Behavior Research Methods & Instrumentation
1977, Vol. 9 (2), 193-201

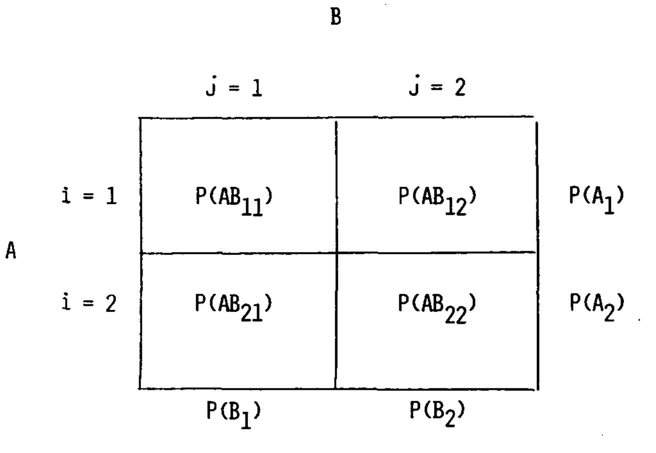
Monte Carlo simulations and the chi-square test of independence

DRAKE R. BRADLEY and STEVEN CUTCOMB
Bates College, Lewiston, Maine 04240

A Monte Carlo program for sampling 2 by 2 contingency tables from a user-specified population is discussed. Applications include computer-assisted instruction (CAI) of statistics, evaluation of actual vs nominal Type I error rates of the chi-square test of independence when expected frequencies are less than 10, and estimation of the power of the chi-square test.

In the present paper, we describe a Monte Carlo program, CHI-PHI, which randomly samples 2 by 2 contingency tables from a user-specified population. Figure 1a presents the symbolic notation used for depicting the population. The parameters requiring specification are the row and column marginal probabilities [$P(A_i)$, $P(B_j)$], the degree of association between the two categorical variables (ϕ), and the size and number of samples to be drawn (N, K). The marginal probabilities and the phi coefficient uniquely determine the joint probability distribution [$P(AB_{ij})$] of the population.[1] These parameters are specified on-line when the program is run (Figure 1b). CHI-PHI then randomly samples N observations for each of the K samples such that the probability of obtaining an observation in Cell_{ij} is equal to the joint probability for that cell. In the special case where $\phi = 0$, the variables are independent and the program sets the probability of obtaining an observation in Cell_{ij} equal to the product of the corresponding marginal probabilities [$P(AB_{ij}) = P(A_i)P(B_j)$]. When $\phi \neq 0$, the program sets the probability of obtaining an observation in Cell_{ij} equal to the product of the corresponding marginal probabilities incremented or decremented by whatever amount is necessary to produce the degree of association specified by the user [$P(AB_{ij}) = P(A_i)P(B_j) \pm \Delta p$]. For each sample of N observations drawn from the user-specified population, the program computes the chi-square statistic and the phi coefficient, and determines whether or not the observed chi square is significant ($p < .05$). CHI-PHI then outputs the following information: (1) the theoretical expected frequency of Cell_{ij}, assuming independence between the two variables, obtained by multiplying sample size by $P(A_i)P(B_j)$; (2) the mean expected frequency of Cell_{ij}, obtained by averaging the K expected frequencies for that cell (each computed from the empirically sampled marginals); (3) the mean observed frequency of Cell_{ij}, obtained by averaging the K observed frequencies for that cell; (4) the proportion of significant chi-square statistics obtained across the K samples; (5) the average phi coefficient based on the K sample estimates.

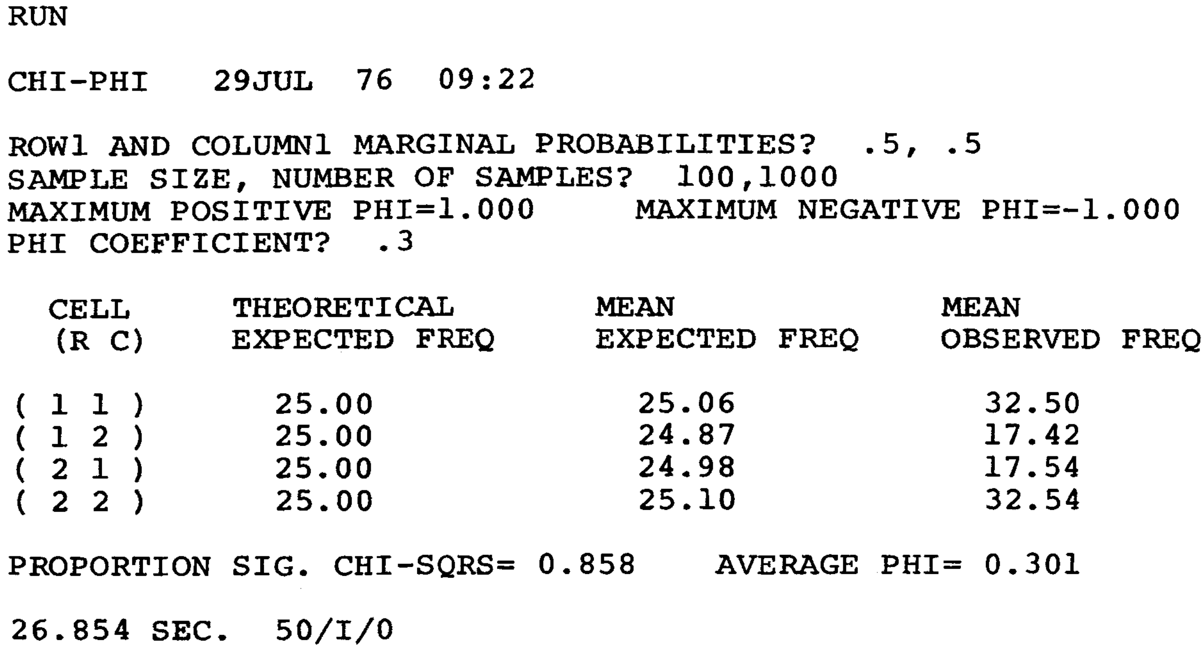
CHI-PHI permits the user to specify either negative or positive association between the two categorical variables. However, depending on the marginal probabilities specified in defining the population, the maximum possible values of the phi coefficient may or may not achieve an upper theoretical limit of ±1. The maximum values are therefore indicated by the program (Figure 1b), and the user simply selects a value of phi within or including these limits. If the variables composing the 2 by 2 contingency table achieve at least an ordinal level of measurement (Row 1 indicating "less" of some characteristic than Row 2, and Column 1 "less" than Column 2), then the phi coefficient is interpretable as a correlation coefficient. Specifically, phi is the fourfold point correlation coefficient between the row and column variables. Positive values of phi



		B		
		j = 1	j = 2	
A	i = 1	$P(AB_{11})$	$P(AB_{12})$	$P(A_1)$
	i = 2	$P(AB_{21})$	$P(AB_{22})$	$P(A_2)$
		$P(B_1)$	$P(B_2)$	

Figure 1a. Symbolic notation used to denote the marginal and joint probabilities of the population.

The computer simulations reported in this paper were conducted on the Dartmouth Time-Sharing System, using the recently installed Honeywell 66/40 duplex computer system. We wish to express our appreciation to Dartmouth College for the excellent service provided by the Kiewit Computation Center. Also, we thank Susan T. Dumais and Lt. Col. T. D. Bradley for their assistance in conducting various phases of this project. Reprint requests should be addressed to Drake R. Bradley, Department of Psychology, Bates College, Lewiston, Maine 04240.



```
RUN

CHI-PHI   29JUL  76  09:22

ROW1 AND COLUMN1 MARGINAL PROBABILITIES?  .5, .5
SAMPLE SIZE, NUMBER OF SAMPLES?  100,1000
MAXIMUM POSITIVE PHI=1.000    MAXIMUM NEGATIVE PHI=-1.000
PHI COEFFICIENT?  .3

 CELL     THEORETICAL        MEAN                 MEAN
 (R C)    EXPECTED FREQ      EXPECTED FREQ        OBSERVED FREQ

( 1 1 )     25.00              25.06                 32.50
( 1 2 )     25.00              24.87                 17.42
( 2 1 )     25.00              24.98                 17.54
( 2 2 )     25.00              25.10                 32.54

PROPORTION SIG. CHI-SQRS= 0.858    AVERAGE PHI= 0.301

26.854 SEC.  50/I/O
```

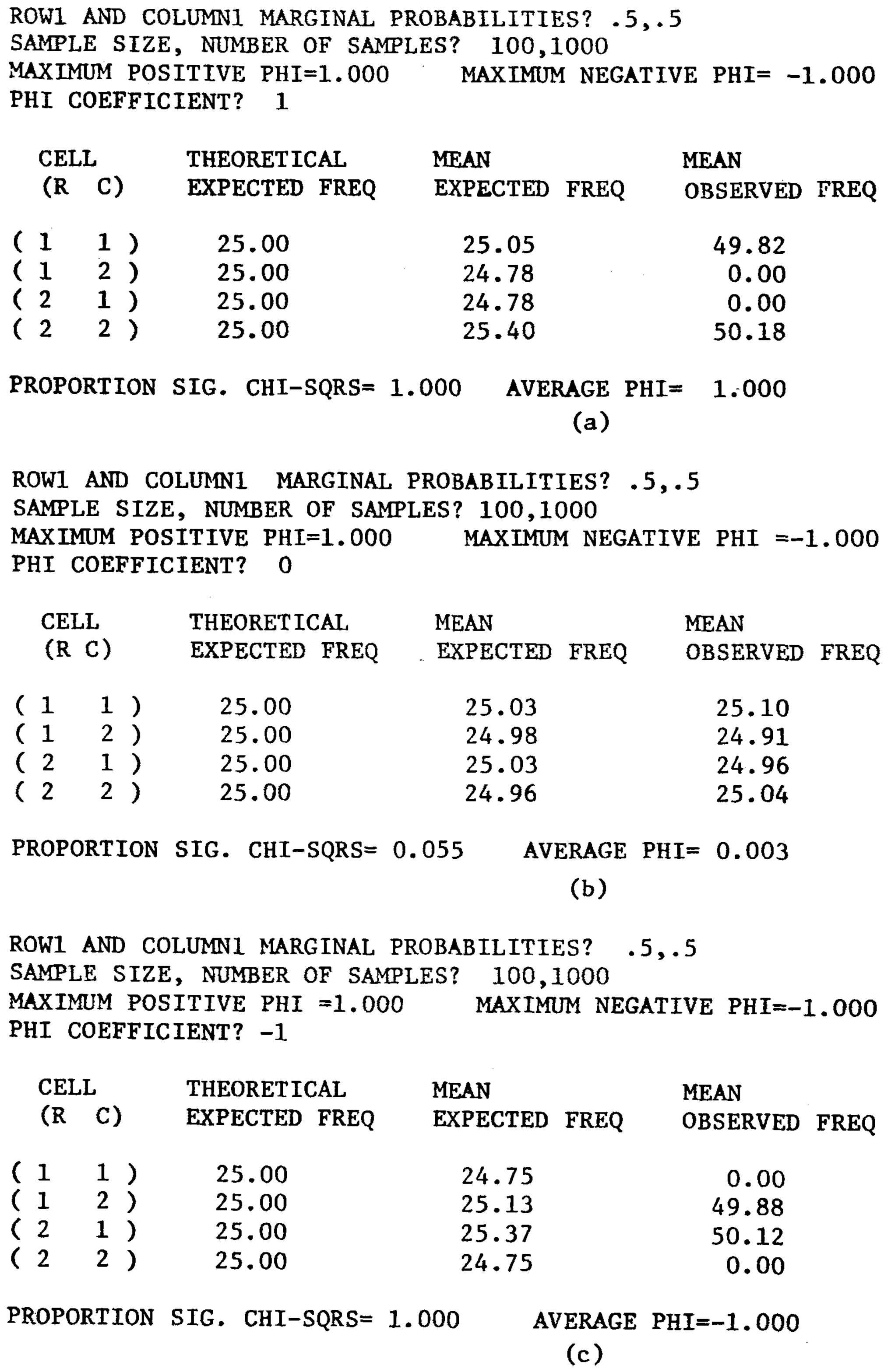
Figure 1b. A sample interactive sequence of the Monte Carlo program CHI-PHI.

will cause observations to cluster in the upper left and lower right cells of the contingency table.[2] In tables where Row 1 and Column 1 signify the absence of the row and column attributes, and Row 2 and Column 2 signify the presence of these attributes, then a positive phi coefficient indicates that observations cluster in the agreement cells (Cell 11 and Cell 22), whereas a negative phi coefficient indicates that observations cluster in the disagreement cells (Cell 12 and Cell 21). Finally, in situations where the row and column variables have no such quasiquantitative interpretation, then positive and negative values of phi simply differentiate the case where observations cluster in the upper left and lower right cells (positive) from the case where they cluster in the lower left and upper right cells (negative).

Figure 2 presents examples of the output generated by CHI-PHI when all marginal probabilities are set equal to .5, and the value of phi is set equal to (a) 1 , (b) 0, and (c) −1. Figure 3 shows a simulation in which the row probabilities are set equal to .4 and .6, the column probabilities equal to .3 and .7, and the value of phi equal to (a) .802, (b) 0, and (c) −.535. In all cases, N = 100, K = 1,000, and the nonzero values of phi are equal to the maximum values possible, given the configuration of marginal probabilities selected. Note that, in Figures 2a and 2c, the mean observed frequencies cluster entirely in one set of diagonal cells, with no observed frequencies occurring in the opposing set of diagonal cells. This is consistent with the fact that the variables are perfectly correlated ($\phi = \pm 1$). In Figure 3, however, the marginal probabilities chosen prevent the selection of phi equal to the upper theoretical limit of ±1. When the maximum possible values are used in the simulation (Figures 3a and 3c), the mean observed cell frequencies cluster in the appropriate diagonals, but only one of the opposing diagonal cells has a zero mean observed frequency. In tables having configurations of this type, it is impossible to achieve a perfect positive or negative correlation between the two categorical variables.[3]

We now consider several possible applications of CHI-PHI. As a general purpose program for simulating sampling from a joint-probability distribution, CHI-PHI may be used by instructors for CAI applications in statistics (demonstrations, labs), or by researchers for estimating possible inflation of the Type I error rate when minimum expected cell-frequency requirements are violated. CHI-PHI may also be used in lieu of tables of noncentral chi square to estimate the power of the chi-square test under various configurations of marginal probabilities, sample size, and a priori assumptions concerning the degree of association present in the population. In the remaining portions of this paper, we provide examples of each of these applications.

One of the most useful demonstrations provided by CHI-PHI for the beginning student in statistics is a clear and obvious delineation of the four possible outcomes of a statistical decision: (1) accepting the null hypothesis when it is true $(1 - \alpha)$; (2) rejecting the null hypothesis when it is true (α, the Type I error rate); (3) accepting the null hypothesis when it is false (β, the Type II error rate); and (4) rejecting the null hypothesis when it is false $(1 - \beta)$. Since the student defines the population at run time, he or she knows

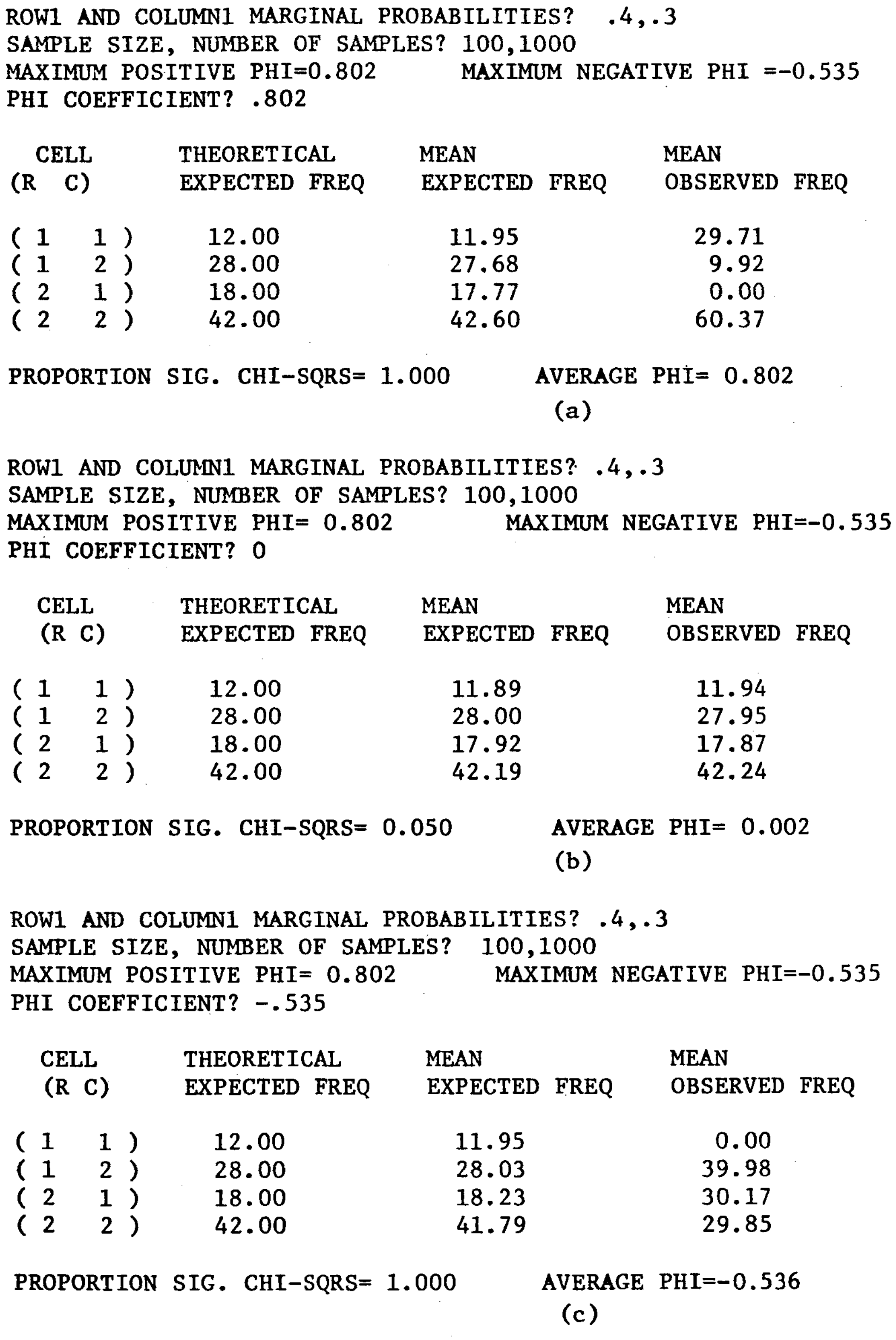


```
ROW1 AND COLUMN1 MARGINAL PROBABILITIES? .5,.5
SAMPLE SIZE, NUMBER OF SAMPLES?  100,1000
MAXIMUM POSITIVE PHI=1.000      MAXIMUM NEGATIVE PHI= -1.000
PHI COEFFICIENT?  1

  CELL        THEORETICAL       MEAN               MEAN
  (R  C)      EXPECTED FREQ     EXPECTED FREQ      OBSERVED FREQ

( 1   1 )       25.00             25.05              49.82
( 1   2 )       25.00             24.78               0.00
( 2   1 )       25.00             24.78               0.00
( 2   2 )       25.00             25.40              50.18

PROPORTION SIG. CHI-SQRS= 1.000    AVERAGE PHI=  1.000
                                             (a)
```

```
ROW1 AND COLUMN1  MARGINAL PROBABILITIES? .5,.5
SAMPLE SIZE, NUMBER OF SAMPLES? 100,1000
MAXIMUM POSITIVE PHI=1.000      MAXIMUM NEGATIVE PHI =-1.000
PHI COEFFICIENT?  0

  CELL        THEORETICAL       MEAN               MEAN
  (R C)       EXPECTED FREQ     EXPECTED FREQ      OBSERVED FREQ

( 1   1 )       25.00             25.03              25.10
( 1   2 )       25.00             24.98              24.91
( 2   1 )       25.00             25.03              24.96
( 2   2 )       25.00             24.96              25.04

PROPORTION SIG. CHI-SQRS= 0.055    AVERAGE PHI= 0.003
                                             (b)
```

```
ROW1 AND COLUMN1 MARGINAL PROBABILITIES?  .5,.5
SAMPLE SIZE, NUMBER OF SAMPLES?  100,1000
MAXIMUM POSITIVE PHI =1.000      MAXIMUM NEGATIVE PHI=-1.000
PHI COEFFICIENT? -1

  CELL        THEORETICAL       MEAN               MEAN
  (R  C)      EXPECTED FREQ     EXPECTED FREQ      OBSERVED FREQ

( 1   1 )       25.00             24.75               0.00
( 1   2 )       25.00             25.13              49.88
( 2   1 )       25.00             25.37              50.12
( 2   2 )       25.00             24.75               0.00

PROPORTION SIG. CHI-SQRS= 1.000     AVERAGE PHI=-1.000
                                             (c)
```

Figure 2. Monte Carlo simulations selecting 1,000 samples of N = 100 from a population having marginal probabilities equal to .5, .5, .5, .5, and phi equal to (a) 1, (b) 0, and (c) –1.



```
ROW1 AND COLUMN1 MARGINAL PROBABILITIES?  .4,.3
SAMPLE SIZE, NUMBER OF SAMPLES? 100,1000
MAXIMUM POSITIVE PHI=0.802      MAXIMUM NEGATIVE PHI =-0.535
PHI COEFFICIENT? .802

  CELL        THEORETICAL      MEAN              MEAN
 (R  C)       EXPECTED FREQ    EXPECTED FREQ     OBSERVED FREQ

 ( 1   1 )      12.00            11.95             29.71
 ( 1   2 )      28.00            27.68              9.92
 ( 2   1 )      18.00            17.77              0.00
 ( 2   2 )      42.00            42.60             60.37

 PROPORTION SIG. CHI-SQRS= 1.000      AVERAGE PHI= 0.802
                                          (a)
```

```
ROW1 AND COLUMN1 MARGINAL PROBABILITIES? .4,.3
SAMPLE SIZE, NUMBER OF SAMPLES? 100,1000
MAXIMUM POSITIVE PHI= 0.802       MAXIMUM NEGATIVE PHI=-0.535
PHI COEFFICIENT? 0

  CELL        THEORETICAL      MEAN              MEAN
  (R C)       EXPECTED FREQ    EXPECTED FREQ     OBSERVED FREQ

 ( 1   1 )      12.00            11.89             11.94
 ( 1   2 )      28.00            28.00             27.95
 ( 2   1 )      18.00            17.92             17.87
 ( 2   2 )      42.00            42.19             42.24

 PROPORTION SIG. CHI-SQRS= 0.050       AVERAGE PHI= 0.002
                                           (b)
```

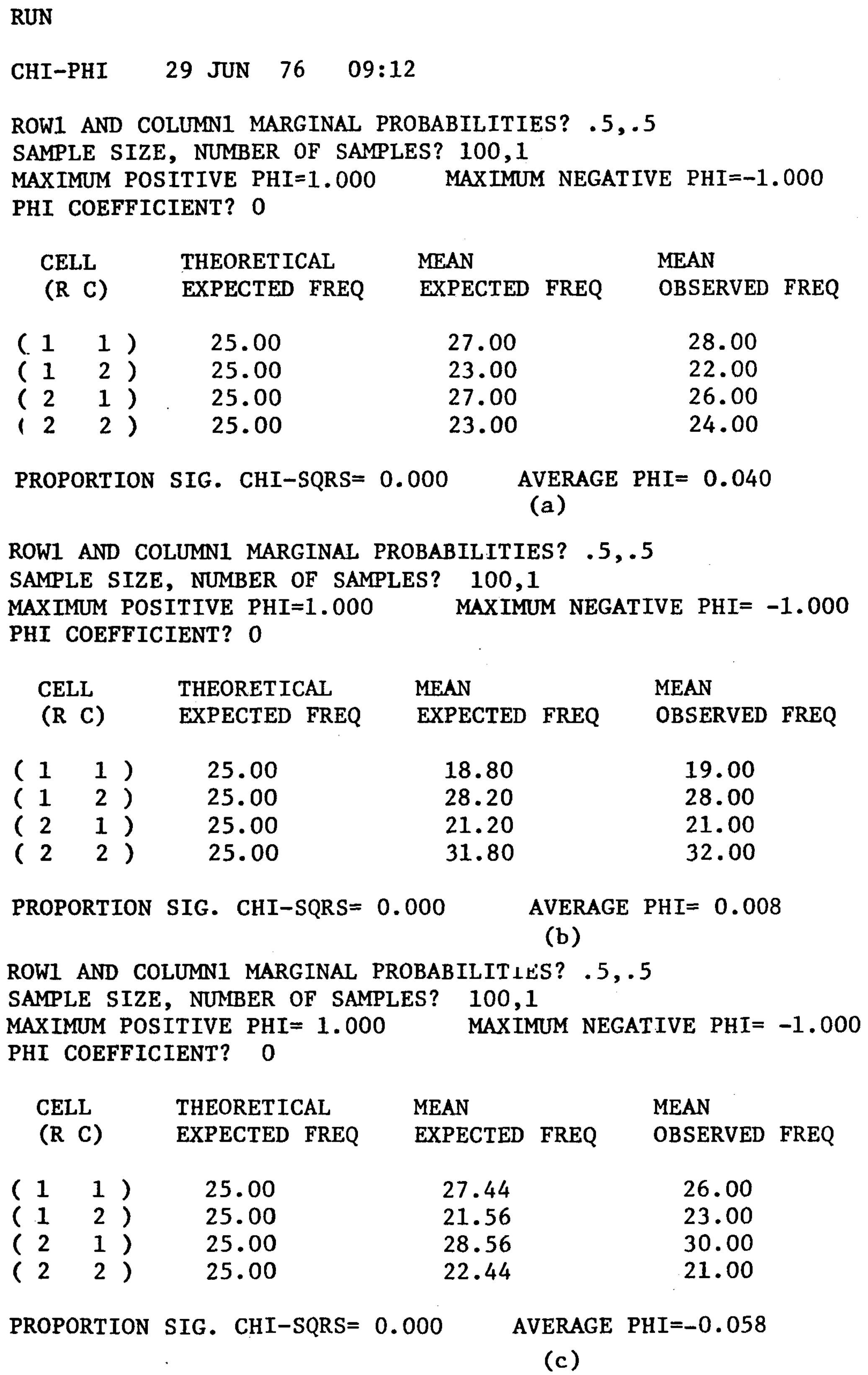
```
ROW1 AND COLUMN1 MARGINAL PROBABILITIES? .4,.3
SAMPLE SIZE, NUMBER OF SAMPLES?  100,1000
MAXIMUM POSITIVE PHI= 0.802       MAXIMUM NEGATIVE PHI=-0.535
PHI COEFFICIENT? -.535

  CELL        THEORETICAL      MEAN              MEAN
  (R C)       EXPECTED FREQ    EXPECTED FREQ     OBSERVED FREQ

 ( 1   1 )      12.00            11.95              0.00
 ( 1   2 )      28.00            28.03             39.98
 ( 2   1 )      18.00            18.23             30.17
 ( 2   2 )      42.00            41.79             29.85

 PROPORTION SIG. CHI-SQRS= 1.000      AVERAGE PHI=-0.536
                                          (c)
```

Figure 3. Monte Carlo simulations selecting 1,000 samples of N = 100 from a population having marginal probabilities of .4, .6, .3, .7, and phi equal to (a) .802, (b) 0, and (c) −.535.

in advance whether the null hypothesis is true or false (the null hypothesis is false whenever $\phi \neq 0$). The student can then determine how frequently the chi-square test of independence yields a conclusion consistent with the true state of affairs in the population. In Figures 2b and 3b, for example, the user specified that the categorical variables be entirely independent ($\phi = 0$). However, of the 1,000 samples selected from each of these two populations, 55 in the first instance and 50 in the second resulted in a decision to reject the null hypothesis (Type I error). The empirical Type I error rate is therefore about .05; that is, the same as the α level used in evaluating the significance of each chi square. In the remaining 95% of the cases $(1 - \alpha)$, a correct decision to accept the null hypothesis occurred. This illustrates that, in the absence of precise information about population parameters, all decisions made on the basis of sample data are probabilistic in nature. The student finds that this is also the case when the null hypothesis is false. Figure 1 illustrates what happens when samples are drawn from a population in which $\phi = .30$. In this case, 838 out of 1,000 samples resulted in a decision to reject the null hypothesis, and since the null hypothesis was indeed false, these decisions were correct. A reasonable estimate of the probability of correctly rejecting the null hypothesis (i.e., the power of the test) is given by the proportion of significant chi-square tests; $1 - \beta = .858$. In the remaining 142 samples, the null hypothesis was accepted (Type II error) and a reasonable estimate of the Type II error rate is $\beta = .142$. The student may also observe that increasing the sample size, the magnitude of association in the population, or both, increases the power of the chi-square test. Figures 2a, 2c, 3a, and 3c each illustrate an extreme instance of this, in which phi is set equal to its maximum positive or negative value and N is large (N = 100). In all cases, the power of the chi-square test is 1. A computer lab can be conducted which requires the student to generate and plot power curves by repeatedly running CHI-PHI with the same parameters, except for phi, which is stepped in .10 increments from the maximum negative value to the maximum positive value (see below).

Another CAI application of CHI-PHI is to demonstrate the sample-to-sample variation in 2 by 2 contingency tables sampled from the same population. If CHI-PHI is run several times holding all parameters constant and with K = 1, then each output represents the outcome of drawing one sample from the population, rather than the average results across K samples (as above). Consequently, the frequencies output underneath the "mean expected frequency" and "mean observed frequency" headings of the output table are not averages, but the actual expected and observed frequencies of the one sample drawn. Likewise, the "average phi" is simply the phi coefficient computed on that one sample. Figure 4 illustrates several successive runs of this type. This kind of exercise provides the student with a direct "feel" for how the cell and marginal frequencies of the contingency table vary from one sample to the next as a result of sampling error. As a further demonstration, the instructor may run CHI-PHI without telling the students the value of phi being input to the program, again sampling only one contingency table from the population. The students are required to decide whether or not the population value of phi is nonzero, based only on a visual inspection of the observed and expected frequencies of the table. If this procedure is repeated many times, they can then compare their "hit" rates (correct detections of nonzero phi) and error rates (false alarms, misses) with those of the chi-square test, and in so doing see that statistical decision procedures balance the successful detections of a relation against a known probability of committing a Type I error. The student may also note that, while his or her hit rate can, in some instances, be higher than that of the chi-square statistic, this is only possible at the expense of an increase in false alarms (Type I errors). Conversely, the student might have a lower Type I error rate than the chi-square statistic, but only at the expense of an increase in "misses" (Type II errors). In actuality, this exercise is a signal-detection task in which the observer (student) has to detect a signal (nonzero phi) masked by noise (sampling error).

There are, of course, many other applications of Monte Carlo sampling programs such as CHI-PHI for computer-assisted instruction of statistics. The examples given above serve to illustrate the flexibility of such general purpose programs for providing a large variety of laboratory demonstrations and exercises. Of equal importance, however, is the use of such programs by professional researchers to evaluate the consequences of violating the formal assumptions of the statistical models being applied to their data. In the case of the chi-square test of independence in 2 by 2 contingency tables, it is widely asserted that this test is valid only if all expected cell frequencies are 10 or greater (Hays, 1963, pp. 596, 613). Since the statistical model assumes a continuous distribution, and since frequency data in 2 by 2 tables are discrete, the accuracy of the approximation (chi-square probabilities for exact multinomial probabilities) is presumably adequate only if N is large and expected frequencies at least 10. Unfortunately, for many research applications this minimum expected cell-frequency requirement will not be met, even though one might plausibly expect the chi-square test to be valid anyway. Table 1 illustrates this point. The first four columns of the table list the row and column marginal probabilities defining various populations from which contingency tables might be sampled. The next two columns list the maximum positive and negative values of phi which can be achieved given the particular configuration of marginal probabilities listed



```
RUN

CHI-PHI    29 JUN  76   09:12

ROW1 AND COLUMN1 MARGINAL PROBABILITIES? .5,.5
SAMPLE SIZE, NUMBER OF SAMPLES? 100,1
MAXIMUM POSITIVE PHI=1.000     MAXIMUM NEGATIVE PHI=-1.000
PHI COEFFICIENT? 0

  CELL        THEORETICAL        MEAN                MEAN
  (R C)       EXPECTED FREQ      EXPECTED FREQ       OBSERVED FREQ

( 1   1 )      25.00              27.00               28.00
( 1   2 )      25.00              23.00               22.00
( 2   1 )      25.00              27.00               26.00
( 2   2 )      25.00              23.00               24.00

PROPORTION SIG. CHI-SQRS= 0.000      AVERAGE PHI= 0.040
                                       (a)

ROW1 AND COLUMN1 MARGINAL PROBABILITIES? .5,.5
SAMPLE SIZE, NUMBER OF SAMPLES?  100,1
MAXIMUM POSITIVE PHI=1.000      MAXIMUM NEGATIVE PHI= -1.000
PHI COEFFICIENT? 0

  CELL        THEORETICAL        MEAN                MEAN
  (R C)       EXPECTED FREQ      EXPECTED FREQ       OBSERVED FREQ

( 1   1 )      25.00              18.80               19.00
( 1   2 )      25.00              28.20               28.00
( 2   1 )      25.00              21.20               21.00
( 2   2 )      25.00              31.80               32.00

PROPORTION SIG. CHI-SQRS= 0.000       AVERAGE PHI= 0.008
                                        (b)

ROW1 AND COLUMN1 MARGINAL PROBABILITIES? .5,.5
SAMPLE SIZE, NUMBER OF SAMPLES?  100,1
MAXIMUM POSITIVE PHI= 1.000      MAXIMUM NEGATIVE PHI= -1.000
PHI COEFFICIENT?  0

  CELL        THEORETICAL        MEAN                MEAN
  (R C)       EXPECTED FREQ      EXPECTED FREQ       OBSERVED FREQ

( 1   1 )      25.00              27.44               26.00
( 1   2 )      25.00              21.56               23.00
( 2   1 )      25.00              28.56               30.00
( 2   2 )      25.00              22.44               21.00

PROPORTION SIG. CHI-SQRS= 0.000     AVERAGE PHI=-0.058

                                      (c)
```

Figure 4. A demonstration of the effect of sampling error on the variation in observed and expected frequencies of 2 by 2 tables sampled from the same population: N = 100, $\phi = 0$, and all marginal probabilities are .5.

Table 1
Error Rate Table

Marginal Probabilities				Maximum Phi	Sample Size						
R1	R2	C1	C2	Positive/Negative	4	10	20	40	60	80	100
.5	.5	.5	.5	1.00 /−1.00	.1089^{4}	.0526^{4}	.0474^{4}	.0492^{0}	.0558^{0}	.0566^{0}	.0528^{0}
.5	.5	.6	.4	.816/− .816	.1035^{4}	.0487^{4}	.0549^{4}	.0543^{2}	.0500^{0}	.0544^{0}	.0471^{0}
.5	.5	.7	.3	.655/− .655	.0921^{4}	.0407^{4}	.0501^{4}	.0571^{2}	.0527^{2}	.0453^{0}	.0523^{0}
.5	.5	.8	.2	.500/− .500	.0707^{4}	.0309^{4}	.0383^{4}	.0488^{2}	.0484^{2}	.0509^{2}	.0506^{0}
.5	.5	.9	.1	.333/− .333	.0462^{4}	.0157^{4}	.0211^{4}	.0359^{2}	.0505^{2}	.0477^{2}	.0520^{2}
.6	.4	.6	.4	1.00 /− .667	.0984^{4}	.0534^{4}	.0509^{4}	.0555^{3}	.0501^{1}	.0492^{0}	.0497^{0}
.6	.4	.7	.3	.802/− .535	.0958^{4}	.0463^{4}	.0476^{4}	.0520^{2}	.0521^{1}	.0480^{1}	.0533^{0}
.6	.4	.8	.2	.612/− .408	.0732^{4}	.0334^{4}	.0456^{4}	.0479^{2}	.0510^{2}	.0471^{2}	.0530^{1}
.6	.4	.9	.1	.408/− .272	.0492^{4}	.0214^{4}	.0287^{3}	.0390^{2}	.0418^{2}	.0477^{2}	.0528^{2}
.7	.3	.7	.3	1.00 /− .429	.0817^{4}	.0450^{4}	.0451^{4}	.0518^{3}	.0520^{1}	.0550^{1}	.0517^{1}
.7	.3	.8	.2	.764/− .327	.0674^{4}	.0451^{4}	.0425^{3}	.0430^{3}	.0494^{2}	.0483^{1}	.0472^{1}
.7	.3	.9	.1	.509/− .218	.0379^{4}	.0360^{4}	.0371^{3}	.0379^{2}	.0434^{2}	.0418^{2}	.0502^{2}
.8	.2	.8	.2	1.00 /− .250	.0535^{4}	.0551^{4}	.0475^{3}	.0372^{3}	.0485^{3}	.0493^{1}	.0482^{1}
.8	.2	.9	.1	.667/− .167	.0347^{4}	.0534^{4}	.0589^{3}	.0447^{3}	.0385^{2}	.0388^{2}	.0423^{2}
.9	.1	.9	.1	1.00 /− .111	.0237^{4}	.0481^{4}	.0721^{3}	.0568^{3}	.0492^{3}	.0437^{3}	.0425^{3}
				Simulations Using Extremely Unbalanced Marginal Probabilities							
.5	.5	.99	.01	.101/− .101	.0052^{4}	.0013^{4}	.0006^{4}	.0011^{2}	.0009^{2}	.0011^{2}	.0026^{2}
.6	.4	.99	.01	.123/− .082	.0055^{4}	.0029^{4}	.0021^{3}	.0025^{2}	.0028^{2}	.0045^{2}	.0049^{2}
.7	.3	.99	.01	.154/− .066	.0049^{4}	.0045^{4}	.0082^{3}	.0072^{2}	.0104^{2}	.0115^{2}	.0159^{2}
.8	.2	.99	.01	.201/− .050	.0049^{4}	.0087^{4}	.0174^{3}	.0286^{3}	.0361^{2}	.0417^{2}	.0445^{2}
.9	.1	.99	.01	.302/− .034	.0028^{4}	.0071^{4}	.0171^{3}	.0309^{3}	.0381^{3}	.0522^{3}	.0531^{3}
.99	.01	.99	.01	1.00 /− .010	.0003^{4}	.0008^{4}	.0020^{3}	.0035^{3}	.0056^{3}	.0086^{3}	.0118^{3}

at the left. The remaining seven columns list the empirically determined Type I error rates (proportion of significant chi-square tests) for samples ranging from N = 4 to N = 100, where $\phi = 0$ in all cases. These were obtained using CHI-PHI as in Figures 2b and 3b, except that each proportion was based on K = 10,000 contingency tables being sampled from the user-specified population.

The superscript listed adjacent to each error rate in Table 1 indicates the number of cells out of four in which the expected frequency was less than 10. Only 13 out of the 147 combinations tested (21 sets of marginal probabilities by 7 sample sizes) yield expected frequencies of 10 or greater in all four cells. Nevertheless, the combinations selected in Table 1 would seem to cover most applications of interest to the researcher, insofar as chi-square tests of independence in psychology rarely involve samples larger than 100 or smaller than 10 (we include N = 4 as a limiting case). Furthermore, the marginal probabilities listed in the first four columns cover the entire range of *unique* combinations of such probabilities, incremented in .10 steps, which might arise in a population.[4] Yet, the superscripts in the table show that very few instances in practice justify the use of a chi-square test of independence on 2 by 2 tables, if we take seriously the requirement of a minimum expected cell frequency of 10.

Fortunately, the Type I error rates listed in Table 1 provide virtually no support for the assumption that expected cell frequencies must be 10 or more for the chi-square test to provide an adequate approximation.[5] Excluding N = 4 for the moment, the empirically obtained error rates in Table 1 are never seriously inflated relative to the nominal error rate of .05. In particular, note the error rates listed in the N = 10 column of the table: In each and every case, all four cells of the contingency table have expected frequencies less than 10, and yet the error rates never exceed the nominal level of $\alpha = .05$ by more than .0051. Simulations using extremely unbalanced marginal probabilities demonstrate a strong *negative* bias; that is, the actual error rates are usually much smaller than the nominal error rate. The only strong positive bias observed occurs for certain combinations of marginal probabilities when N = 4, and this is certainly not a realistic sample size for most research applications. Furthermore, in 75 out of the 147 combinations tested, one or more cells of the 2 by 2 table had expected frequencies less than 1, again with no serious inflation in error rate except when N = 4. Consequently, the chi-square test of independence would appear to be an exceptionally robust test, far more so than is generally supposed. This conclusion is reached empirically, via the results of a Monte Carlo sampling program (CHI-PHI), rather than derivationally, a far more tedious process.[6]

Another valuable application of CHI-PHI, for either the student or the experienced researcher, is the use of this program for estimating the power of the chi-square test. If the simulations summarized in Table 1 had been conducted with $\phi \neq 0$, then the numbers entered in the seven right-hand columns would be hit rates (proportion of significant chi-square tests), rather than Type I error rates. In this case, the table would be a power

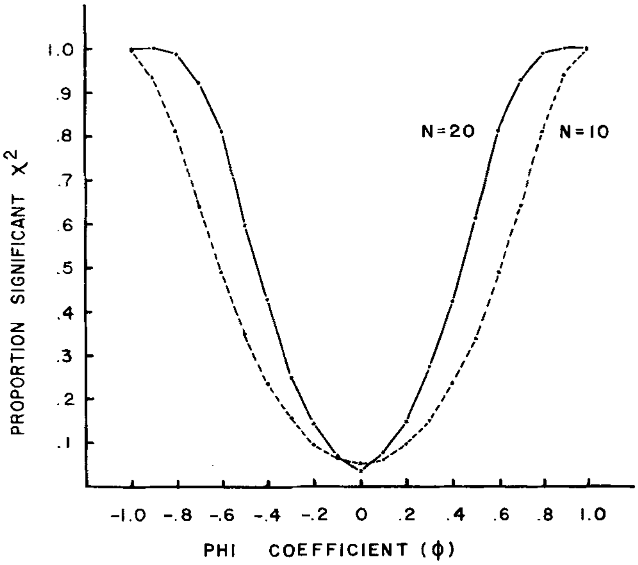


Figure 5. Empirically determined power levels for the chi-square test of independence on 2 by 2 tables as a function of sample size (N = 10 vs N = 20) and magnitude of association (phi ranging from −1 to 1).

table rather than an error-rate table. If a researcher is interested in estimating power for a particular application of chi square, then CHI-PHI may be used to establish power levels for the different sample sizes and values of phi under consideration. The researcher might then select that sample size which is adequate for detecting the minimum degree of association of practical or theoretical importance 75% of the time. Figure 5 presents two power curves (N = 10, N = 20) obtained by successively running CHI-PHI with marginal probabilities all equal to .5, and phi ranging from −1 to 1 in .10 increments. Each data point represents the proportion of significant chi-square tests obtained in that simulation. The proportions plotted in the N = 10 curve are based on K = 10,000 samples, whereas those plotted in the N = 20 curve are based on K = 1,000 samples. These results show, as expected, that power falls off as the degree of association approaches zero, and as sample size decreases (N = 10 vs N = 20). An association between attributes of $\phi = \pm.60$ will be detected about 50% of the time with a sample of N = 10, whereas samples of N = 20 increase the hit rate to around 81%. This general procedure permits the researcher to find a sample size which is adequate for detecting whatever minimum degree of association is of practical import for any particular configuration of marginal probabilities desired.

In conclusion, Monte Carlo sampling programs such as CHI-PHI have many practical uses, both for research and CAI applications. CHI-PHI is just one of a package of such programs we have been using in a CAI system for teaching introductory statistics. Other programs we have written generate power curves for the F test in the one-way analysis of variance, and evaluate the consequences of violating the assumptions of homogeneity of variance and normality of distribution of within-groups errors (Bradley, Hotchkiss, Dumais, & Shea, 1976). The use of empirical simulations to study the behavior and operating characteristics of inferential statistics under various conditions, including those representing substantial departures from the assumptions of the statistical models being applied, would appear to be a useful pedagogical device and a highly valuable research tool.

REFERENCES

Bradley, D. R., Hotchkiss, C. M., Dumais, S. T., & Shea, S. L. Computer assisted instruction in the small college. *Proceedings of the Seventh Conference on Computers in the Undergraduate Curricula*, 1976, 205-213.

Hays, W. L. *Statistics*. New York: Holt, Rinehart, & Winston, 1963.

Tate, M. W., & Hyer, L. A. Inaccuracy of the χ^2 test of goodness of fit when expected frequencies are small. *Journal of the American Statistical Association*, 1973, **68**, 836-841.

NOTES

1. Since the population value of phi is defined (after Hays, 1963, p. 604) as

$$\phi = \sqrt{\Sigma\Sigma\,\frac{[P(AB_{ij}) - P(A_i)P(B_j)]^2}{P(A_i)P(B_j)}}\,,$$

fixing the marginal probabilities $\{P(A_i), P(B_j)\}$ and the degree of association in the population (ϕ) defines the required joint-probability distribution $[P(AB_{ij})]$. Substituting Δp for $P(AB_{ij}) - P(A_i)P(B_j)$ in the formula above, and expanding, simplifying, and rearranging terms, we obtain:

$$\Delta p = \phi\sqrt{\frac{abcd}{bcd + acd + abd + abc}}\,,$$

where $a = P(A_1)P(B_1)$, $b = P(A_1)P(B_2)$, $c = P(A_2)P(B_1)$, and $d = P(A_2)P(B_2)$. The probability increment/decrement factor $[\Delta p = P(AB_{ij}) - P(A_i)P(B_j)]$ is the size of the difference required in each cell to produce the amount of association (ϕ) entered in the formula. Consequently, the joint-probability distribution is simply: $P(AB_{ij}) = P(A_i)P(B_j) \pm \Delta p$. Note that Δp is the same for all four cells because of the linear constraints pertaining to 2 by 2 tables (df = 1). Furthermore, Δp must be added in one set of diagonal cells and subtracted in the other in order to produce the desired distribution of joint probabilities, and to insure that they sum (within rows and columns) to equal the corresponding marginal probabilities (Figure 1a).

2. The program produces positive association by adding Δp to $P(A_1)P(B_1)$ and $P(A_2)P(B_2)$, and subtracting Δp from $P(A_1)P(B_2)$ and $P(A_2)P(B_1)$. Negative association is produced by the reverse operation. In either case, the resulting joint-probability distribution $[P(AB_{ij})]$ is used to construct a probability space from which a random-number generator samples the N observations.

3. Maximum values of −1 or +1 are possible only if $P(A_1)P(B_1) = P(A_2)P(B_2)$ or $P(A_1)P(B_2) = P(A_2)P(B_1)$, respectively, so that $P(AB_{ij}) = P(A_i)P(B_j) - \Delta p = 0$ in two

diagonal cells. This condition is met whenever both sets of marginal probabilities are balanced in the same way: $P(A_1)/P(A_2) = P(B_1)/P(B_2)$ or $P(A_1)/P(A_2) = P(B_2)/(B_1)$, since the arrangement of rows and columns is arbitrary except for determining the sign of phi. When $P(A_1) = P(A_2) = P(B_1) = P(B_2) = .50$, the maximum possible values of phi are ±1; otherwise, if the marginal probabilities are unequal but balanced similarly for rows and columns, then the maximum value of phi can be 1 or −1, but not both. Finally, when marginal probabilities are unequal and are not balanced in the same way for rows and columns, the maximum values of phi cannot equal either 1 or −1. Table 1 demonstrates these points by listing a variety of marginal probabilities, along with the maximum positive and negative values of phi which are possible for each particular combination.

4. Although the marginal probabilities in Table 1 do not include all possible *ordered* combinations, they do represent all possible joint-probability distributions which can result from marginal probabilities incremented in .10 steps, provided order is disregarded. Hence, marginal probabilities of .6, .4, .3, and .7, of .3, .7, .6, and .4, and of .3, .7, .4, and .6 all result in a joint-probability distribution (assuming independence) consisting of the following four probabilities: .42, .28, .18, and .12. The only difference between them is the cells to which the probabilities are assigned in the 2 by 2 table. However, as far as testing for possible inflation in the Type I error rate, it is inconsequential which cells receive which joint probabilities. Therefore, if an investigator has an estimate of the marginal probabilities in the population and wishes to determine the empirical Type I error rate for various sample sizes, all that is necessary is that he consult Table 1 and find the set of marginal probabilities most closely approximating his own, disregarding order. Hence, any of the above three sets of marginal probabilities can be evaluated for inflation in error rate by consulting the .6, .4, .7, .3 row of the table.

5. This is true with regard to using the continuous chi-square distribution to approximate a discrete multinomial distribution in terms of the areas in the tails of each distribution ($p < .05$). This does not imply that the approximation is adequate throughout the entire range of cumulative probabilities (Tate & Hyer, 1973). Since we are concerned here with the use of chi square as an inferential test, adequacy of approximation is of concern only for those α levels typically used in inference (i.e., $\alpha = .05$, $\alpha = .01$, $\alpha = .10$). We have conducted additional Monte Carlo simulations (reproducing Table 1) for tests conducted at the $\alpha = .01$ and $\alpha = .10$ levels of significance. The same overall results were obtained as reported in the text for $\alpha = .05$.

6. As a check on the accuracy of the empirically determined error rates reported in Table 1, several exact probabilities were computed for small N by manual expansion of the multinomial. In each case, the exact values were closely approximated by the empirical values. For example, manual expansion of the multinomial for marginal probabilities of .5, .5, .5, .5, and $N = 4$ produced an exact Type I error rate of .1094, which is very close to the empirical value (Table 1) of .1089; in fact, 95% confidence limits constructed about the former value include the latter. A similar check conducted on marginal probabilities of .9, .1, .9, .1, and $N = 4$ produced an exact value of .0226, as compared to the empirical value of .0237.