l \left(n_{ijk}|R_i, C_j, S_k, N\right) \ ,$$

where

$$\Psi_l \left(n_{ijk}|R_i, C_j, S_k, N\right) = \begin{cases} p(n_{ijk}|R_i, C_j, S_k, N) & \text{if } \kappa \geq \kappa_{\text{o}} \ , \\ 0 & \text{otherwise} \ , \end{cases}$$

and M denotes the total number of possible, equally-likely cell frequency arrangements in the reference set of all possible arrangements of cell frequencies, given the observed marginal frequency distributions. When M is very large, as is typical with multi-way contingency tables, exact tests are impractical and Monte Carlo resampling procedures become necessary. In such cases, a random sample of the M possible, equally-likely arrangements of cell frequencies provides a comparison of κ test statistics calculated on L random multi-way tables with the κ test statistic calculated on the observed multi-way contingency table.

An efficient Monte Carlo resampling algorithm to generate random cell frequency arrangements for multi-way contingency tables with fixed marginal frequency distributions was developed by Mielke, Berry, and Johnston in 2007 [66, pp. 19–20]. For a three-way contingency table with r rows, c columns, and s slices, the resampling algorithm is given in 12 simple steps.

STEP 1. Construct an $r \times c \times s$ contingency table from the observed data.

STEP 2. Obtain the fixed marginal frequency totals $R_1, \ldots, R_r$, $C_1, \ldots, C_c$, $S_1, \ldots, S_s$, and frequency total N. Set a resampling counter $JL = 0$, and set L equal to the number of samples desired.

STEP 3. Set the resampling counter $JL = JL + 1$.

STEP 4. Set the marginal frequency counters $JR_i = R_i$ for $i = 1, \ldots, r$; $JC_j = C_j$ for $j = 1, \ldots, c$; $JS_k = S_k$ for $k = 1, \ldots, s$, and $M = N$.

STEP 5. Set $n_{ijk} = 0$ for $i = 1, \ldots, r$, $j = 1, \ldots, c$, and $k = 1, \ldots, s$, and set row, column, and slice counters IR, IC, and IS equal to zero.

STEP 6. Create cumulative probability distributions PR_i, PC_j, and PS_k from the adjusted marginal frequency totals JR_i, JC_j, and JS_k for $i = 1, \ldots, r$, $j = 1, \ldots, c$, and $k = 1, \ldots, s$, where

$$PR_1 = JR_1/M \quad \text{and} \quad PR_i = PR_{i-1} + JR_i/M$$

for $i = 1, \ldots, r$,

$$PC_1 = JC_1/M \quad \text{and} \quad PC_j = PC_{j-1} + JC_j/M$$

for $j = 1, \ldots, c$, and

$$PS_1 = JS_1/M \quad \text{and} \quad PS_k = PS_{k-1} + JS_k/M$$

for $k = 1, \ldots, s$.

STEP 7. Generate three uniform pseudorandom numbers U_r, U_c, and U_s over [0, 1) and set row, column, and slice indices $i = j = k = 1$, respectively.

STEP 8. If $U_r \leq PR_i$, then $IR = i$, $JR_i = JR_i - 1$, and go to STEP 9; otherwise, $i = i + 1$ and repeat STEP 8.

STEP 9. If $U_c \leq PC_j$, then $IC = j$, $JC_j = JC_j - 1$, and go to STEP 10; otherwise, $j = j + 1$ and repeat STEP 9.

STEP 10. If $U_s \leq PS_k$, then $IS = k$, $JS_k = JS_k - 1$, and go to STEP 11; otherwise, $k = k + 1$ and repeat STEP 10.

STEP 11. Set $M = M - 1$ and $n_{IR,IC,IS} = n_{IR,IC,IS} + 1$. If $M > 0$, go to STEP 4; otherwise, obtain the required test statistic.

STEP 12. If $JL < L$, go to STEP 3; otherwise, stop.

At the conclusion of the resampling procedure, Cohen's κ, as given in Eq. (4.9) on p. 172, is obtained for each of the L random three-way contingency tables, given fixed marginal frequency distributions. Let κ_o denote the observed value of κ, then under the null hypothesis the resampling approximate probability value for κ_o is given by:

$$P(\kappa_o) = \frac{1}{L} \sum_{l=1}^{L} \Psi_l(\kappa) \ ,$$

where

$$\Psi_l(\kappa) = \begin{cases} 1 & \text{if } \kappa \geq \kappa_o \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Table 4.29 Classification of $N = 93$ objects by three independent judges into one of three disjoint, unordered categories: A, B, or C, with disagreement weights in parentheses

		Judge 3		
Judge 1	Judge 2	A	B	C
A	A	6 (0)	4 (1)	2 (1)
	B	3 (1)	5 (1)	4 (1)
	C	2 (1)	3 (1)	4 (1)
B	A	4 (1)	5 (1)	3 (1)
	B	5 (1)	8 (0)	4 (1)
	C	3 (1)	2 (1)	3 (1)
C	A	1 (1)	3 (1)	4 (1)
	B	3 (1)	2 (1)	2 (1)
	C	1 (1)	2 (1)	5 (0)

4.5.5 *Example Analysis with Multiple Judges*

The calculation of unweighted kappa and the resampling procedure for obtaining a probability value with multiple judges can be illustrated with a sparse data set. Consider $b = 3$ independent judges who classify $N = 93$ objects into one of $c = 3$ disjoint, unordered categories: A, B, or C. Table 4.29 lists the c^3 cross-classified frequencies and corresponding disagreement weights, where the cell disagreement weights are given in parentheses.

For the frequency data listed in Table 4.29, the observed value of kappa is $\kappa = +0.1007$, indicating approximately 10% agreement among the $b = 3$ judges above that expected by chance. If κ_{o} denotes the observed value of κ, the approximate resampling probability value based on $L = 1{,}000{,}000$ random arrangements of the observed data is

$$P(\kappa \geq \kappa_{\text{o}}|H_0) = \frac{\text{number of } \kappa \text{ values } \geq \kappa_{\text{o}}}{L} = \frac{8{,}311}{1{,}000{,}000} = 0.0083 \; .$$

4.6 McNemar's Q Test for Change

In 1947, psychologist Quinn McNemar proposed a test for change that was derived from the matched-pairs t test for proportions [63]. A typical application is to analyze binary responses, coded (0, 1), at $g = 2$ time periods for each of $N \geq 2$ subjects, such as Success and Failure, Yes and No, Agree and Disagree, or Pro and Con. If the four cells are identified as in Table 4.30, then McNemar's test for change is given by:

$$Q = \frac{(B - C)^2}{B + C} \; ,$$

Table 4.30 Notation for a 2×2 cross-classification for McNemar's Q test for change

	Time 2		
Time 1	Pro	Con	Total
Pro	A	B	$A+B$
Con	C	D	$C+D$
Total	$A+C$	$B+D$	N

where $N = A + B + C + D$ and B and C represent the two cells of change, i.e., from Pro to Con and from Con to Pro.

Alternatively, McNemar's Q test can be thought of as a chi-squared goodness-of-fit test with two categories, where the observed frequencies, O_1 and O_2, correspond to cells B and C, respectively, and the expected frequencies, E_1 and E_2, are given by $E_1 = E_2 = (B + C)/2$, i.e., half the subjects are expected to change in one direction (e.g., from Pro to Con) and half in the other direction (e.g., from Con to Pro), under the null hypothesis of no change from Time 1 to Time 2. Let

$$E = \frac{B+C}{2}$$

denote an expected value where, by chance, half of the changes are from Pro to Con and half are from Con to Pro. Then, a chi-squared goodness of fit for the two categories of change is given by:

$$\chi^2 = \frac{(B-E)^2}{E} + \frac{(C-E)^2}{E} = \frac{B^2}{E} + \frac{C^2}{E} + 2E - 2B - 2C \ .$$

Substituting $(B + C)/2$ for E yields

$$\begin{aligned}
&\frac{2B^2}{B+C} + \frac{2C^2}{B+C} + B + C - 2B - 2C \\
&= \frac{2B^2}{B+C} + \frac{2C^2}{B+C} - B - C \\
&= \frac{2B^2 + 2C^2 - B(B+C) - C(B+C)}{B+C} \\
&= \frac{B^2 - 2BC + C^2}{B+C} \\
&= \frac{(B-C)^2}{B+C} \ .
\end{aligned}$$

4.6.1 Example 1

To illustrate McNemar's test for change, consider the frequency data given in Table 4.31, where $N = 50$ objects have been recorded as either Pro or Con on a specified issue at Time 1 and again on the same issue at Time 2. For the frequency data given in Table 4.31, the observed value of McNemar's Q test statistic is

$$Q = \frac{(B-C)^2}{B+C} = \frac{(5-25)^2}{5+25} = 13.3333\ .$$

Alternatively, $O_1 = B = 5$, $O_2 = C = 25$, $E_1 = E_2 = (O_1 + O_2)/2 = (5+25)/2 = 15$, and

$$\chi_1^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(5-15)^2}{15} + \frac{(25-15)^2}{15} = 13.3333\ .$$

The exact probability value of an observed value of Q, under the null hypothesis, is given by the sum of the hypergeometric point probability values associated with the Q values that are equal to or greater than the observed value of Q. For the frequency data listed in Table 4.31, there are only $M = 31$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the two cell frequencies of change, 5 and 25, and only 12 Q values are equal to or greater than the observed value of $Q = 13.3333$.

Since $M = 31$ is a reasonably small number of arrangements, it will be illustrative to list the complete set of Q values and the associated hypergeometric point probability values in Table 4.32, where rows with hypergeometric point probability values associated with Q values equal to or greater than the observed value of Q are indicated with asterisks. The exact upper-tail probability value of the observed value of Q is the sum of the hypergeometric point probability values that are associated with values of $Q = 13.3333$ or greater. Since the distribution of all possible Q values is symmetrical, the exact two-tailed probability value is

$$P = 2\Big(0.1327\times10^{-3} + 0.2552\times10^{-4} + 0.3781\times10^{-5} + 0.4051\times10^{-6} + 0.2794\times10^{-7} + 0.9313\times10^{-9}\Big) = 0.3429\times10^{-3}\ .$$

Table 4.31 Example frequency data for McNemar's test for change with $N = 50$ objects

	Time 2		
Time 1	Pro	Con	Total
Pro	15	5	20
Con	25	5	30
Total	40	10	50

Table 4.32 McNemar Q values and exact hypergeometric point probability values for $M = 31$ possible arrangements of the frequency data given in Table 4.31

Number	B	C	Q	Probability
1∗	0	30	30.0000	0.9313×10^{-9}
2∗	1	29	26.1333	0.2794×10^{-7}
3∗	2	28	22.5333	0.4051×10^{-6}
4∗	3	27	19.2000	0.3781×10^{-5}
5∗	4	26	16.1333	0.2552×10^{-4}
6∗	5	25	13.3333	0.1327×10^{-3}
7	6	24	10.8000	0.5530×10^{-3}
8	7	23	8.5333	0.1896×10^{-2}
9	8	22	6.5333	0.5451×10^{-2}
10	9	21	4.8000	0.1333×10^{-1}
11	10	20	3.3333	0.2798×10^{-1}
12	11	19	2.1333	0.5088×10^{-1}
13	12	18	1.2000	0.8055×10^{-1}
14	13	17	0.5333	0.1115
15	14	16	0.1333	0.1354
16	15	15	0.0000	0.1445
17	16	14	0.1333	0.1354
18	17	13	0.5333	0.1154
19	18	12	1.2000	0.8055×10^{-1}
20	19	11	2.1333	0.5088×10^{-1}
21	20	10	3.3333	0.2798×10^{-1}
22	21	9	4.8000	0.1333×10^{-1}
23	22	8	6.5333	0.5451×10^{-2}
24	23	7	8.5333	0.1896×10^{-2}
25	24	6	10.8000	0.5530×10^{-3}
26∗	25	5	13.3333	0.1327×10^{-3}
27∗	26	4	16.1333	0.2552×10^{-4}
28∗	27	3	19.2000	0.3781×10^{-5}
29∗	28	2	22.5333	0.4051×10^{-6}
30∗	29	1	26.1333	0.2794×10^{-7}
31∗	30	0	30.0000	0.9313×10^{-9}
Sum				1.0000

4.6.2 *Example 2*

For a second example of McNemar's Q test, consider the frequency data given in Table 4.33, where $N = 190$ objects have been recorded as either Pro or Con on a specified issue at Time 1 and again at Time 2. For the frequency data given in Table 4.33, the observed value of McNemar's Q test statistic is

$$Q = \frac{(B - C)^2}{B + C} = \frac{(59 - 37)^2}{59 + 37} = 5.0417\ .$$

Table 4.33 Example frequency data for McNemar's test for change with $N = 190$ objects

	Time 2		
Time 1	Pro	Con	Total
Pro	73	59	132
Con	37	21	58
Total	110	80	190

Alternatively, $O_1 = B = 59$, $O_2 = C = 37$, $E_1 = E_2 = (O_1 + O_2)/2 = (59 + 37)/2 = 48$, and

$$\chi_1^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(59 - 48)^2}{48} + \frac{(37 - 48)^2}{48} = 5.0417\ .$$

The exact probability value of an observed value of Q, under the null hypothesis, is given by the sum of the hypergeometric point probability values associated with the Q values that are equal to or greater than the observed value of Q. For the frequency data listed in Table 4.33, there are only $M = 97$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the two cell frequencies of change, 59 and 37, and only 76 Q values are equal to or greater than the observed value of $Q = 5.0417$. The exact upper-tail probability value of the observed Q value is $P = 0.0315$, i.e., the sum of the hypergeometric point probability values that are associated with values of $Q = 5.0417$ or greater.

4.7 Cochran's Q Test for Change

The ubiquitous dichotomous variable plays a large role and has many applications in research and measurement. Conventionally, a value of one is assigned to each test item that a subject answers correctly and a zero is assigned to each incorrect answer. A common example application occurs when subjects are placed into an experimental situation, observed as to whether or not some specified response is elicited, and scored appropriately [56].

In 1950, William Cochran published an article on "The comparison of percentages in matched samples" [22]. In this brief but formative article, Cochran described a test for equality of matched proportions that is now widely used in educational and psychological research. The matching may be based on the characteristics of different subjects or on the same subjects under different conditions. The Cochran Q test may be viewed as an extension of the McNemar [63] test to three or more treatment conditions. For a typical application, suppose that a sample of $N \geq 2$ subjects is observed in a situation wherein each subject performs individually under each of $k \geq 1$ different experimental conditions. The performance is scored as a success (1) or as a failure (0). The research question evaluates whether the true proportion of successes is constant over the k time periods.

Cochran's Q test for the analysis of k treatment conditions (columns) and N subjects (rows) is given by:

$$Q = \frac{(k-1)\left(k\sum_{j=1}^{k} C_j^2 - A^2\right)}{kA - B} \;, \tag{4.11}$$

where

$$C_j = \sum_{i=1}^{N} x_{ij}$$

is the number of 1s in the jth of k columns,

$$R_i = \sum_{j=1}^{k} x_{ij}$$

is the number of 1s in the ith of N rows,

$$A = \sum_{i=1}^{N} R_i \;, \quad B = \sum_{i=1}^{N} R_i^2 \;,$$

and x_{ij} denotes the cell entry of either 0 or 1 associated with the ith of N rows and the jth of k columns. The null hypothesis stipulates that each of the

$$M = \prod_{i=1}^{N} \binom{k}{R_i}$$

distinguishable arrangements of 1s and 0s within each of the N rows occurs with equal probability, given that the values of $R_1, \ldots, R_N$ are fixed [65].

4.7.1 Example 1

For an example analysis of Cochran's Q test, consider the binary-coded data listed in Table 4.34 consisting of responses (1 or 0) for $N = 10$ subjects evaluated over

Table 4.34 Successes (1) and failures (0) of $N = 10$ subjects on a series of $k = 5$ time periods

	Time					
Subject	1	2	3	4	5	R_i
1	0	1	1	0	0	2
2	1	0	1	0	1	3
3	0	1	1	0	0	2
4	1	1	0	0	0	2
5	1	0	1	1	0	3
6	0	1	1	0	0	2
7	0	1	0	1	0	2
8	0	0	1	0	0	1
9	0	1	0	1	0	2
10	1	1	1	0	0	3
C_j	4	7	7	3	1	22

$k = 5$ time periods, where a 1 denotes success on a prescribed task and a 0 denotes failure. For the binary-coded data listed in Table 4.34,

$$\sum_{j=1}^{k} C_j^2 = 4^2 + 7^2 + 7^2 + 3^2 + 1^2 = 124 \;,$$

$$A = \sum_{i=1}^{N} R_i = 2 + 3 + 2 + 2 + 3 + 2 + 2 + 1 + 2 + 3 = 22 \;,$$

$$B = \sum_{i=1}^{N} R_i^2 = 2^2 + 3^2 + 2^2 + 2^2 + 3^2 + 2^2 + 2^2 + 1^2 + 2^2 + 3^2 = 52 \;,$$

and, following Eq. (4.11) on p. 180, the observed value of Cochran's Q is

$$Q = \frac{(k-1)\left(k\sum_{j=1}^{k} C_j^2 - A^2\right)}{kA - B} = \frac{(5-1)[(5)(124) - 22^2]}{(5)(22) - 52} = 9.3793 \;.$$

For the binary-coded data listed in Table 4.34, there are

$$M = \prod_{i=1}^{N} \binom{k}{R_i} = \binom{5}{1}^1 \binom{5}{2}^6 \binom{5}{3}^3 = (5)(10^6)(10^3) = 5{,}000{,}000{,}000$$

possible, equally-likely arrangements of the observed data, making an exact permutation analysis prohibitive and a Monte Carlo resampling analysis necessary. Based on $L = 1{,}000{,}000$ random arrangements of the observed data, there are 54,486 Q values equal to or greater than the observed value of $Q = 9.3793$. If Q_{o} denotes the observed value of Q, the approximate resampling probability value of the observed data is

$$P\left(Q \geq Q_{\text{o}}|H_0\right) = \frac{\text{number of } Q \text{ values } \geq Q_{\text{o}}}{L} = \frac{54{,}486}{1{,}000{,}000} = 0.0545 \ .$$

For comparison, under the null hypothesis Cochran's Q is approximately distributed as chi-squared with $k - 1$ degrees of freedom. The approximate probability of $Q = 9.3793$ with $k - 1 = 5 - 1 = 4$ degrees of freedom is $P = 0.0523$.

4.7.2 Example 2

For a second example of Cochran's Q test, consider the binary-coded data listed in Table 4.35 consisting of responses (1 or 0) for $N = 9$ subjects evaluated over $k = 3$ time periods, where a 1 indicates success on a prescribed task and a 0 indicates failure. For the binary-coded data listed in Table 4.35,

$$A = \sum_{i=1}^{N} R_i = 1 + 1 + 1 + 1 + 2 + 1 + 2 + 1 + 2 = 12 \ ,$$

$$B = \sum_{i=1}^{N} R_i^2 = 1^2 + 1^2 + 1^2 + 1^2 + 2^2 + 1^2 + 2^2 + 1^2 + 2^2 = 18 \ ,$$

$$\sum_{j=1}^{g} C_j^2 = 4^2 + 7^2 + 1^2 = 66 \ ,$$

Table 4.35 Successes (1) and failures (0) of $N = 9$ subjects on a series of $k = 3$ time periods

	Time			
Subject	1	2	3	R_i
1	0	1	0	1
2	0	1	0	1
3	1	0	0	1
4	0	1	0	1
5	1	0	1	2
6	0	1	0	1
7	1	1	0	2
8	0	1	0	1
9	1	1	0	2
C_j	4	7	1	12

and, following Eq. (4.11) on p. 180, the observed value of Cochran's Q is

$$Q = \frac{(k-1)\left(k\sum_{j=1}^{k} C_j^2 - A^2\right)}{kA - B} = \frac{(3-1)[(3)(66)-12^2]}{(3)(12)-18} = 6.00\ .$$

For the binary-coded data listed in Table 4.35, there are only

$$M = \prod_{i=1}^{N} \binom{k}{R_i} = \binom{3}{1}^6 \binom{3}{2}^3 = (3^6)(3^3) = 19{,}683$$

possible, equally-likely arrangements of the observed data in the reference set of all possible arrangements, making an exact permutation analysis easily accomplished. Based on $M = 19{,}683$ equally-likely, possible arrangements of the observed data, there are 1,056 Q values equal to or greater than the observed value of $Q = 6.00$. If Q_{o} denotes the observed value of Q, the exact upper-tail probability value of the observed data is

$$P\big(Q \geq Q_{\text{o}}|H_0\big) = \frac{\text{number of } Q \text{ values } \geq Q_{\text{o}}}{M} = \frac{1{,}056}{19{,}683} = 0.0537\ .$$

For comparison, under the null hypothesis Cochran's Q is approximately distributed as chi-squared with $k-1$ degrees of freedom. The approximate probability of $Q = 86.00$ with $k-1 = 3-1 = 2$ degrees of freedom is $P = 0.0498$.

4.8 A Measure of Effect Size for Cochran's Q Test

Measures of effect size are increasingly important in reporting research outcomes. The American Psychological Association (APA) has long recommended measures of effect size for articles published in APA journals. For example, as far back as 1994 the 4th edition of the *APA Publication Manual* strongly encouraged reporting measures of effect size in conjunction with probability values. In 1999, the APA Task Force on Statistical Inference, under the direction of Leland Wilkinson, noted that "reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" [87, p. 599]. In 2016, the American Statistical Association (ASA) recommended that measures of effect size be included in future publications in ASA journals [84]. Unfortunately, measures of effect size do not exist for a number of common statistical tests. In this section, a chance-corrected measure of effect size is presented for Cochran's Q test for related proportions [9].

Consider an alternative approach to Cochran's Q test where g treatments are applied independently to each of N subjects with the result of each treatment application recorded as either 1 or 0, representing any suitable dichotomization of the treatment results, i.e., a randomized-block design where the subjects are the blocks and the treatment results are registered as either 1 or 0. Let x_{ij} denote the recorded 1 and 0 response measurements for $i = 1, \ldots, N$ and $j = 1, \ldots, g$. Then, Cochran's test statistic can be defined as:

$$Q = \frac{g-1}{2\sum_{i=1}^{N} p_i(1-p_i)} \left[2\left(\sum_{i=1}^{N} p_i\right)\left(N - \sum_{i=1}^{N} p_i\right) - N(N-1)\,\delta \right] ,$$

where

$$\delta = \left[g\binom{N}{2} \right]^{-1} \sum_{k=1}^{g} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left| x_{ik} - x_{jk} \right| \qquad (4.12)$$

and

$$p_i = \frac{1}{g} \sum_{j=1}^{g} x_{ij} \qquad \text{for } i = 1, \ldots, N ,$$

that is, the proportion of 1 values for the ith of N subjects. Note that in this representation the variation of Q is totally dependent on δ.

In 1979, Acock and Stavig [1] proposed a maximum value for Q given by:

$$Q_{\max} = N(g-1) . \qquad (4.13)$$

Acock and Stavig's maximum value of Q in Eq. (4.13) was employed by Serlin, Carr, and Marascuilo [77] to provide a measure of effect size for Cochran's Q given by:

$$\hat{\eta}_Q^2 = \frac{Q}{Q_{\max}} = \frac{Q}{N(g-1)} ,$$

which standardized Cochran's Q by a maximum value. Unfortunately, the value of $Q_{\max} = N(g-1)$ advocated by Acock and Stavig is achieved only when each subject g-tuple is identical and there is at least one 1 and one 0 in each g-tuple. Thus, $\hat{\eta}_Q^2$ is a "maximum-corrected" measure of effect size and $0 \leq \hat{\eta}_Q^2 \leq 1$ only under these rare conditions.

Assume $0 < p_i < 1$ for $i = 1, \ldots, N$ since $p_i = 0$ and $p_i = 1$ are uninformative. If p_i is constant for $i = 1, \ldots, N$, then $Q_{\max} = N(g-1)$. However, for the vast majority of cases when $p_i \neq p_j$ for $i \neq j$, $Q_{\max} < N(g-1)$. Thus, the routine use of setting $Q_{\max} = N(g-1)$ is problematic and leads to questionable results.

It should also be noted that $\hat{\eta}_Q^2$ is a member of the V family of measures of nominal association based on Cramér's V^2 test statistic given by:

$$V^2 = \frac{\chi^2}{\chi^2_{\max}} = \frac{\chi^2}{N\big[\min(r-1, c-1)\big]} ,$$

where r and c denote the number of rows and columns in an $r \times c$ contingency table [1]. Other members of the V family are Pearson's ϕ^2 for 2×2 contingency tables [70] and Tschuprov's T^2 for $r \times c$ contingency tables where $r = c$ [82]. The difficulties in interpreting V^2 extend to $\hat{\eta}_Q^2$.

As noted in Chap. 3, Wickens observed that Cramér's V^2 lacks an intuitive interpretation other than as a scaling of chi-squared, which limits its usefulness [86, p. 226]. Also, Costner noted that V^2 and other measures based on Pearson's chi-squared lack any interpretation at all for values other than 0 and 1, or the maximum, given the observed marginal frequency distributions [27]. Agresti and Finlay also noted that Cramér's V^2 is very difficult to interpret and recommended other measures [2, p. 284]. Blalock noted that "all measures based on chi square are somewhat arbitrary in nature, and their interpretations leave a lot to be desired…they all give greater weight to those columns or rows having the smallest marginals rather than to those with the largest marginals" [17, 18, p. 306]. Ferguson discussed the problem of using idealized marginal frequencies [30, p. 422], and Guilford noted that measures such as Pearson's ϕ^2, Tschuprov's T^2, and Cramér's V^2 necessarily underestimate the magnitude of association present [42, p. 342]. Berry, Martin, and Olson considered these issues with respect to 2×2 contingency tables [10, 12], and Berry, Johnston, and Mielke discussed in some detail the problems with using Pearson's ϕ^2, Tschuprov's T^2, and Cramér's V^2 as measures of effect size [8]. Since $\hat{\eta}_Q^2$ is simply a special case of Cramér's V^2, it presents the same problems of interpretation. For a detailed assessment of Pearson's ϕ^2, Tschuprov's T^2, and Cramér's V^2, see Chap. 3.

4.8.1 A Chance-Corrected Measure of Effect Size

Chance-corrected measures of effect size have much to commend them over maximum-corrected measures. A chance-corrected measure of effect size is a measure of agreement among the N subjects over g treatments, corrected for chance. A number of researchers have advocated chance-corrected measures of effect size, including Brennan and Prediger [20], Cicchetti, Showalter, and Tyrer [21], Conger [26], and Krippendorff [50]. A chance-corrected measure is zero under chance conditions, unity when agreement among the N subjects is perfect, and negative under conditions of disagreement. Some well-known chance-corrected measures are Scott's coefficient of inter-coder agreement [76], Kendall and Babington Smith's u measure of agreement [48], Cohen's unweighted and weighted coefficients of inter-rater agreement [23, 24], and Spearman's footrule measure [79, 80]. Under certain conditions, Spearman's rank-order correlation coefficient [79, 80] is also a chance-corrected measure of agreement, i.e., when variables x and y consist of ranks from 1 to N with no tied values, or when variable x includes tied values and variable y is a permutation of variable x, then Spearman's rank-order correlation coefficient is both a measure of correlation and a chance-corrected measure of agreement [50, p. 144].

Let x_{ij} denote the (0, 1) response measurements for $i = 1, \ldots, N$ blocks and $j = 1, \ldots, g$ treatments, then

$$\delta = \left[g \binom{N}{2} \right]^{-1} \sum_{k=1}^{g} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left| x_{ik} - x_{jk} \right| .$$

Under the null hypothesis that the distribution of δ assigns equal probability to each of

$$M = \left(g! \right)^{N}$$

possible allocations of the g dichotomous response measurements to the g treatment positions for each of the N subjects, the average value of δ is given by:

$$\mu_\delta = \frac{2}{N(N-1)} \left[\left(\sum_{i=1}^{N} p_i \right) \left(N - \sum_{i=1}^{N} p_i \right) - \sum_{i=1}^{N} p_i (1 - p_i) \right] ,$$

where

$$p_i = \frac{1}{g} \sum_{i=1}^{g} x_{ij} \qquad \text{for } i = 1, \ldots, N .$$

Then, a chance-corrected measure of effect size may be defined as:

$$\Re = 1 - \frac{\delta}{\mu_\delta} .$$

4.8.2 Example

Consider a sample of $N = 6$ psychology graduate students enrolled in a seminar designed to hone skills in assessing patients with various disorders. The seminar includes a clinical aspect whereby actors, provided with different scripts, present symptoms that the students then diagnose. There are $g = 8$ scripts for a variety

Table 4.36 Example data for Cochran's Q test of related proportions with $N = 6$ subjects and $g = 8$ treatments

	Treatment							
Subject	1	2	3	4	5	6	7	8
1	0	1	1	1	0	0	1	0
2	1	1	1	0	0	1	1	1
3	0	1	0	1	1	0	1	1
4	1	1	1	1	0	1	1	1
5	0	1	1	0	0	0	1	1
6	1	1	1	1	0	1	1	0

Table 4.37 Summations for p_i and $p_i(1 - p_i)$ for $i = 1, \ldots, N$

i	p_i	$1 - p_i$	$p_i(1 - p_i)$
1	0.5000	0.5000	0.2500
2	0.7500	0.2500	0.1875
3	0.6250	0.3750	0.2344
4	0.8750	0.1250	0.1094
5	0.5000	0.5000	0.2500
6	0.7500	0.2500	0.1875
Total	4.0000		1.2188

of symptoms including eating disorders, anxiety, depression, oppositional defiant behavior, obsessive-compulsive disorder, and post-traumatic stress disorders, any of which may be presented over the course of the seminar. The "patients" present at random intervals during the semester and the students are assessed as to whether or not the correct diagnosis was made. Table 4.36 lists the data with a 1 (0) indicating a correct (false) diagnosis. For the binary data listed in Table 4.36, Table 4.37 illustrates the calculation of

$$\sum_{i=1}^{N} p_i \quad \text{and} \quad \sum_{i=1}^{N} p_i(1 - p_i) \ ,$$

where

$$p_1 = \frac{1}{g}\sum_{j=1}^{g} x_{1j} = \frac{0+1+1+1+0+0+1+0}{8} = 0.5000 \ ,$$

$$p_2 = \frac{1}{g}\sum_{j=1}^{g} x_{2j} = \frac{1+1+1+0+0+1+1+1}{8} = 0.7500 \ ,$$

$$p_3 = \frac{1}{g}\sum_{j=1}^{g} x_{3j} = \frac{0+1+0+1+1+0+1+1}{8} = 0.6250 \ ,$$

$$p_4 = \frac{1}{g}\sum_{j=1}^{g} x_{4j} = \frac{1+1+1+1+0+1+1+1}{8} = 0.8750\ ,$$

$$p_5 = \frac{1}{g}\sum_{j=1}^{g} x_{5j} = \frac{0+1+1+0+0+0+1+1}{8} = 0.5000\ ,$$

and

$$p_6 = \frac{1}{g}\sum_{j=1}^{g} x_{6j} = \frac{1+1+1+1+0+1+1+0}{8} = 0.7500\ .$$

Table 4.38 illustrates the calculation of the $|x_{ik} - x_{jk}|$ values, $i = 1, \ldots, N-1$ and $j = i+1, \ldots, N$, for Treatments $1, 2, \ldots, 8$. Then,

$$\delta = \left[g \binom{N}{2} \right]^{-1} \sum_{k=1}^{g} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left| x_{ik} - x_{jk} \right|$$
$$= \left[8 \binom{6}{2} \right]^{-1} (9+0+5+8+5+9+0+8) = 0.3667\ ,$$

$$Q = \frac{g-1}{2\displaystyle\sum_{i=1}^{N} p_i(1-p_i)} \left[2 \left(\sum_{i=1}^{N} p_i \right) \left(N - \sum_{i=1}^{N} p_i \right) - N(N-1)\,\delta \right]$$
$$= \frac{8-1}{2(1.2188)} \big[2(4.00)(6-4.00) - 6(6-1)(0.3667) \big] = 14.3590\ ,$$

$$\mu_\delta = \frac{2}{N(N-1)} \left[\left(\sum_{i=1}^{N} p_i \right) \left(N - \sum_{i=1}^{N} p_i \right) - \sum_{i=1}^{N} p_i(1-p_i) \right]$$
$$= \frac{2}{6(6-1)} \big[(4.00)(6-4.00) - 1.2188 \big] = 0.4521\ ,$$

and

$$\Re = 1 - \frac{\delta}{\mu_\delta} = 1 - \frac{0.3667}{0.4521} = +0.1889\ ,$$

Table 4.38 Summation totals for $|x_{ik} - x_{jk}|$ for $k = 1, 2, \ldots, 7, 8$ treatments, $i = 1, \ldots, N-1$, and $j = i + 1, \ldots, N$

	Treatment				
	1	2	$\cdots$	7	8
i	$\lvert x_{i1} - x_{j1}\rvert$	$\lvert x_{i2} - x_{j2}\rvert$	$\cdots$	$\lvert x_{i7} - x_{j7}\rvert$	$\lvert x_{i8} - x_{j8}\rvert$
1	$\lvert 0-1\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 0-1\rvert = 1$
2	$\lvert 0-0\rvert = 0$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 0-1\rvert = 1$
3	$\lvert 0-1\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 0-1\rvert = 1$
4	$\lvert 0-0\rvert = 0$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 0-1\rvert = 1$
5	$\lvert 0-1\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 0-0\rvert = 0$
6	$\lvert 1-0\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$
7	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$
8	$\lvert 1-0\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$
9	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-0\rvert = 1$
10	$\lvert 0-1\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$
11	$\lvert 0-0\rvert = 0$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$
12	$\lvert 0-1\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-0\rvert = 1$
13	$\lvert 1-0\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$
14	$\lvert 1-1\rvert = 0$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-0\rvert = 1$
15	$\lvert 0-1\rvert = 1$	$\lvert 1-1\rvert = 0$	$\cdots$	$\lvert 1-1\rvert = 0$	$\lvert 1-0\rvert = 1$
Total	9	0	$\cdots$	0	8

indicating approximately 19% agreement above that expected by chance. For comparison, the maximum-corrected measure of effect size proposed by Serlin et al. [77] is

$$\hat{\eta}_Q^2 = \frac{Q}{Q_{\max}} = \frac{Q}{N(g-1)} = \frac{14.3590}{6(8-1)} = 0.3419.$$

4.8.3 Advantages of the $\Re$ Measure of Effect Size

Chance-corrected measures of effect size, such as $\Re$, possess distinct advantages in interpretation over maximum-corrected measures of effect size, such as $\hat{\eta}_Q^2$. The problem with $\hat{\eta}_Q^2$ lies in the manner in which $\hat{\eta}_Q^2$ is maximized. The denominator of $\hat{\eta}_Q^2$, $Q_{\max} = N(g-1)$, standardizes the observed value of Q for the sample size (N) and the number of treatments (g). Unfortunately, $N(g-1)$ does not standardize Q for the data on which Q is based, but rather standardizes Q on another unobserved hypothetical set of data.

Consider a simple example with $N = 10$ subjects and $g = 2$ treatments. The observed data are given in Table 4.39, where at Time 1 seven subjects were classified

Table 4.39 Example 2×2 cross-classification for Cochran's Q test for change

	Time 2		
Time 1	Pro	Con	Total
Pro	5	2	7
Con	0	3	3
Total	5	5	10

Table 4.40 Four possible arrangements of the data given in Table 4.39 with fixed observed row and column marginal frequency distributions, {7, 3} and {5, 5}, respectively

	Table A		Table B		Table C		Table D	
	Pro	Con	Pro	Con	Pro	Con	Pro	Con
Pro	5	2	4	3	3	4	2	5
Con	0	3	1	2	2	1	3	0

as Pro and three subjects were classified as Con, and at Time 2 five subjects were classified as Pro and five subjects were classified as Con.

Given the observed data in Table 4.39, only four values of Q are possible. Table 4.40 displays the four possible arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, {7, 3} and {5, 5}, respectively. Table A in Table 4.40 (the observed table) yields $Q = 2.00$, Table B yields $Q = 1.00$, Table C yields $Q = 0.6667$, and Table D yields $Q = 0.50$. Thus, for the observed data given in Table 4.40, $Q = 2.00$ is the maximum value of Q possible, given the observed marginal frequency distributions. Note that $Q_{\text{max}} = N(g-1) = 10(2-1) = 10$ cannot be achieved with these data. For the data given in Table A in Table 4.40 with $Q = 2.00$, $\hat{\eta}^2_Q$ is only 0.20, while $\mathfrak{R} = 1.00$, indicating the proper maximum-corrected effect size.

$\mathfrak{R}$ is a preferred alternative to $\hat{\eta}^2_Q$ as a measure of effect size for two reasons. First, $\mathfrak{R}$ can achieve an effect size of unity for the observed data, while this is often impossible for $\hat{\eta}^2_Q$. Second, $\mathfrak{R}$ is a chance-corrected measure of effect size, meaning that $\mathfrak{R}$ is zero under chance conditions, unity when agreement among the N subjects is perfect, and negative under conditions of disagreement. Therefore, $\mathfrak{R}$ possesses a clear interpretation corresponding to Cohen's coefficient of inter-rater agreement and other chance-corrected measures that are familiar to most researchers. On the other hand, $\hat{\eta}^2_Q$ possesses no meaningful interpretation except for the limiting values of $Q = 0$ and $Q = 1$.

4.9 Leik and Gove's d_N^c Measure of Association

In 1971, Robert Leik and Walter Gove proposed a new measure of nominal association based on pairwise comparisons of differences between observations [53]. Dissatisfied with the existing measures of nominal association, Leik and Gove suggested a proportional-reduction-in-error measure of association that was corrected for the true maximum amount of association, given the observed marginal frequency distributions. The new measure was denoted by d_N^c, where d indicated the index, following other indices such as Somers' d_{yx} and d_{xy}; the subscript N indicated the relevance of d to a nominal dependent variable; and the superscript c indicated that the measure was corrected for the constraints imposed by the marginal frequency distributions [53, p. 287].

Like d_N^c, many measures of association for two variables have been based on pairwise comparisons of differences between observations. Consider two nominal-level variables that have been cross-classified into an $r \times c$ contingency table, where r and c denote the number of rows and columns, respectively. Let $n_{i.}$, $n_{.j}$, and n_{ij} denote the row marginal frequency totals, column marginal frequency totals, and number of objects in the ijth cell, respectively, for $i = 1, \ldots, r$ and $j = 1, \ldots, c$, and let N denote the total number of objects in the $r \times c$ contingency table. If y and x represent the row and column variables, respectively, there are $N(N-1)/2$ pairs of objects in the table that can be partitioned into five mutually exclusive, exhaustive types of pairs: concordant pairs, discordant pairs, pairs tied on variable y but differing on variable x, pairs tied on variable x but differing on variable y, and pairs tied on both variables x and y.

For an $r \times c$ contingency table, concordant pairs (pairs of objects that are ranked in the same order on both variable x and variable y) are given by:

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \right) ,$$

discordant pairs (pairs of objects that are ranked in one order on variable x and the reverse order on variable y) are given by:

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \left(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \right) ,$$

pairs of objects tied on variable x but differing on variable y are given by:

$$T_x = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=j+1}^{c} n_{ik} \right) ,$$

pairs of objects tied on variable y but differing on variable x are given by:

$$T_y = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} \left(\sum_{k=i+1}^{r} n_{kj} \right) ,$$

Table 4.41 Example observed values in a 3×3 contingency table with $N = 100$ observations

	x			
y	x_1	x_2	x_3	Total
y_1	15	5	0	20
y_2	15	25	10	50
y_3	0	10	20	30
Total	30	40	30	100

and pairs of objects tied on both variable x and variable y are given by:

$$T_{xy} = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \left(n_{ij} - 1 \right) .$$

Then,

$$C + D + T_x + T_y + T_{xy} = \frac{N(N-1)}{2} .$$

To illustrate the calculation of Leik and Gove's d_N^c measure, consider first an example 3×3 contingency table, such as given in Table 4.41, where $N = 100$ observations are cross-classified into variable x and variable y, each with $r = c = 3$ categories labeled x_1, x_2, x_3 and y_1, y_2, y_3, respectively.

4.9.1 Observed Contingency Table

For the frequency data given in Table 4.41, consider all possible pairs of observed cell frequency values that have been partitioned into concordant pairs,

$$\begin{aligned} C &= \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \right) \\ &= (15)(25+10+10+20) + (5)(10+20) + (15)(10+20) + (25)(20) \\ &= 2{,}075 , \end{aligned}$$

all discordant pairs of observed cell frequency values,

$$\begin{aligned} D &= \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \left(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \right) \\ &= (0)(15+25+0+10) + (5)(15+0) + (10)(0+10) + (25)(0) \\ &= 175 , \end{aligned}$$

all pairs of observed cell frequency values tied on variable x,

$$T_x = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} \left(\sum_{k=i+1}^{r} n_{kj} \right)$$

$$= (15)(15+0) + (15)(0) + (5)(25+10) + (25)(10) + (0)(10+20) + (10)(20) = 850 \ ,$$

all pairs of observed cell frequency values tied on variable y,

$$T_y = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=j+1}^{c} n_{ik} \right)$$

$$(15)(5+0) + (5)(0) + (15)(25+10) + (25)(10) + (0)(10+20) + (10)(20) = 1{,}050 \ ,$$

and all pairs of observed cell frequency values tied on both variables x and y,

$$T_{xy} = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \left(n_{ij} - 1 \right)$$

$$= \frac{1}{2} \big[(15)(15-1) + (5)(5-1) + (15)(15-1) + (25)(25-1) + (10)(10-1) + (10)(10-1) + (20)(20-1) \big] = 800 \ .$$

Then,

$$C + D + T_x + T_y + T_{xy} = \frac{N(N-1)}{2}$$

and, for the observed frequency data given in Table 4.41,

$$2{,}075 + 175 + 850 + 1{,}050 + 800 = \frac{100(100-1)}{2} = 4{,}950 \ .$$

4.9.2 Expected Contingency Table

Now, consider Table 4.41 expressed as expected cell values, as given in Table 4.42, where an expected value is given by:

$$E_{ij} = \frac{n_{i.} n_{.j}}{N} \qquad \text{for } i = 1, \ldots, r \text{ and } j = 1, \ldots, c \ .$$

Table 4.42 Example expected values in a 3×3 contingency table with $N = 100$ observations

	x			
y	x_1	x_2	x_3	Total
y_1	6	8	6	20
y_2	15	20	15	50
y_3	9	12	9	30
Total	30	40	30	100

For example,

$$E_{11} = \frac{(20)(30)}{100} = 6 \quad \text{and} \quad E_{12} = \frac{(20)(40)}{100} = 8\ .$$

Following Leik and Gove, let a prime (′) indicate a sum of pairs calculated on the expected cell frequency values. Then, for the expected cell frequency values given in Table 4.42, consider all possible pairs of expected values partitioned into concordant pairs,

$$C' = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \right)$$
$$= (6)(20+15+12+9) + (8)(15+9) + (15)(12+9)$$
$$+ (20)(9) = 1{,}023\ ,$$

all discordant pairs of expected cell frequency values,

$$D' = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \left(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \right)$$
$$= (6)(15+20+9+12) + (8)(15+9) + (15)(9+12)$$
$$+ (20)(9) = 1{,}023\ ,$$

all pairs of expected cell frequency values tied on variable x,

$$T'_x = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} \left(\sum_{k=i+1}^{r} n_{kj} \right)$$
$$= (6)(15+9) + (15)(9) + (8)(20+12) + (20)(12)$$
$$+ (6)(15+9) + (15)(9) = 1{,}054\ ,$$

all pairs of expected cell frequency values tied on variable y,

$$T_y' = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=j+1}^{c} n_{ik} \right)$$

$$(6)(8+6) + (8)(6) + (15)(20+15) + (20)(15) + (9)(12+9) + (12)(9) = 1{,}254 \,,$$

and all pairs of expected cell frequency values tied on both variables x and y,

$$\begin{aligned} T_{xy}' &= \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \left(n_{ij} - 1 \right) \\ &= \frac{1}{2} \big[(6)(6-1) + (8)(8-1) + (6)(6-1) + (15)(15-1) \\ &\quad + (20)(20-1) + (15)(15-1) + (9)(9-1) + (12)(12-1) \\ &\quad + (9)(9-1) \big] = 596 \,. \end{aligned}$$

Then,

$$C' + D' + T_x' + T_y' + T_{xy}' = \frac{N(N-1)}{2}$$

and, for the expected frequency data given in Table 4.42,

$$1{,}023 + 1{,}023 + 1{,}054 + 1{,}254 + 596 = \frac{100(100-1)}{2} = 4{,}950 \,.$$

Fortunately, there is a more convenient way to calculate C', D', T_x', T_y', and T_{xy}' without first calculating the expected values. First, given the observed row and column marginal frequency distributions in Table 4.41, $\{20, 50, 30\}$ and $\{30, 40, 30\}$, respectively, calculate the number of pairs of expected cell frequency values tied on both variables x and y,

$$\begin{aligned} T_{xy}' &= \frac{1}{2N^2} \left(\sum_{i=1}^{r} n_{i.}^2 \right) \left(\sum_{j=1}^{c} n_{.j}^2 \right) - \frac{N}{2} \\ &= \frac{1}{2(100^2)} \left(20^2 + 50^2 + 30^2 \right) \left(30^2 + 40^2 + 30^2 \right) - \frac{100}{2} = 596 \,. \end{aligned}$$

Next, calculate the number of pairs of expected cell frequency values tied on variable y,

$$T_y' = \frac{1}{2}\sum_{i=1}^{r} n_{i.}^2 - \frac{N}{2} - T_{xy}' = \frac{1}{2}\left(20^2 + 50^2 + 30^2\right) - \frac{100}{2} - 596 = 1{,}254\ .$$

In like manner, calculate the number of pairs of expected cell frequency values tied on variable x,

$$T_x' = \frac{1}{2}\sum_{j=1}^{c} n_{.j}^2 - \frac{N}{2} - T_{xy}' = \frac{1}{2}\left(30^2 + 40^2 + 30^2\right) - \frac{100}{2} - 596 = 1{,}054\ .$$

Finally, calculate the number of concordant and discordant pairs of expected cell frequency values,

$$\begin{aligned} C' = D' &= \frac{1}{2}\left[\frac{N(N-1)}{2} - T_x' - T_y' - T_{xy}'\right] \\ &= \frac{1}{2}\left[\frac{100(100-1)}{2} - 1054 - 1254 - 596\right] = 1{,}023\ . \end{aligned}$$

It should be noted that C', D', T_x', T_y', and T_{xy}' are all calculated on the observed marginal frequency totals of the observed contingency table, which are invariant under permutation.

4.9.3 Maximized Contingency Table

Test statistic d_N^c is based on three contingency tables: the table of observed values given in Table 4.41, the table of expected values given in Table 4.42, and a table of maximum values to be described next. A contingency table of maximum values is necessary for computing d_N^c. An algorithm for generating an arrangement of cell frequencies in an $r \times c$ contingency table that provides the maximum value of a test statistic was presented in Chap. 3, Sect. 3.2. The algorithm is reproduced here for convenience.

STEP 1: List the observed marginal frequency totals of an $r \times c$ contingency table with empty cell frequencies.

STEP 2: If any pair of marginal frequency totals, one from each set of marginals, are equal to each other, enter that value in the table as n_{ij} and subtract the value from the two marginal frequency totals. For example, if the marginal frequency total for Row 2 is equal to the marginal frequency total for Column 3, enter the marginal frequency total in the table as n_{23} and subtract the value of n_{23} from the marginal frequency totals of Row 2 and Column 3.

Repeat Step 2 until no two marginal frequency totals are equal. If all marginal frequency totals have been reduced to zero, go to Step 5; otherwise, go to Step 3.

Step 3: Find the largest remaining marginal frequency totals in each set and enter the smaller of the two values in n_{ij}. Then, subtract that (smaller) value from the two marginal frequency totals. Go to Step 4.

Step 4: If all marginal frequency totals have been reduced to zero, go to Step 5; otherwise, go to Step 2.

Step 5: Set any remaining n_{ij} values to zero, $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

To illustrate the algorithmic procedure, consider the 3×3 contingency table given in Table 4.41 on p. 192, replicated in Table 4.43 for convenience. Then, the procedure is:

Step 1: List the observed row and column marginal frequency totals, leaving the cell frequencies empty, as in Table 4.44.

Step 2: For the two sets of marginal frequency totals given in Table 4.44, three marginal frequency totals are equal to 30, one for Row 3, one for Column 1, and one for Column 3, i.e., $n_{3.} = n_{.1} = n_{.3} = 30$. Set $n_{31} = 30$ and subtract 30 from the two marginal frequency totals. The adjusted row and column marginal frequency totals are now $\{20, 50, 0\}$ and $\{0, 40, 30\}$, respectively. No other two marginal frequency totals are identical, so go to Step 3.

Step 3: The two largest remaining marginal frequency totals are 50 in Row 2 and 50 in Column 2, i.e., $n_{2.} = 50$ and $n_{.2} = 40$. Set $n_{22} = 40$, the smaller of the two marginal frequency totals, and subtract 40 from the two adjusted marginal frequency totals. The adjusted row and column marginal frequency totals are now $\{20, 10, 0\}$ and $\{0, 0, 30\}$, respectively. Go to Step 4.

Step 4: Not all marginal frequency totals have been reduced to zero, so go to Step 2.

Table 4.43 Example observed values in a 3×3 contingency table with $N = 100$ observations

	x			
y	x_1	x_2	x_3	Total
y_1	15	5	0	20
y_2	15	25	10	50
y_3	0	10	20	30
Total	30	40	30	100

Table 4.44 Empty 3×3 contingency table with observed row marginal frequency distribution $\{20, 50, 30\}$ and observed column marginal frequency distribution $\{30, 40, 30\}$

	x			
y	x_1	x_2	x_3	Total
y_1	–	–	–	20
y_2	–	–	–	50
y_3	–	–	–	30
Total	30	40	30	100

STEP 2: No two marginal frequency totals are identical, so go to STEP 3.

STEP 3: The two largest marginal frequency totals are 20 in Row 1 and 30 in Column 3, i.e., $n_{1.} = 20$ and $n_{.3} = 30$. Set $n_{13} = 20$, the smaller of the two marginal frequency totals and subtract 20 from the two adjusted marginal frequency totals. The adjusted row and column marginal frequency totals are now {0, 10, 0} and {0, 0, 10}. Go to STEP 4.

STEP 4: Not all marginal frequency totals have been reduced to zero, so go to STEP 2.

STEP 2: Two marginal frequency totals are equal to 10, one for Row 2 and one for Column 3, i.e., $n_{2.} = n_{.3} = 10$. Set $n_{23} = 10$ and subtract 10 from the two adjusted marginal frequency totals. The adjusted row and column marginals are now {0, 0, 0} and {0, 0, 0}. All adjusted marginal frequency totals are now zero, so go to STEP 5.

STEP 5: Set any remaining n_{ij} values to zero; in this case, n_{11}, n_{12}, n_{21}, n_{32}, and n_{33} are set to zero.

The completed contingency table is given in Table 4.45. When there are tied values in a marginal distribution, e.g., $n_{.1} = n_{.3} = 30$, there may be alternative cell locations for the non-zero entries, meaning that more than one arrangement of cell frequencies may satisfy the conditions, but the nine cell frequency values {0, 0, 20, 0, 40, 10, 30, 0, 0} must be included in the 3×3 maximized contingency table.

Let a double prime (″) indicate a sum of pairs calculated on the maximized cell frequency values. Then, for the maximized frequency data given in Table 4.45, the number of concordant pairs of maximized cell frequency values is

$$C'' = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \right)$$
$$= (0)(40 + 10 + 0 + 0) + (0)(10 + 0) + (0)(0 + 0) + (20)(0) = 0 \,,$$

Table 4.45 Completed 3×3 contingency table with row marginal frequency distribution {20, 50, 30} and column marginal frequency distribution {30, 40, 30}

	x			
y	x_1	x_2	x_3	Total
y_1	0	0	20	20
y_2	0	40	10	50
y_3	30	0	0	30
Total	30	40	30	100

the number of discordant pairs of maximized cell frequency values is

$$D'' = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \left(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \right)$$

$$= (20)(0+40+30+0) + (0)(0+30) + (10)(30+0) + (40)(30) = 2{,}900\ ,$$

the number of pairs of maximized cell frequency values tied on variable x is

$$T_x'' = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} \left(\sum_{k=i+1}^{r} n_{kj} \right)$$

$$= (0)(0+20) + (0)(20) + (0)(40+10) + (40)(10) + (30)(0+0) + (0)(0) = 400\ ,$$

the number of pairs of maximized cell frequency values tied on variable y is

$$T_y'' = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=j+1}^{c} n_{ik} \right)$$

$$(0)(0+30) + (0)(30) + (0)(40+0) + (40)(0) + (20)(10+0) + (10)(0) = 200\ ,$$

and the number of pairs of maximized cell frequency values tied on both variables x and y is

$$T_{xy}'' = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \left(n_{ij} - 1 \right)$$

$$= \frac{1}{2}\big[(20)(20-1) + (40)(40-1) + (10)(10-1) + (30)(30-1)\big] = 1{,}450\ .$$

Then,

$$C'' + D'' + T_x'' + T_y'' + T_{xy}'' = \frac{N(N-1)}{2}$$

Table 4.46 Values for C, D, T_x, T_y, and T_{xy} obtained from the observed, expected, and maximized frequency tables

	Frequency table		
Pairs	Observed	Expected	Maximized
C	2,075	1,023	0
D	175	1,023	2,900
T_x	850	1,054	200
T_y	1,050	1,254	400
T_{xy}	800	596	1,450
Total	4,950	4,950	4,950

and for the maximized data given in Table 4.45,

$$C'' + D'' + T_x'' + T_y'' + T_{xy}''$$
$$= 0 + 2{,}900 + 200 + 400 + 1{,}450 = \frac{100(100-1)}{2} = 4{,}950\ .$$

Note that the maximized contingency table given in Table 4.45 occurs only when as few cells as possible contain non-zero entries. Thus, either C'' or D'' is maximized and the other is minimized; in this case, $C'' = 0$ is the minimum value of C possible, given the observed marginal frequency distributions, and $D'' = 2{,}900$ is the maximum value of D possible, given the observed marginal frequency distributions. Also, $T_x'' = 200$ and $T_y'' = 400$ are the minimum values of T_x and T_y possible, given the observed marginal frequency distributions. On the other hand, $T_{xy}'' = 1{,}450$ is the maximum value of T_{xy} possible, given the observed marginal frequency distributions.

Table 4.46 summarizes the C, D, T_x, T_y, and T_{xy} values obtained from the observed, expected, and maximized contingency tables.

4.9.4 Calculation of Leik and Gove's d_N^c

Given the observed, expected, and maximized values for C, D, T_x, T_y, and T_{xy} in Table 4.46, errors of the first kind (E_1)—the variation between independence and maximum association—are given by:

$$E_1 = T_y' - T_y'' = 1{,}254 - 400 = 854$$

and errors of the second kind (E_2)—the variation between the observed table and the table of maximum association—are given by:

$$E_2 = T_y - T_y'' = 1{,}050 - 400 = 650\ .$$

Then, in the manner of proportional-reduction-in-error measures of association,

$$d_N^c = \frac{E_1 - E_2}{E_1} = \frac{(T_y' - T_y'') - (T_y - T_y'')}{T_y' - T_y''} = \frac{T_y' - T_y}{T_y' - T_y''}$$
$$= \frac{1{,}254 - 1{,}050}{1{,}254 - 400} = 0.2389\ .$$

Because d_N^c is a symmetrical measure, the number of tied values on variable x can be used in place of the number of tied values on variable y. Thus,

$$d_N^c = \frac{T_x' - T_x}{T_x' - T_x''} = \frac{1{,}054 - 850}{1{,}054 - 200} = 0.2389\ .$$

Alternatively, d_N^c can be defined in terms of the number of values tied on both x and y. Thus,

$$d_N^c = \frac{T_{xy}' - T_{xy}}{T_{xy}' - T_{xy}''} = \frac{596 - 800}{596 - 1{,}450} = 0.2389\ .$$

Because the data are categorical, C and D can be considered as grouped together. Thus,

$$d_N^c = \frac{\left(C' + D'\right) - \left(C + D\right)}{\left(C' + D'\right) - \left(C'' + D''\right)} = \frac{(1{,}023 + 1{,}023) - (2{,}075 + 175)}{(1{,}023 + 1{,}023) - (0 + 2{,}900)}$$
$$= 0.2389\ .$$

Finally,

$$d_N^c = \frac{T_y' - T_y}{T_y' - T_y''} = \frac{T_x' - T_x}{T_x' - T_x''} = \frac{T_{xy}' - T_{xy}}{T_{xy}' - T_{xy}''} = \frac{\left(C' + D'\right) - \left(C + D\right)}{\left(C' + D'\right) - \left(C'' + D''\right)}\ .$$

As noted by Leik and Gove, for an aid in interpreting the relationship between variables x and y, it would be preferable to explicitly determine the number of pairs lost to the marginal requirements of the contingency table. Association can then be defined within those limits, enabling the index to reach unity if cell frequencies are as close to a perfect pattern as the marginal distributions allow [53, p. 286]. Thus, for the frequency data given in Table 4.41 on p. 192, the proportion of cases being considered is

$$1 - \frac{2\left(T_x'' + T_y''\right)}{N(N-1)} = 1 - \frac{2(200 + 600)}{100(100 - 1)} = 0.8384\ .$$

4.9.5 A Permutation Test for d_N^c

Leik and Gove did not provide a standard error for test statistic d_N^c [52]. On the other hand, permutation tests neither assume nor require knowledge of standard errors. Consider the expression

$$d_N^c = \frac{T_y' - T_y}{T_y' - T_y''} \ .$$

It is readily apparent that T_y' and T_y'' are invariant under permutation. Therefore, the probability of d_N^c under the null hypothesis can be determined by the discrete permutation distribution of T_y alone, which is easily obtained from the observed contingency table. Exact permutation statistical methods are highly efficient when only the variable portion of the defined test statistic is calculated on each of the M possible arrangements of the observed data; in this case, T_y.

For the frequency data given in Table 4.41 on p. 192, there are only $M = 96{,}151$ possible, equally-likely arrangements in the reference set of all permutations of cell frequencies given the observed row and column marginal frequency distributions, $\{20, 50, 30\}$ and $\{30, 40, 30\}$, respectively, making an exact permutation analysis feasible. If all $M = 96{,}151$ arrangements occur with equal chance, the exact probability value of d_N^c under the null hypothesis is the sum of the hypergeometric point probability values associated with $d_N^c = 0.2389$ or greater. Based on the underlying hypergeometric probability distribution, the exact upper-tail probability value is $P = 0.1683 \times 10^{-11}$.

4.10 A Matrix Occupancy Problem

In many research situations, it is necessary to examine a sequence of observations on a small group of subjects, where each observation is classified in one of two ways. Suppose, for example, a Success (1) or Failure (0) is recorded for each of $N \geq 2$ subjects on each of $k \geq 2$ tasks. The standard test in such cases is Cochran's Q test, as described in Sect. 4.7.

However, when the number of subjects is small, e.g., $2 \leq N \leq 6$, and the number of treatments is large, e.g., $20 \leq k \leq 400$, an alternative test may be preferable to Cochran's Q test. Such research conditions arise for a number of reasons. First, a long-term panel study is proposed, but few subjects are willing to make a research commitment due to the extended time of the research, or the treatment is either distasteful or time-intensive for the subjects. Second, a longitudinal study begins with an adequate number of subjects, but there is a high drop-out rate and survival analysis cannot be justified. Third, very few subjects satisfy the research protocol. Fourth, the cost of each observation/treatment is expensive for the researcher. Fifth, subjects are very expensive, as in primate studies. Sixth, a pilot study with a small number of subjects may be implemented to establish the validity of the research prior to applying for funding for a larger study.

Consider an $N \times k$ occupancy matrix with N subjects (rows) and k treatment conditions (columns). Let x_{ij} denote the observation of the ith subject ($i = 1, \ldots, N$) in the jth treatment condition ($j = 1, \ldots, k$), where a success is coded 1 and a failure is coded 0. For any subject, a success might result from the treatment administered or it might result from some other cause or a random response, i.e., a false positive. Therefore, a successful treatment response is counted only when all N subjects score a success, i.e., a full column of 1 values. Clearly, this approach does not generalize well to a great number of subjects since it is unrealistic for a large number of subjects to respond in concert. The Q test of Cochran is preferable when N is large.

In 1965, Mielke and Siddiqui presented an exact permutation procedure for the matrix occupancy problem in *Journal of the American Statistical Association* that is appropriate for small samples (N) and a large number of treatments (k) [68]. Let

$$R_i = \sum_{j=1}^{k} x_{ij}$$

for $i = 1, \ldots, N$ denote subject (row) totals, let

$$M = \prod_{i=1}^{N} \binom{k}{R_i}$$

denote the number of equally-likely distinguishable $N \times k$ occupancy matrices in the reference set, under the null hypothesis, and let $v = \min(R_1, \ldots, R_N)$. The null hypothesis stipulates that each of the M distinguishable configurations of 1s and 0s within each of the N rows occurs with equal probability, given that the $R_1, \ldots, R_N$ values are fixed. If U_g is the number of distinct configurations where exactly k treatment conditions (columns) are filled with successes (1s), then

$$U_v = \binom{k}{v} \prod_{i=1}^{N} \binom{k-v}{R_i - v}$$

is the initial value of the recursive relation

$$U_g = \binom{k}{g} \left[\prod_{i=1}^{N} \binom{k-g}{R_i - g} - \sum_{j=g+1}^{v} \binom{k-g}{j-g} \frac{U_j}{\binom{k}{j}} \right] ,$$

where $0 \leq g \leq v - 1$. If $g = 0$, then

$$M = \sum_{g=0}^{v} U_g$$

and the exact probability of observing s or more treatment conditions (columns) completely filled with successes (1s) is given by:

$$P = \frac{1}{M} \sum_{g=s}^{v} U_g \ ,$$

where $0 \leq s \leq v$.

In 1972, Eicker, Siddiqui, and Mielke described extensions to the matrix occupancy problem [28]. In 1974, Mantel [58] observed that the solution to the matrix occupancy problem was also the solution to the "committee problem" considered by Mantel and Pasternack in 1968 [59], Gittelsohn in 1969 [36], Sprott in 1969 [81], and White in 1971 [85]. Whereas the matrix occupancy problem considers N subjects and k treatments, scoring a success by a subject for a specific treatment as a 1 and a failure as a 0, the committee problem considers N committees and k individuals, scoring a 1 if an individual is not a member of a specified committee and 0 otherwise. The committee problem is concerned with the number of individuals belonging to no committees, which is equivalent to the concern of the matrix occupancy problem with the number of treatments associated with successes among all subjects.

4.10.1 Example Analysis

Consider an experiment with $N = 6$ subjects and $k = 8$ treatment conditions, such as given in Table 4.47. For the binary data listed in Table 4.47, the R_i totals are $\{4, 6, 5, 7, 4, 6\}$, the minimum of R_i, $i = 1, \ldots, N$, is $v = 4$, the number of

Table 4.47 Successes (1s) and failures (0s) of $N = 6$ subjects on a series of $k = 8$ treatments

	Treatment								
Subject	1	2	3	4	5	6	7	8	R_i
1	0	1	1	1	0	0	1	0	4
2	1	1	1	0	0	1	1	1	6
3	0	1	0	1	1	0	1	1	5
4	1	1	1	1	0	1	1	1	7
5	0	1	1	0	0	0	1	1	4
6	1	1	1	1	0	1	1	0	6

treatment conditions filled with 1s is $s = 2$ (treatments 2 and 7),

$$\sum_{g=s}^{v} U_g = \sum_{g=2}^{4} U_g = 149{,}341{,}920 + 6{,}838{,}720 + 40{,}320 = 156{,}220{,}960\ ,$$

the number of $N \times k$ occupancy matrices in the reference set of all possible occupancy matrices, under the null hypothesis, is

$$M = \prod_{i=1}^{N} \binom{k}{R_i} = \binom{8}{4}\binom{8}{6}\binom{8}{5}\binom{8}{7}\binom{8}{4}\binom{8}{6}$$

$$= 70 \times 28 \times 56 \times 8 \times 70 \times 28 = 1{,}721{,}036{,}800\ ,$$

and the exact probability of observing $s = 2$ or more treatment conditions completely filled with 1s is

$$P = \frac{1}{M} \sum_{g=s}^{v} U_g = \frac{156{,}220{,}960}{1{,}721{,}036{,}800} = 0.0908\ .$$

It is also possible to define a maximum-corrected measure of effect size as $R = s/k$ that varies between 0 when no treatments (columns) are completely filled with 1s, to a maximum of 1 when all k columns are filled with 1s; in this example,

$$R = \frac{s}{k} = \frac{2}{8} = 0.25.$$

4.11 Fisher's Exact Probability Test

While Fisher's exact probability (FEP) test is, strictly speaking, not a measure of association between two nominal-level variables, it has assumed such importance in the analysis of 2×2 contingency tables that excluding Fisher's exact test from consideration would be a serious omission. That said, however, Fisher's exact probability test provides the probability of association rather than a measure of the strength of association. The Fisher exact probability test was independently developed by R.A. Fisher, Frank Yates, and Joseph Irwin in the early 1930s [32, 47, 89]. Consequently, the test is often referred to as the Fisher–Yates or the Fisher–Irwin exact probability test.[5]

[5] In this research monograph "Fisher exact probability test" is used throughout.

Although the Fisher exact probability test was originally designed for 2×2 contingency tables and is used almost exclusively for this purpose, in this section the test is extended to apply to other contingency tables such as 2×3, 3×3, 3×4, 2×2×2, and other larger contingency tables. For ease of calculation and to avoid large factorial expressions, a recursion procedure with an arbitrary initial value provides an efficient method to obtain exact probability values; for a detailed description of recursion procedures, see Chap. 2, Sects. 2.6.1 and 2.6.2.

4.11.1 Fisher's Exact Analysis of a 2×2 Table

Consider a 2×2 contingency table with N cases, where x_o denotes the observed frequency of any cell and r and c represent the row and column marginal frequency totals, respectively, corresponding to x_o. Table 4.48 illustrates the notation for a 2×2 contingency table.

If $H(x|r,c,N)$ is a recursively defined positive function in which

$$\begin{aligned} H(x|r,c,N) &= D \times \binom{r}{x}\binom{N-r}{c-x}\binom{N}{c}^{-1} \\ &= D \times \frac{r!\,c!\,(N-r)!\,(N-c)!}{N!\,x!\,(r-x)!\,(c-x)!\,(N-r-c+x)!}\ , \end{aligned}$$

where $D > 0$ is an unknown constant, then solving the recursive relation

$$H(x+1|r,c,N) = H(x|r,c,N) \times g(x)$$

yields

$$g(x) = \frac{(r-x)(c-x)}{(x+1)(N-r-c+x+1)}\ .$$

The algorithm may then be employed to enumerate all values of

$$H(x|r,c,N)\ ,$$

Table 4.48 Example notation for a 2×2 contingency table

	A_1	A_2	Total
B_1	x	$r-x$	r
B_2	$c-x$	$N-r-c+x$	$N-r$
Total	c	$N-c$	N

where $a \leq x \leq b$, $a = \max(0, r + c - N)$, $b = \min(r, c)$, and $H(a|N, r, c)$ is initially set to some small positive value [14]. The total over the entire distribution may be found by:

$$T = \sum_{k=a}^{b} H(k|r, c, N) \ .$$

To calculate the probability value of x_o, given the observed marginal frequency distributions, the point probability of the observed table must be determined. This value, designated by $U_2 = H(x|r, c, N)$, is found recursively. Next, the tail of the probability distribution associated with U_2 must be identified. Let

$$U_1 = \begin{cases} H(x_o - 1|r, c, N) & \text{if } x_o > a \ , \\ 0 & \text{if } x_o = a \ , \end{cases}$$

and

$$U_3 = \begin{cases} H(x_o + 1|r, c, N) & \text{if } x_o < b \ , \\ 0 & \text{if } x_o = b \ . \end{cases}$$

If $U_1 > U_3$, U_2 is located in the right tail of the distribution; otherwise, U_2 is defined to be in the left tail of the distribution, and the one-tailed (S_1) and two-tailed (S_2) subtotals may be found by:

$$S_1(x_o|r, c, N) = \sum_{k=a}^{b} K_k H(k|r, c, N)$$

and

$$S_2(x_o|r, c, N) = \sum_{k=a}^{b} L_k H(k|r, c, N) \ ,$$

respectively, where

$$K_k = \begin{cases} 1 & \text{if } U_1 \leq U_3 \text{ and } k \leq x_o \text{ or if } U_1 > U_2 \text{ and } k \geq x_o \ , \\ 0 & \text{otherwise} \ , \end{cases}$$

and

$$L_k = \begin{cases} 1 & \text{if } H(k|r, c, N) \leq U_2 \ , \\ 0 & \text{otherwise} \ , \end{cases}$$

for $k = a, \ldots, b$. The one- and two-tailed exact probability values are then given by:

$$P_1 = \frac{S_1}{T} \quad \text{and} \quad P_2 = \frac{S_2}{T} \ ,$$

respectively.

A 2×2 Contingency Table Example

To illustrate the calculation of Fisher's exact probability test for a fourfold contingency table, consider the 2×2 contingency table given in Table 4.49 with $x_o = 6$, $r = 9$, $c = 8$, $N = 20$,

$$a = \max(0, r + c - N) = \max(0, 9 + 8 - 20) = \max(0, -3) = 0 \ ,$$

$$b = \min(r, c) = \min(9, 8) = 8 \ ,$$

and $b - a + 1 = 8 - 0 + 1 = 9$ possible table configurations in the reference set of all permutations of cell frequencies, given the observed row and column marginal frequency distributions, {9, 11} and {8, 12}, respectively.

Table 4.50 lists the nine possible values of x in the first column. The second column of Table 4.50 lists the exact point probability values for $x = 0, \ldots, 8$ calculated from the conventional hypergeometric probability expression given by:

$$p(x|r, c, N) = \binom{r}{x}\binom{N-r}{c-x}\binom{N}{c}^{-1}$$

$$= \frac{r!\,(N-r)!\,c!\,(N-c)!}{N!\,x!\,(r-x)!\,(c-x)!\,(N-r-c+x)!} \ .$$

Table 4.49 Example 2×2 contingency table

	A_1	A_2	Total
B_1	6	3	9
B_2	2	9	11
Total	8	12	20

Table 4.50 Example of statistical recursion with an arbitrary initial value

x	Probability	$H(x\|r, c, N)$	$H(x\|r, c, N)/T$
0	0.001310	1	0.001310
1	0.023577	18	0.023577
2	0.132032	100.80	0.132032
3	0.308073	235.20	0.308073
4	0.330079	252	0.330079
5	0.165039	126	0.165039
6	0.036675	28	0.036675
7	0.003144	2.40	0.003144
8	0.000071	0.054545	0.000071
Total	1.000000	763.454545	1.000000

The third column of Table 4.50 contains the recursion values where, for $x = 0$, the initial (starting) value is arbitrarily set to 1 for this example analysis. Then,

$$1\left[\frac{(9)(8)}{(1)(4)}\right] = 18 \;,$$

$$18\left[\frac{(8)(7)}{(2)(5)}\right] = 100.80 \;,$$

$$100.80\left[\frac{(7)(6)}{(3)(6)}\right] = 235.20 \;,$$

$$235.20\left[\frac{(6)(5)}{(4)(7)}\right] = 252 \;,$$

$$252\left[\frac{(5)(4)}{(5)(8)}\right] = 126 \;,$$

$$126\left[\frac{(4)(3)}{(6)(9)}\right] = 28 \;,$$

$$28\left[\frac{(3)(2)}{(7)(10)}\right] = 2.40 \;,$$

$$2.40\left[\frac{(2)(1)}{(8)(11)}\right] = 0.054545 \;.$$

The total of $H(x|r, c, N)$ for $x = 0, \ldots, 8$ is

$$\begin{aligned} T &= 1 + 18 + 100.80 + 235.20 + 252 + 126 + 28 + 2.40 + 0.054545 \\ &= 763.454545 \;. \end{aligned}$$

The fourth column of Table 4.50 corrects the entries of the third column by dividing each entry by T. For the frequency data given in Table 4.41 on p. 192,

$$U_2 = H(x_o|r, c, N) = H(6|9, 8, 20) = 28 \ .$$

Because $x_o > a$, i.e., $6 > 1$,

$$U_1 = H(x_o - 1|r, v, N) = H(5|9, 8, 20) = 126$$

and because $x_o < b$, i.e., $6 < 8$,

$$U_3 = H(x_o + 1|r, c, N) = H(7|9, 8, 20) = 2.40 \ .$$

Thus, $U_2 = 28$ is located in the right tail of the distribution since $U_1 > U_3$, i.e., $126 > 2.40$. Then, the one- and two-tailed subtotals are

$$S_1 = 28 + 2.40 + 0.054545 = 30.454545$$

and

$$S_2 = 1 + 18 + 28 + 2.40 + 0.054545 = 49.454545 \ ,$$

respectively, and the one- and two-tailed exact probability values are

$$P_1 = \frac{S_1}{T} = \frac{30.454545}{763.454545} = 0.039890$$

and

$$P_2 = \frac{S_2}{T} = \frac{49.454545}{763.454545} = 0.064777 \ ,$$

respectively.

4.11.2 Larger Contingency Tables

Although Fisher's exact probability test has largely been limited to the analysis of 2×2 contingency tables in the literature, it is not difficult to extend Fisher's exact test to larger contingency tables, although such extensions may be computationally intensive [71, pp. 127–130, 296–298]. Consider an example 2×3 contingency table with N cases, where x_o denotes the observed frequency of the cell in the first row and first column, y_o denotes the observed frequency of the cell in the second row and first column, and r_1, r_2, and c_1 are the observed marginal frequency totals in the first row, second row, and first column, respectively. If $H(x, y)$, given N, r_1,

r_2, and c_1, is a recursively defined positive function, then solving the recursive relation

$$H(x, y+1) = H(x, y) \times g_1(x, y)$$

yields

$$g_1(x, y) = \frac{(c_1 - x - y)(r_2 - y)}{(1 + y)(N - r_1 - r_2 - c_1 + 1 + x + y)} \; . \tag{4.14}$$

If $y = \min(r_2, c_1 - x)$, then $H(x + 1, y) = H(x, y) \times g_2(x, y)$, where

$$g_2(x, y) = \frac{(c_1 - x - y)(r_1 - x)}{(1 + x)(N - r_1 - r_2 - c_1 + 1 + x + y)} \; , \tag{4.15}$$

given that $\max(0, r_1 + r_2 + c_1 - N - x) = 0$. However, if $y = \min(r_2, c_1 - x)$ and $\max(0, r_1 + r_2 + c_1 - N - x) > 0$, then $H(x + 1, y - 1) = H(x, y) \times g_3(x, y)$, where

$$g_3(x, y) = \frac{y(r_1 - x)}{(1 + x)(r_2 + 1 - y)} \; . \tag{4.16}$$

The three recursive expressions given in Eqs. (4.14), (4.15), and (4.16) may be employed to completely enumerate the distribution of $H(x, y)$, where $a \leq x \leq b$, $a = \max(0, r_1 + c_1 - N)$, $b = \min(r_1, c_1)$, $c(x) \leq y \leq d(x)$, $c(x) = \max(0, r_1 + r_2 + c_1 - N + x)$, $d(x) = \min(r_2, c_1 - x)$, and $H[a, c(x)]$ is initially set to some small positive value [15]. The total over the completely enumerated distribution may be found by:

$$T = \sum_{x=a}^{b} \sum_{y=c(x)}^{d(x)} H(x, y) \; .$$

To calculate the probability value of (x_o, y_o), given the observed marginal frequency distributions, the hypergeometric point probability value of the observed 2×3 contingency table must be obtained; this value may also be found recursively. Next, the probability of a result this extreme or more extreme must be found. The subtotal is given by:

$$S = \sum_{x=a}^{b} \sum_{y=c(x)}^{d(x)} J_{x,y} H_{x,y} \; ,$$

Table 4.51 Example 2×3 contingency table

	A_1	A_2	A_3	Total
B_1	5	3	2	10
B_2	8	4	7	19
Total	13	7	9	29

where

$$J_{x,y} = \begin{cases} 1 & \text{if } H(x, y) \leq H(x_o, y_o) \ , \\ 0 & \text{otherwise} \ , \end{cases}$$

for $x = a, \ldots, b$ and $y = c(x), \ldots, d(x)$. The exact probability value for independence associated with the observed cell frequencies, x_o and y_o is given by $P = S/T$.

A 2×3 Contingency Table Example

To illustrate the calculation of Fisher's exact probability test for a 2×3 contingency table, consider the frequency data given in Table 4.51 where $x_o = 5$, $y_o = 3$, $r_1 = 10$, $c_1 = 13$, $c_2 = 7$, and $N = 29$. For the frequency data given in Table 4.51, there are only $M = 59$ arrangements[6] of cell frequencies that are consistent with the observed row and column marginal frequency distributions, {10, 19} and {13, 7, 9}, respectively, and exactly 56 of the arrangements $M = 59$ have hypergeometric point probability values equal to or less than the point probability value of the observed table ($p = 0.8096 \times 10^{-1}$), yielding an exact probability value of $P = 0.6873$. Since the 2×3 table in Table 4.51 has only two degrees of freedom, Table 4.52 lists the $M = 59$ values for n_{11} and n_{12} for each possible arrangement of cell frequencies, given the observed marginal frequency totals, and the associated hypergeometric point probability values. Row 56 contains the observed values of $n_{11} = 5$ and $n_{12} = 3$ indicated by an asterisk.

A 2×6 Contingency Table Example

Fisher's exact probability test is easily extended to any 2×c contingency table. For example, consider the 2×6 contingency table given in Table 4.53 where $v_o = 1$, $w_o = 4$, $x_o = 3$, $y_o = 4$, $z_o = 8$, $r_1 = 6$, $r_2 = 5$, $r_3 = 10$, $r_4 = 9$, $r_5 = 10$,

[6] Although it is relatively simple to calculate the number of possible arrangements of cell frequencies (M) for a 2×2 contingency tables prior to analysis, it is considerably more difficult to calculate M for larger contingency tables; thus, M is usually determined at the conclusion of the analysis. For an algorithm to approximate the number of possible arrangements of cell frequencies, see a 1977 article in *Journal of the American Statistical Association* by Gail and Mantel [35].

Table 4.52 Listing of the $M = 59$ possible cell arrangements for the data given in Table 4.51 with cell frequencies n_{11}, n_{12}, and associated exact hypergeometric point probability values

Table	n_{11}	n_{12}	Probability	Table	n_{11}	n_{12}	Probability
1	0	1	0.3495×10^{-6}	31	6	4	0.2999×10^{-2}
2	1	0	0.6490×10^{-6}	32	7	3	0.2999×10^{-2}
3	0	7	0.4194×10^{-5}	33	8	1	0.4048×10^{-2}
4	0	2	0.9436×10^{-5}	34	4	5	0.6747×10^{-2}
5	10	0	0.1428×10^{-4}	35	2	2	0.6869×10^{-2}
6	3	7	0.1428×10^{-4}	36	2	5	0.6869×10^{-2}
7	1	7	0.2336×10^{-4}	37	7	0	0.7196×10^{-2}
8	2	0	0.3505×10^{-4}	38	5	0	0.8096×10^{-2}
9	2	7	0.3505×10^{-4}	39	3	1	0.8396×10^{-2}
10	1	1	0.4089×10^{-4}	40	3	5	0.1079×10^{-1}
11	0	6	0.4404×10^{-4}	41	6	0	0.1079×10^{-1}
12	0	3	0.6291×10^{-4}	42	7	2	0.1619×10^{-1}
13	0	5	0.1321×10^{-3}	43	2	3	0.1717×10^{-1}
14	0	4	0.1468×10^{-3}	44	2	4	0.1717×10^{-1}
15	4	6	0.2499×10^{-3}	45	5	4	0.2024×10^{-1}
16	9	1	0.2499×10^{-3}	46	7	1	0.2159×10^{-1}
17	9	0	0.3213×10^{-3}	47	6	3	0.2699×10^{-1}
18	1	6	0.3816×10^{-3}	48	4	1	0.3148×10^{-1}
19	1	2	0.4907×10^{-3}	49	3	2	0.3778×10^{-1}
20	3	0	0.5140×10^{-3}	50	3	4	0.4198×10^{-1}
21	3	6	0.8996×10^{-3}	51	4	4	0.4498×10^{-1}
22	2	6	0.9813×10^{-3}	52	6	1	0.5037×10^{-1}
23	2	1	0.9813×10^{-3}	53	5	1	0.5667×10^{-1}
24	5	5	0.1349×10^{-2}	54	3	3	0.6297×10^{-1}
25	8	2	0.1349×10^{-2}	55	6	2	0.6447×10^{-1}
26	1	5	0.1717×10^{-2}	56*	5	3	0.8096×10^{-1}
27	1	3	0.1908×10^{-2}	57	4	2	0.9445×10^{-1}
28	8	0	0.2313×10^{-2}	58	4	3	0.1049
29	1	4	0.2862×10^{-2}	59	5	2	0.1133
30	4	0	0.2999×10^{-2}				

Table 4.53 Example 2×6 contingency table

	A_1	A_2	A_3	A_4	A_5	A_6	Total
B_1	1	4	3	4	8	9	29
B_2	5	1	7	5	2	3	23
Total	6	5	10	9	10	12	52

$c_1 = 29$, and $N = 52$. For the frequency data given in Table 4.53, $M = 33{,}565$ arrangements of cell frequencies are consistent with the observed row and column marginal frequency distributions, {29, 23} and {6, 5, 10, 9, 10, 12}, respectively, and exactly 27,735 of the $M = 33{,}565$ arrangements have hypergeometric point probability values equal to or less than the point probability value of the observed table ($p = 0.1159 \times 10^{-3}$), yielding an exact probability value of $P = 0.0338$.

A 3×3 Contingency Table Example

Fisher's exact probability test can also be applied to larger contingency tables, although calculation time increases substantially as the number of rows and columns increase. In this section, Fisher's exact probability test is applied to a 3×3 contingency table. Consider the 3×3 contingency table given in Table 4.54 where $w_o = 3$, $x_o = 5$, $y_o = 2$, $z_o = 9$, $r_1 = 10$, $r_2 = 14$, $c_1 = 13$, $c_2 = 16$, and $N = 40$. For the frequency data given in Table 4.54, $M = 4{,}818$ arrangements of cell frequencies are consistent with the observed row and column marginal frequency distributions, {10, 14, 16} and {13, 16, 11}, respectively, and exactly 3,935 of the $M = 4{,}818$ arrangements have hypergeometric point probability values equal to or less than the point probability value of the observed table ($p = 0.1273 \times 10^{-4}$), yielding an exact probability value of $P = 0.0475$.

A 3×4 Contingency Table Example

Finally, consider the sparse 3×4 contingency table given in Table 4.55. For the frequency data given in Table 4.55, only $M = 706$ arrangements of cell frequencies are consistent with the observed row and column marginal frequency distributions, {5, 5, 4} and {4, 3, 4, 3}, respectively, and 168 of the $M = 706$ arrangements have hypergeometric point probability values equal to or less than the point probability value of the observed table ($p = 0.1903 \times 10^{-3}$), yielding an exact probability value of $P = 0.0187$.

Table 4.54 Example 3×3 contingency table

	A_1	A_2	A_3	Total
B_1	3	5	2	10
B_2	2	9	3	14
B_3	8	2	6	16
Total	13	16	11	40

Table 4.55 Example 3×4 contingency table

	A_1	A_2	A_3	A_4	Total
B_1	3	0	0	2	5
B_2	0	3	1	1	5
B_3	1	0	3	0	4
Total	4	3	4	3	14

4.12 Analyses of 2×2×2 Tables

Fisher's exact probability test is not limited to two-way contingency tables. Consider a 2×2×2 contingency table, such as depicted in Fig. 4.1, where n_{ijk} denotes the cell frequency of the ith row, jth column, and kth slice for $i, j, k = 1, 2$. Denote by a dot $(\cdot)$ the partial sum of all rows, all columns, or all slices, depending on the position of the $(\cdot)$ in the subscript list. If the $(\cdot)$ is in the first subscript position, the sum is over all rows, if the $(\cdot)$ is in the second subscript position, the sum is over all columns, and if the $(\cdot)$ is in the third subscript position, the sum is over all slices. Thus, $n_{i..}$ denotes the marginal frequency total of the ith row, $i = 1, \ldots, r$, summed over all columns and slices; $n_{.j.}$ denotes the marginal frequency total of the jth column, $j = 1, \ldots, c$, summed over all rows and slices; and $n_{..k}$ denotes the marginal frequency total of the kth slice, $k = 1, \ldots, s$, summed over all rows and columns. Therefore, $A = n_{1..}$, $B = n_{.1.}$, $C = n_{..1}$, and $N = n_{...}$ denote the observed marginal frequency totals of the first row, first column, first slice, and entire table, respectively, such that $1 \leq A \leq B \leq C \leq N/2$. Also, let $w = n_{111}$, $x = n_{112}$, $y = n_{121}$, and $z = n_{211}$ denote cell frequencies of the 2×2×2 contingency table. Then, the probability for any w, x, y, and z is given by:

$$
\begin{aligned}
P(w, x, y, z | A, B, C, N) = \\
&\big[A!(N-A)!\,B!\,(N-B)!\,C!(N-C)!\big] \\
\times &\big[(N!)^2\,w!\,x!\,y!\,z!\,(A-w-x-y)!\,(B-w-x-z)! \\
&(C-w-y-z)!\,(N-A-B-C+2w+x+y+z)!\big]^{-1}
\end{aligned}
$$

[67]. An algorithm to compute Fisher's exact probability test involves a nested looping structure and requires two distinct passes. The first pass yields the exact

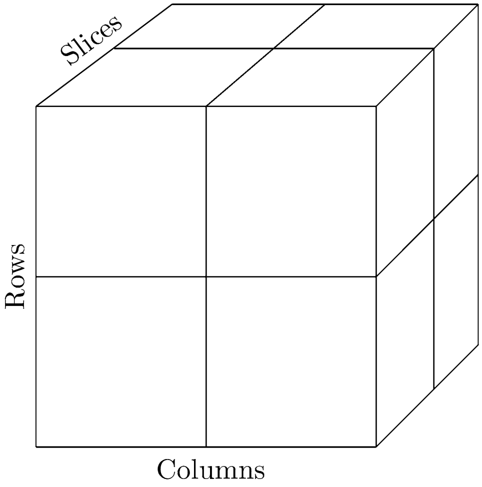

Fig. 4.1 Graphic depiction of a 2×2×2 contingency table

probability, U, of the observed 2×2×2 contingency table and is terminated when U is obtained. The second pass yields the exact probability value of all tables with hypergeometric point probability values equal to or less than the point probability of the observed contingency table. The four nested loops within each pass are over the cell frequency indices w, x, y, and z, respectively. The bounds for w, x, y, and z are

$$\begin{aligned} 0 \leq & w \leq M_w \;, \\ 0 \leq & x \leq M_x \;, \\ 0 \leq & y \leq M_y \;, \end{aligned}$$

and

$$L_x \leq z \leq M_z \;,$$

respectively, where $M_w = A$, $M_x = A - w$, $M_y = A - w - x$, $M_z = \min(B - w - x, C - w - y)$, and $L_z = \max(0, A + B + C - N - 2w - x - y)$.

The recursion method can be illustrated with the fourth (inner) loop over z, given w, x, y, A, B, C, and N because the inner loop yields both U on the first pass and the exact probability value on the second pass. Let $H(w, x, y, z)$ be a recursively defined positive function given A, B, C, and N, satisfying

$$H(w, x, y, z + 1) = H(w, x, y, z) \times g(w, x, y, z) \;,$$

where

$$g(w, x, y, z) = \frac{(B - w - x - z)(C - w - z)}{(z + 1)(N - A - B - C + 2w + x + y + z + 1)} \;.$$

The remaining three loops of each pass initialize $H(w, x, y, z)$ for continued enumerations. Let $I_x = \max(0, A + B + C - N)$ and set the initial value of $H(0, 0, 0, I_z)$ to an arbitrary small positive constant. Then, the total over the completely enumerated distribution is found by:

$$T = \sum_{w=0}^{M_w} \sum_{x=0}^{M_x} \sum_{y=0}^{M_y} \sum_{z=L_x}^{M_x} H(w, x, y, z) \;.$$

If w_o, x_o, y_o, and z_o are the values of w, x, y, and z in the observed 2×2×2 contingency table, then U and the exact probability value (P) are given by:

$$U = H(w_o, x_o, y_o, z_o)/T$$

and

$$P = \sum_{w=0}^{M_w} \sum_{x=0}^{M_x} \sum_{y=0}^{M_y} \sum_{z=L_x}^{M_x} H(w, x, y, z)\, \psi(w, x, y, z,)/T \ .$$

respectively, where

$$\psi(w, x, y, z) = \begin{cases} 1 & \text{if } H(w, x, y, z) \leq H(w_o, x_o, y_o, z_o) \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

4.12.1 A 2×2×2 *Contingency Table Example*

Consider a scenario in which $N = 1{,}663$ respondents were asked if they agreed with the statement that women should have equal pay for the same job as men (No, Yes). The respondents were then classified by region of the country (North, South) and by year of the survey (2000, 2010). For the frequency data given in Table 4.56, $M = 3{,}683{,}159{,}504$ arrangements of cell frequencies are consistent with the observed row, column, and slice marginal frequency distributions, {623, 1040}, {1,279, 384}, and {1,039, 624}, respectively. Exactly 2,761,590,498 of the arrangements have hypergeometric point probability values equal to or less than the point probability value of the observed table ($p = 0.1684{\times}10^{-72}$), yielding an exact probability value of $P = 0.1684{\times}10^{-65}$.

4.12.2 A 3×4×2 *Contingency Table Example*

Fisher's exact probability test is not limited to multi-way contingency tables with only two categories in each dimension. Consider the $r{\times}c{\times}s$ contingency table given in Table 4.57 with $r = 3$ rows, $c = 4$ columns, and $s = 2$ slices. In general, it is not efficient to analyze complex multi-way tables with exact permutation procedures, as there are usually too many arrangements of cell frequencies in the reference set of all possible arrangements of cell frequencies. For the frequency data given in Table 4.57 with row, column, and slice marginal frequency distributions, {71, 31},

Table 4.56 Cross-classification of responses (No, Yes), categorized by year and region

	Region			
	North		South	
Year	No	Yes	No	Yes
2000	410	56	126	31
2010	439	374	64	163

Table 4.57 Three-way contingency table with $r = 3$ rows, $c = 4$ columns, and $s = 2$ slices

		C_1	C_2	C_3	C_4
S_1	R_1	3	4	1	6
	R_2	7	8	4	9
	R_3	7	8	9	5
S_2	R_1	2	6	5	2
	R_2	0	2	6	1
	R_3	2	4	0	1

$\{21, 32, 25, 24\}$, and $\{29, 37, 36\}$, respectively, the approximate resampling probability value based on $L = 1{,}000{,}000$ random arrangements of cell frequencies is

$$P = \frac{29{,}600}{1{,}000{,}000} = 0.0296 \;.$$

4.13 Coda

Chapter 3 applied permutation statistical methods to measures of association for two nominal-level variables that are based on Pearson's chi-squared test statistic. Chapter 4 applied exact and resampling permutation statistical methods to measures of association for two nominal-level variables that are not based on Pearson's chi-squared test statistic. Included in Chap. 4 were Goodman and Kruskal's asymmetric λ_a, λ_b, t_a, and t_b measures, Cohen's unweighted chance-corrected κ coefficient, McNemar's and Cochran's Q measures of change, Leik and Gove's d_N^c measure, Mielke and Siddiqui's exact probability for the matrix occupancy problem, and Fisher's exact probability test, extended to cover a variety of contingency tables. For each test, examples illustrated the measures and either exact or resampling probability values based on the appropriate permutation analysis were provided.

Chapter 5 applies permutation statistical methods to a variety of measures of association designed for ordinal-level variables that are based on all possible paired comparisons. Included in Chap. 5 are Kendall's τ_a and τ_b and Stuart's τ_c measures of ordinal association, Somers' asymmetric d_{yx} and d_{xy} measures, Kim's $d_{y.x}$ and $d_{x.y}$ measures, Wilson's e measure, and Cureton's rank-biserial correlation coefficient.

References

1. Acock, A.C., Stavig, G.R.: A measure of association for nonparametric statistics. Social Forces **57**, 1381–1386 (1979)
2. Agresti, A., Finlay, B.: Statistical Methods for the Social Sciences. Prentice–Hall, Upper Saddle River, NJ (1997)
3. Armitage, P., Blendis, L.M., Smyllie, H.C.: The measurement of observer disagreement in the recording of signs. J. R. Stat. Soc. A Gen. **129**, 98–109 (1966)

4. Bartko, J.J.: The intraclass correlation coefficient as a measure of reliability. Psychol. Rep. **19**, 3–11 (1966)
5. Bartko, J.J., Carpenter, W.T.: On the methods and theory of reliability. J. Nerv. Ment. Dis. **163**, 307–317 (1976)
6. Berkson, J.: Some difficulties of interpretation encountered in the application of the chi-square test. J. Am. Stat. Assoc. **33**, 526–536 (1938)
7. Berry, K.J., Jacobsen, R.B., Martin, T.W.: Clarifying the use of chi-square: Testing the significance of Goodman and Kruskal's gamma. Soc. Sci. Quart. **57**, 687–690 (1976)
8. Berry, K.J., Johnston, J.E., Mielke, P.W.: Exact and resampling probability values for measures associated with ordered R by C contingency tables. Psychol. Rep. **99**, 231–238 (2006)
9. Berry, K.J., Johnston, J.E., Mielke, P.W.: An alternative measure of effect size for Cochran's Q test for related proportions. Percept. Motor Skill **104**, 1236–1242 (2007)
10. Berry, K.J., Martin, T.W., Olson, K.F.: A note on fourfold point correlation. Educ. Psychol. Meas. **34**, 53–56 (1974)
11. Berry, K.J., Martin, T.W., Olson, K.F.: A note on fourfold point correlation. Educ. Psychol. Meas. **34**, 53–56 (1974)
12. Berry, K.J., Martin, T.W., Olson, K.F.: Testing theoretical hypotheses: A PRE statistic. Social Forces **53**, 190–196 (1974)
13. Berry, K.J., Mielke, P.W.: Goodman and Kruskal's tau-b statistic: A nonasymptotic test of significance. Sociol Method Res. **13**, 543–550 (1985)
14. Berry, K.J., Mielke, P.W.: Subroutines for computing exact chi-square and Fisher's exact probability tests. Educ. Psychol. Meas. **45**, 153–159 (1985)
15. Berry, K.J., Mielke, P.W.: Exact chi-square and Fisher's exact probability test for 3 by 2 cross-classification tables. Educ. Psychol. Meas. **47**, 631–636 (1987)
16. Berry, K.J., Mielke, P.W.: A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educ. Psychol. Meas. **48**, 921–933 (1988)
17. Blalock, H.M.: Probabilistic interpretations for the mean square contingency. J. Am. Stat. Assoc. **53**, 102–105 (1958)
18. Blalock, H.M.: Social Statistics, 2nd edn. McGraw–Hill, New York (1979)
19. Böhning, D., Holling, H.: A Monte Carlo study on minimizing chi-square distances under the hypothesis of homogeneity or independence for a two-way contingency table. Statistics **20**, 55–70 (1989)
20. Brennan, R.L., Prediger, D.J.: Coefficient kappa: Some uses, misuses, and alternatives. Educ. Psychol. Meas. **41**, 687–699 (1981)
21. Cicchetti, D.V., Showalter, D., Tyrer, P.J.: The effect of number of rating scale categories on levels of interrater reliability. Appl. Psychol. Meas. **9**, 31–36 (1985)
22. Cochran, W.G.: The comparison of percentages in matched samples. Biometrika **37**, 256–266 (1950)
23. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**, 37–46 (1960)
24. Cohen, J.: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. **70**, 213–220 (1968)
25. Conger, A.J.: Integration and generalization of kappas for multiple raters. Psychol. Bull. **88**, 322–328 (1980)
26. Conger, A.J.: Kappa reliabilities for continuous behaviors and events. Educ. Psychol. Meas. **45**, 861–868 (1985)
27. Costner, H.L.: Criteria for measures of association. Am. Sociol. Rev. **30**, 341–353 (1965)
28. Eicker, P.J., Siddiqui, M.M., Mielke, P.W.: A matrix occupancy problem. Ann. Math. Stat. **43**, 988–996 (1972)
29. Feinstein, A.R.: Clinical Biostatistics XXIII: The role of randomization in sampling, testing, allocation, and credulous idolatry (Part 2). Clin. Pharmacol. Ther. **14**, 898–915 (1973)
30. Ferguson, G.A.: Statistical Analysis in Psychology and Education, 5th edn. McGraw–Hill, New York (1981)

31. Fisher, R.A.: Statistical Methods for Research Workers, 5th edn. Oliver and Boyd, Edinburgh (1934)
32. Fisher, R.A.: The logic of inductive inference (with discussion). J. R. Stat. Soc. **98**, 39–82 (1935)
33. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psych. Bull. **76**, 378–382 (1971)
34. Fleiss, J.L., J, C.: The equivalence of weighted kappa and the intraclass coefficient as measures of reliability. Educ. Psychol. Meas. **33**, 613–619 (1973)
35. Gail, M., Mantel, N.: Counting the number of $r \times c$ contingency tables with fixed margins. J. Am. Stat. Assoc. **72**, 859–862 (1977)
36. Gittelsohn, A.M.: An occupancy problem. Am. Stat. **23**, 11–12 (1969)
37. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. J. Am. Stat. Assoc. **49**, 732–764 (1954)
38. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications, II: Further discussion and references. J. Am. Stat. Assoc. **54**, 123–163 (1959)
39. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications, III: Approximate sampling theory. J. Am. Stat. Assoc. **58**, 310–364 (1963)
40. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications, IV: Simplification of asymptotic variances. J. Am. Stat. Assoc. **67**, 415–421 (1972)
41. Graham, P., Jackson, R.: The analysis of ordinal agreement data: Beyond weighted kappa. J. Clin. Epidemiol. **46**, 1055–1062 (1993)
42. Guilford, J.P.: Fundamental Statistics in Psychology and Education. McGraw–Hill, New York (1950)
43. Guttman, L.: An outline of the statistical theory of prediction. In: Horst, P., Wallin, P., Guttman, L., et al. (eds.) The Prediction of Personal Adjustment, pp. 253–318. Social Science Research Council, New York (1941)
44. Hubert, L.J.: Kappa revisited. Psychol. Bull. **84**, 289–297 (1977)
45. Hunter, A.A.: On the validity of measures of association: The nominal-nominal two-by-two case. Am. J. Sociol. **79**, 99–109 (1973)
46. Iachan, R.: Measures of agreement for incompletely ranked data. Educ. Psychol. Meas. **44**, 823–830 (1984)
47. Irwin, J.O.: Tests of significance for differences between percentages based on small numbers. Metron **12**, 83–94 (1935)
48. Kendall, M.G., Babington Smith, B.: On the method of paired comparisons. Biometrika **31**, 324–345 (1940)
49. Kramer, M., Schmidhammer, J.: The chi-squared statistic in ethology: Use and misuse. Animal. Beh. **44**, 833–841 (1992)
50. Krippendorff, K.: Bivariate agreement coefficients for reliability of data. In: Borgatta, E.F. (ed.) Sociological Methodology, pp. 139–150. Jossey–Bass, San Francisco (1970)
51. Landis, J.R., Koch, G.G.: An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics **33**, 363–374 (1977)
52. Leik, R.K., Gove, W.R.: The conception and measurement of asymmetric monotonic relationships in sociology. Am. J. Sociol. **74**, 696–709 (1969)
53. Leik, R.K., Gove, W.R.: Integrated approach to measuring association. In: Costner, H.L. (ed.) Sociological Methodology, pp. 279–301. Jossey Bass, San Francisco, CA (1971)
54. Light, R.J.: Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychol. Bull. **76**, 365–377 (1971)
55. Light, R.J., Margolin, B.H.: An analysis of variance for categorical data. J. Am. Stat. Assoc. **66**, 534–544 (1971)
56. Lunney, G.H.: Using analysis of variance with a dichotomous dependent variable: An empirical study. J. Educ. Meas. **7**, 263–269 (1970)
57. Maclure, M., Willett, W.C.: Misinterpretation and misuse of the kappa statistic. Am. J. Epidemiol. **126**, 161–169 (1987)

58. Mantel, N.: 361: Approaches to a health research occupancy problem. Biometrics **30**, 355–362 (1974)
59. Mantel, N., Pasternack, B.S.: A class of occupancy problems. Am. Stat. **22**, 23–24 (1968)
60. Marascuilo, L.A., McSweeney: Nonparametric and Distribution-free methods in the Social Sciences. Brooks–Cole, Monterey, CA (1977)
61. Margolin, B.H., Light, R.J.: An analysis of variance for categorical data, II: Small sample comparisons with chi square and other competitors. J. Am. Stat. Assoc. **69**, 755–764 (1974)
62. May, R.B., Masson, M.E., Hunter, M.A.: Applications of Statistics in Behavioral Research. Harper & Row, New York (1990)
63. McNemar, Q.: Note on the sampling error of the differences between correlated proportions and percentages. Psychometrika **12**, 153–157 (1947)
64. Mielke, P.W., Berry, K.J.: Cumulant methods for analyzing independence of *r*-way contingency tables and goodness-of-fit frequency data. Biometrika **75**, 790–793 (1988)
65. Mielke, P.W., Berry, K.J.: Nonasymptotic inferences based on Cochran's *Q* test. Percept. Motor Skill **81**, 319–322 (1995)
66. Mielke, P.W., Berry, K.J., Johnston, J.E.: Resampling programs for multiway contingency tables with fixed marginal frequency totals. Psychol. Rep. **101**, 18–24 (2007)
67. Mielke, P.W., Berry, K.J., Zelterman, D.: Fisher's exact test of mutual independence for $2\times2\times2$ cross-classification tables. Educ. Psychol. Meas. **54**, 110–114 (1994)
68. Mielke, P.W., Siddiqui, M.M.: A combinatorial test for independence of dichotomous responses. J. Am. Stat. Assoc. **60**, 437–441 (1965)
69. Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. 5 **50**, 157–175 (1900)
70. Pearson, K.: On the laws of inheritance in man: II. On the inheritance of the mental and moral characters in man, and its comparison with the inheritance of the physical characters. Biometrika **3**, 131–190 (1904)
71. Pierce, A.: Fundamentals of Nonparametric Statistics. Dickenson, Belmont, CA (1970)
72. Robinson, W.S.: Ecological correlations and the behavior of individuals. Am. Soc. Rev. **15**, 351–357 (1950). [Reprinted in Int J Epidem **38**, 337–341 (2009)]
73. Robinson, W.S.: The statistical measurement of agreement. Am. Sociol. Rev. **22**, 17–25 (1957)
74. Robinson, W.S.: The geometric interpretation of agreement. Am. Sociol. Rev. **24**, 338–345 (1959)
75. Särndal, C.E.: A comparative study of association measures. Psychometrika **39**, 165–187 (1974)
76. Scott, W.A.: Reliability of content analysis: The case of nominal scale coding. Public Opin. Quart. **19**, 321–325 (1955)
77. Serlin, R.C., Carr, J., Marascuilo, L.A.: A measure of association for selected non-parametric procedures. Psychol. Bull. **92**, 786–790 (1982)
78. Somers, R.H.: A new asymmetric measure of association for ordinal variables. Am. Sociol. Rev. **27**, 799–811 (1962)
79. Spearman, C.E.: The proof and measurement of association between two things. Am. J. Psychol. **15**, 72–101 (1904)
80. Spearman, C.E.: 'Footrule' for measuring correlation. Brit. J. Psychol. **2**, 89–108 (1906)
81. Sprott, D.A.: A note on a class of occupancy problems. Am. Stat. **23**, 12–13 (1969)
82. Tschuprov, A.A.: Principles of the Mathematical Theory of Correlation. Hodge, London (1939). [Translated by M. Kantorowitsch]
83. Vanbelle, S., Albert, A.: A note on the linearly weighted kappa coefficient for ordinal scales. Stat. Methodol. **6**, 157–163 (2008)
84. Wasserstein, R., Lazar, N.A.: The ASA's statement on p-values: Context, process, and purpose. Am. Stat. **70**, 129–133 (2016)
85. White, C.: The committee problem. Am. Stat. **25**, 25–26 (1971)
86. Wickens, T.D.: Multiway Contingency Tables Analysis for the Social Sciences. Erlbaum, Hillsdale, NJ (1989)

87. Wilkinson, L.: Statistical methods in psychology journals: Guidelines and explanations. Am. Psychol. **54**, 594–604 (1999)
88. Williams, G.W.: Comparing the joint agreement of several raters with another rater. Biometrics **32**, 619–627 (1976)
89. Yates, F.: Contingency tables involving small numbers and the χ^2 test. Suppl. J. R. Stat. Soc. **1**, 217–235 (1934)
90. Yule, G.U.: On the methods of measuring association between two attributes. J. R. Stat. Soc. **75**, 579–652 (1912). [Originally a paper read before the Royal Statistical Society on 23 April 1912]