Statistical Approaches for Analyzing Mutational Spectra: Some Recommendations for Categorical Data

Walter W. Piegorsch* and A. John Bailer†

*Department of Statistics, University of South Carolina, Columbia, South Carolina 29208 and †Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056

Manuscript received March 10, 1993
Accepted for publication October 4, 1993

ABSTRACT

In studies examining the patterns or spectra of mutational damage, the primary variables of interest are expressed typically as discrete counts within defined categories of damage. Various statistical methods can be applied to test for heterogeneity among the observed spectra of different classes, treatment groups and/or doses of a mutagen. These are described and compared via computer simulations to determine which are most appropriate for practical use in the evaluation of spectral data. Our results suggest that selected, simple modifications of the usual Pearson X^2 statistic for contingency tables provide stable false positive error rates near the usual $\alpha = 0.05$ level and also acceptable sensitivity to detect differences among spectra. Extensions to the problem of identifying individual differences within and among mutant spectra are noted.

INCREASINGLY, scientists studying the effects of environmental stimuli are able to identify specific forms of mutagenic damage at the DNA or protein-product level. Experiments to study and compare the spectra of these mutations are performed in a variety of microbial and mammalian systems; recent examples include works by TINDALL, STEIN and HUTCHISON (1988), CARIELLO *et al.* (1990), DEMARINI *et al.* (1992), and LAMBERT *et al.* (1992), among many others. When the study concerns two or more groups of chemicals, different doses of the same chemical, different classes of effect modifiers (such as repair capacity, see the example below), etc., comparisons among the mutational spectra associated with each group or class are desired. These comparisons are facilitated by use of appropriate statistical methods. Such methods can vary, depending on both the nature of the spectral data and on the scientific questions being raised. For example, if the data take the form of scores or ratios of numerical values, standard univariate or multivariate analysis of variance methods can be employed (BENIGNI, PALOMBO and DOGLIOTTI 1992). If, alternatively, the data are counts of how often each mutation is observed for a given group, then statistical methods must take this categorical nature into account. In this case, one often displays the data as R rows of counts comprising a categorical data table, with each row representing a different mutant site. If there are T treatments or groups under study, the table is often referred to as an $R \times T$ contingency table; see, *e.g.*, WEIR (1990) or AGRESTI (1990).

In the special case of $T = 2$ groups under study, the issue becomes one of testing whether or not the responses in the ith row ($i = 1, \ldots, R$) across the two groups are similar for each row category. For example, consider the data in Table 1, which represent a comparison of G:C → A:T transition spectra in strains of the bacterium *Escherichia coli* that vary by excision repair capability after treatment with ethyl methanesulfonate (BURNS *et al.* 1986). Of interest is the comparison of DNA sequences that were ostensibly identical in the different strains prior to treatment. Table 1 presents the mutational counts in a 25 × 2 format (*i.e.*, $R = 25$), where the counts are seen to vary somewhat across nucleotide positions. The table also illustrates a critical aspect of the categorical data: the total number of mutations observed in each group may differ. Any comparison across rows must take these differing totals into account.

The statistical sampling characterization for data such as those in Table 1 involves the multinomial distribution (WEIR 1990, Ch. 2). ADAMS and SKOPEK (1987) describe the multinomial sampling format as applied to comparisons of mutational spectra. As they note, use of the terminology "mutational spectra" to describe such data is rather ambiguous. Technically, the data represent a per-group *sample* of observations from some underlying, unobserved spectrum. As the column totals increase and more mutations are observed in each group, the spectral sample would be expected to better represent the true, underlying spectrum.

In what follows, we describe various statistical methodologies for comparing spectral samples, where the scientific goal is to make inferences about the similarity or difference(s) among the T experimental groups. As will be seen, there are a number of test statistics applicable to $R \times T$ tables; one of our goals will be to

TABLE 1

Mutational spectra of G:C → A:T transitions in *E. coli*

Row, i [position/sequence]	Group 1: strain uvr^+	Group 2: strain $uvrB^-$
1 [42/CGT]	4	2
2 [53/TGT]	0	0
3 [56/CGC]	9	3
4 [57/TGC]	1	7
5 [75/AGA]	2	10
6 [80/TGA]	1	4
7 [84/GGT]	3	6
8 [90/GGA]	2	2
9 [92/CGG]	5	11
10 [93/CGC]	4	6
11 [95/CGT]	0	1
12 [104/TGG]	1	1
13 [113/TGG]	0	0
14 [120/AGA]	0	4
15 [129/CGT]	0	0
16 [140/AGT]	4	5
17 [174/GGG]	6	0
18 [179/CGG]	0	0
19 [185/GGC]	9	4
20 [186/TGC]	1	3
21 [188/TGT]	0	3
22 [191/TGT]	3	4
23 [198/CGC]	0	0
24 [201/GGC]	1	1
25 [206/TGT]	0	2
Total	56	79

Data from BURNS, ALLEN and GLICKMAN (1986). G → A transitions occur at middle G of each listed sequence.

compare their various operating characteristics in order to make specific recommendations for use in this setting.

STATISTICAL NOTATION

In the general $R \times T$ setting, we represent the data counts as variables Y_{ij} ($i = 1, \ldots, R; j = 1, \ldots, T$). The column totals, $N_j = \sum_{i=1}^{R} Y_{ij}$ ($j = 1, \ldots, T$), are assumed fixed and known. We also write the total sample size as $N_+ = \sum_{j=1}^{T} N_j$. For example, in Table 1, $R = 25$ and $T = 2$, while $N_1 = 56$, $N_2 = 79$, and the total sample size is $N_+ = 56 + 79 = 135$.

Under this format, we assume the jth column of data possesses a multinomial sampling distribution, with parameters p_{ij} and N_j; the proportion p_{ij} represents the response probability in the table's (i,j)th cell. A necessary structural aspect of the multinomial sampling model requires $p_{1j} + p_{2j} + \ldots + p_{Rj} = 1.0$, at each $j = 1, \ldots, T$.

An important feature of this multinomial model is the assumed independence among the observations. In particular, it is assumed that each tabled observation represents a sum of independent contributions to the total mutant count. The validity of this assumption will vary from experiment to experiment. For example, independence might be reasonable if cells scored for mutants are derived from a mixed population–such as a tissue–but not if they are derived clonally from a single progenitor cell. In all the examples and statistical methods described herein, we assume independence is valid. [In those cases where it may be invalid, and where additional, extra-multinomial sources of variability are present, more complex, hierarchical statistical models are required. These are discussed, *e.g.*, by BISHOP, FIENBERG and HOLLAND (1975, Ch. 12) or DIACONIS and EFRON (1985). A good practical exposition is given by KRAMER and SCHMIDHAMMER (1992).]

Under multinomial sampling, the statistical hypothesis of homogeneity,

$$H_0: p_{i1} = p_{i2} = \ldots = p_{iT} \quad (\text{for all } i),$$

translates to no differences among spectra across groups. Under this null hypothesis, the maximum likelihood estimates of the cell response probabilities are

$$\hat{p}_i^0 = \frac{\sum_{j=1}^{T} Y_{ij}}{\sum_{j=1}^{T} N_j},$$

$i = 1, \ldots, R$. Notice that the estimates do not require a column subscript under the homogeneity constraint. They satisfy the constraint in the sense that $\hat{p}_i^0$ estimates any p_{ij}, for fixed i, under H_0.

STATISTICAL TESTS OF SPECTRAL HOMOGENEITY

Chi-square tests: To test the significance of the homogeneity hypothesis, H_0, against any departure from homogeneity, a number of test statistics are available. Most common, perhaps, is the usual Pearson chi-squared goodness-of-fit statistic, which takes the form

$$X^2 = \sum_{i=1}^{R} \sum_{j=1}^{T} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}}.$$

The E_{ij} are the *expected* cell counts under the null hypothesis of homogeneity: $E_{ij} = \hat{p}_i^0 N_j$. Another well known test statistic for H_0 is the likelihood ratio statistic

$$G^2 = 2 \sum_{i=1}^{R} \sum_{j=1}^{T} Y_{ij} \log \left\{\frac{Y_{ij}}{E_{ij}}\right\}.$$

Both these forms can be shown to belong to a larger class of statistics, known as the power divergence family (CRESSIE and READ 1989). Another statistic from this class is

$$C^2 = \frac{9}{5} \sum_{i=1}^{R} \sum_{j=1}^{T} Y_{ij} \left\{\left(\frac{Y_{ij}}{E_{ij}}\right)^{2/3} - 1\right\},$$

which we will refer to as the CRESSIE-READ statistic.

All the statistics in the power divergence family possess approximately χ^2 distributions in large samples under the null hypothesis of no spectral differences (CRESSIE and READ 1984). Thus, as the column totals, N_j, become large, each of X^2, G^2, and C^2 may be referred to a χ^2 distribution with $(R-1)(T-1)$ degrees of freedom (d.f.): when the test statistic exceeds a χ^2 table value with $(R-1)(T-1)$ d.f., one rejects the null hypothesis of spectral homogeneity.

An alternative methodology for testing H_0, proposed by MARGOLIN and LIGHT (1974), involves a modified form of analysis of variance. The method mimics the usual ANOVA F-ratio adjusting for application to categorical data. The test statistic is

$$M^2 = \frac{(N_+ - 1)(R-1)\sum_{i=1}^{R}\sum_{j=1}^{T}\frac{(Y_{ij}-E_{ij})^2}{N_j}}{N_+ - \frac{1}{N_+}\sum_{i=1}^{R}\left(\sum_{j=1}^{T} Y_{ij}\right)^2}.$$

In large samples, M^2 is distributed as a weighted sum of χ^2 distributions. In some cases, this somewhat complex sum can itself be approximated by a single χ^2 distribution with $(R-1)(T-1)$ d.f. We will employ this approximation, and reject H_0 when M^2 exceeds a χ^2 table value with $(R-1)(T-1)$ d.f., in similar fashion to the power divergence statistics, above.

Modifying the degrees of freedom for zero row totals: An important consideration noted by MARGOLIN and LIGHT is that the d.f. must be delineated properly. If all the cells of the $R \times T$ table are filled with a non-zero entry (*i.e.*, $Y_{ij} \neq 0$ for all i,j), then there are $(R-1)(T-1)$ d.f. available in the table for statistical manipulation. Indeed, even if selected rows have only one cell with a non-zero entry, the table again provides $(R-1)(T-1)$ d.f. of information. Suppose, however, that there are one or more rows with a zero row total, *i.e.*, there is some row index i' for which

$$Y_{i'1} + Y_{i'2} + \ldots + Y_{i'T} = 0.$$

Then, that i'th row contributes no information to any of the test statistics. As such, one can eliminate those rows from consideration, and consequently *reduce the number of degrees of freedom*. Thus, if there are I rows of the $R \times T$ table that exhibit non-zero totals, then the corresponding $I \times T$ table of informative data will generate the same statistical information, but with a reduced number of d.f.: $(I-1)(T-1)$. Under such a reduction, the analysis is said to be *conditional* on the observed pattern of zero rows, since the statistical inferences from it apply technically to those data tables with the same observed pattern of non-zero row totals. MARGOLIN and LIGHT (1974) provide a series of theorems that help support this conditional reduction in d.f.; also see BISHOP *et al.* (1975, §5.2) or AGRESTI (1990, §7.7).

The effect of reducing the d.f. can be seen with the *E. coli* data from Table 1, where five of the observed row totals are zero. These contribute no information to any of the test statistics, but they nonetheless add $5 \times (2-1) = 5$ d.f. to the unconditional d.f. Under a conditional analysis, however, the d.f. in the table drop from 24 to 19, increasing the sensitivity of the test statistic. (We will illustrate this in more detail below.)

Large-sample tests based on a Normal reference distribution: The reference to a χ^2 distribution for P values or other statistical inferences from these various test statistics may not be appropriate in selected settings. Of greatest concern is the case when the $R \times T$ table exhibits many low or zero counts. This situation is known as "sparseness"; it is more properly characterized as a situation where many of the E_{ij} are small, especially if many are less than 1.0 (LEWONTIN and FELSENSTEIN 1965).

One approach that attempts to correct for sparseness involves determination of an alternative reference distribution for these statistics, or some functions thereof. For example, the Pearson X^2 statistic can be modified to the studentized form

$$Z_X = \frac{X^2 - (R-1)(T-1)}{\sqrt{2(R-1)(T-1)}}$$

(CRESSIE and READ 1984). An updated construction was suggested by ZELTERMAN (1987): take

$$D = X^2 - \sum_{i=1}^{I}\sum_{j=1}^{T}\frac{Y_{ij}}{E_{ij}},$$

and set $\mu_D = \frac{N_+}{N_+ - 1}(I-1)(T-1) - IT$. Also, calculate

$$\sigma_D^2 = \frac{2N_+}{N_+ - 3}\left\{\frac{(I-1)(N_+ - I)}{N_+ - 1} - \frac{N_+\sum_{i=1}^{I}(\sum_{j=1}^{T} Y_{ij})^{-1} - I^2}{N_+ - 2}\right\}$$
$$\left\{\frac{(T-1)(N_+ - T)}{N_+ - 1} - \frac{N_+\sum_{j=1}^{T} N_j^{-1} - T^2}{N_+ - 2}\right\}$$
$$+ \frac{4}{N_+ - 1}\left(\frac{N_+\sum_{i=1}^{I}(\sum_{j=1}^{T} Y_{ij})^{-1} - I^2}{N_+ - 2}\right)$$
$$\left(\frac{N_+\sum_{j=1}^{T} N_j^{-1} - T^2}{N_+ - 2}\right)$$

where I is the number of non-zero rows and the summations over $i = 1, \ldots, I$ are taken over only those non-zero rows. Then, the statistic

$$Z_D = \frac{D - \mu_D}{\sigma_D}$$

may also be employed to test H_0. (Notice that Z_D implicitly conditions its analysis on the number of non-zero rows; *i.e.*, it does not include any information from rows that have zero row totals, and it uses I as the number of rows for calculation.) Both Z_X and Z_D are referenced to a standard normal distribution; *i.e.*, rejection occurs when the absolute value of the statistic exceeds a two-sided standard normal table value.

"Exact" hypergeometric tests: One additional statistical construction that may be applied in testing spectral homogeneity involves so-called "exact" tests (WEIR 1990, pp. 76–77). The principle derives from the theories of R. A. FISHER (1935), who argued that testing for homogeneity in the special case of 2 × 2 contingency tables can be performed without call to large sample reference distributions such as χ^2. One simply conditions on the observed pattern of column *and* row totals to construct a test statistic.

Note that "conditioning" here refers to the same concept as described above for the case of zero row totals: one limits the statistical inferences to only those cases that exhibit the same conditional structure–patterns of zero row totals, or patterns of row and column totals, etc.–as that seen with the current set of data.

FISHER's test reports as its P value the probability of recovering a tabular configuration more extreme than that actually observed. The P value is based on the probability distribution of the tabular data after fixing both the row and column totals. This is known as a hypergeometric distribution, and is readily calculated in the 2 × 2 setting; see YATES (1984). An important point of clarification here is that this test does not insure a false positive error rate of exactly α, even when the test is performed nominally at α-level significance. (One typically constructs the test to insure that the true false positive rate is no larger than α; a "conservative" test of H_0.) The test is often called "exact," due to its exact calculation–as opposed to a large-sample χ^2 "approximation"–of the P value.

In the more general case of an $R \times T$ table, FISHER's "exact" hypergeometric construction may be extended in a straightforward manner: Simply identify some measure that integrates departure from homogeneity into a single quantity (such as X^2), and compute this measure for all possible $R \times T$ tables with the same row and column totals as the observed table. The "exact" P value is the sum of the hypergeometric probabilities corresponding to all those tables whose departure measures are more extreme (larger, *e.g.*, if the measure is X^2) than the observed table's.

When sample sizes grow very large, however, the calculations required for the exact P value can become unwieldy. Research into computational algorithms that improve the computational efficiency of the $R \times T$ exact test has produced a so-called "network" algorithm for calculation of the exact P value (MEHTA and PATEL 1983, 1986), and implementation of this algorithm is recommended when appropriate computer programming resources are available.

Alternatively, an approximation to the exact test involves computer simulation: one generates randomly a large number of $R \times T$ tables with the same marginal hypergeometric structure–*i.e.*, the same row and column totals–as the observed $R \times T$ table. Each random table is then compared to the original, observed table to determine if the random table exhibits greater departure from H_0. The proportion of randomly generated tables that exhibit this departure is an estimate of the true, "exact" P value (AGRESTI *et al.* 1979). This is known as a Monte Carlo approximation to FISHER's exact test. Many modern computer packages employ or recommend use of a Monte Carlo approximation when an $R \times T$ table becomes too large. A useful rule-of-thumb is to employ the Monte Carlo approximation when the total sample size exceeds $5(R - 1)(T - 1)$ (SAS Institute Inc.1985, PROC FREQ). Since this condition occurs fairly often in many of the settings we explore below, we will center attention on the Monte Carlo approximation to FISHER's exact test, as describe by ADAMS and SKOPEK (1987) or ROFF and BENTZEN (1989). Following ADAMS and SKOPEK (1987), we will refer to this as Fisher's hypergeometric test.

One important issue to note with FISHER's exact test in either its fully-computed form, or in its approximate, Monte Carlo form, is that it also conditions the analysis on the number of non-zero rows. This is a consequence of the test's conditioning on the observed pattern of row and column totals: since all those rows with zero row totals will be preserved under the hypergeometric structure, they add no information to the statistical analysis. Hence, the exact test imparts no additional information to any zero rows. Computations based on this information are restricted implicitly to the $(I - 1)(T - 1)$ d.f. provided by the non-zero rows.

COMPUTER SIMULATION COMPARISONS

All the statistical approaches for analyzing spectral homogeneity discussed above provide relatively stable and similar statistical inferences as sample sizes grow large without limit. In order to identify relative strengths and weaknesses among these approaches in small samples, we compared them by simulating various sets of $R \times T$ data tables with and without underlying differences in the spectral patterns among the T groups. Similar studies were reported, *e.g.*, by ROSCOE and BYARS (1971) and RUDAS (1986). For simplicity's sake, we limited the investigation to the case $T = 2$.

We applied each of the statistical methods to our simulated data, and recorded how often each method rejected the null hypothesis of homogeneity. When the null hypothesis of homogeneity was true, the

TABLE 2

Probability values for non-uniform mutational spectra in computer simulations

R	i:1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	0.05	0.05	0.2	0.2	0.5															
10	0.02	0.02	0.03	0.03	0.05	0.075	0.075	0.1	0.2	0.4										
15	0.01	0.01	0.02	0.02	0.03	0.03	0.03	0.05	0.05	0.075	0.075	0.1	0.1	0.1	0.3					
20	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.05	0.05	0.05	0.07	0.075	0.075	0.075	0.075	0.1	0.2

simulations provided an empirical estimate of each procedure's false positive rate. In those cases, the empirical rejection rates should be near the nominal significance rate, or "α level." We chose $\alpha = 5\%$. When the null hypothesis was false, however, the empirical rejection rates should be larger than the empirical significance level, and preferably larger than α. (This is statistical *power* to detect departures from the null hypothesis.)

To generate the simulated data sets, we selected four possible values for the number of categories R: $R = 5, 10, 15, 20$, and four possible values for the total sample size N_+: $N_+ = 20, 50, 100, 150$. We designed the computer simulations so that the two groups always had equal numbers of total mutants, *i.e.*, $N_1 = N_2 = N_+/2$, and chose two different patterns of spectral response for the p_{ij}: (a) uniform $p_{ij} = 1/R$ for all $i = 1, \ldots, R$ (j fixed), and (b) non-uniform p_{ij}'s varying from between 0.01 and 0.05 to between 0.2 and 0.5, depending on R. The non-uniform probabilities employed under pattern (b) are an attempt to mimic approximately spectral patterns seen in practice; they are given in Table 2. Notice that the sum of the individual p_{ij} across $i = 1, \ldots, R$ equals 1.0 for both the uniform and non-uniform patterns.

To compare false positive performance, both groups were assigned the same multinomial probabilities, either under the uniform or the non-uniform patterns. To compare powers, group $j = 1$ was assigned the uniform probability pattern and group $j = 2$ was assigned the non-uniform pattern from Table 2 (for fixed R). This led to 30 different homogeneous parameter configurations for comparison: four sample sizes × four row sizes × two probability patterns (when $N_+ = 20$, we did not simulate responses with $R = 20$), and to 15 heterogeneous "power" configurations. For each configuration, we generated 10,000 simulated $R \times 2$ tables under a multinomial sampling assumption, and counted the number of rejections of H_0 that were observed for each of the statistical methods discussed above. The resulting proportion of rejections is an estimate of each statistic's false positive rate under homogeneity between groups, or of its power under heterogeneity between groups. With 10,000 simulated samples per setting, the empirical false positive error rates themselves possess approximate standard errors of $\pm\sqrt{\alpha(1 - \alpha)/1000}$; near $\alpha = 0.05$, this is ± 0.002. For empirical powers, near true power = 50%, the approximate standard error is ± 0.005; near power = 90%, it is ± 0.003.

The multinomial variates were generated via an algorithm given by DEVROYE (1986, p. 559). For FISHER's hypergeometric test, we employed only the first 1000 simulated tables, and followed AGRESTI *et al.* (1979) to generate 1700 Monte Carlo hypergeometric tables for each of these 1000 simulations. To generate the hypergeometric tables we used the algorithm due to BOYETT (1979).

One additional feature incorporated into the simulations involved the issue of zero row totals and the effect of conditioning thereupon to reduce d.f. For all approaches except Fisher's hypergeometric test and Zelterman's Z_D (both of which condition implicitly on the number of non-zero rows), we computed the proportion of rejections under both an unconditional and a conditional number of d.f. That is, for the unconditional analyses we set the degrees of freedom equal to $(R - 1)(T - 1)$, which is simply $R - 1$ for these simulations. For the conditional analyses, we counted the number of non-zero rows, denoted by I, and set instead the d.f. and any other computations involving the number of rows for that table to $I - 1$.

The simulation results for estimated false positive errors under homogeneity are presented in Table 3. The table reports results for the following statistics: the three power divergence statistics X^2, G^2, C^2; the Margolin-Light statistic M^2; the studentized statistics Z_X and Z_D; and FISHER's hypergeometric test using either X^2 as the measure of departure in each $R \times 2$ table (denoted *hg-X*) as employed by ROFF and BENTZEN (1989), or the table's actual hypergeometric probability as the measure of departure (denoted *hg-P*) as employed by ADAMS and SKOPEK (1987). (We also evaluated FISHER's hypergeometric test using the C^2 statistic as the measure of departure. The empirical false positive rates under this measure were almost identical to those using X^2, and we do not present the hypergeometric C^2 rates here.) Table 3 also reports the average number of zero cells per $R \times 2$ table observed among the 10,000 simulated tables for each of the parameter configurations considered. This is a simple measure of the sparseness evidenced under each configuration. Larger numbers, relative to the

TABLE 3

Empirical false positive error rates under spectral homogeneity; nominal α-level = 5%

R	N_+		X^2	G^2	C^2	M^2	Z_X	Z_D	*hg-X*	*hg-P*	Sparse
(a) Inter-group homogeneity; uniform pattern											
5	20	unconditional d.f.	0.033	0.107	0.041	0.045	0.030				1.14
		conditional d.f.	0.035	0.110	0.044	0.044	0.037	0.036	0.034	0.051	
	50	unconditional d.f.	0.047	0.071	0.052	0.047	0.042				0.03
		conditional d.f.	0.047	0.071	0.052	0.047	0.042	0.046	0.045	0.041	
	100	unconditional d.f.	0.053	0.061	0.053	0.055	0.048				0
		conditional d.f.	0.053	0.061	0.053	0.055	0.048	0.046	0.048	0.050	
	150	unconditional d.f.	0.050	0.056	0.051	0.053	0.047				0
		conditional d.f.	0.050	0.056	0.051	0.053	0.047	0.046	0.044	0.044	
10	20	unconditional d.f.	0.004	0.076	0.005	0.048	0.004				6.91
		conditional d.f.	0.011	0.141	0.019	0.037	0.010	0.023	0.037	0.080	
	50	unconditional d.f.	0.035	0.117	0.044	0.049	0.026				1.46
		conditional d.f.	0.036	0.119	0.045	0.049	0.030	0.040	0.043	0.043	
	100	unconditional d.f.	0.045	0.079	0.051	0.052	0.037				0.10
		conditional d.f.	0.045	0.079	0.051	0.052	0.037	0.042	0.040	0.042	
	150	unconditional d.f.	0.048	0.065	0.052	0.051	0.042				0.01
		conditional d.f.	0.048	0.065	0.052	0.051	0.037	0.043	0.046	0.052	
15	20	unconditional d.f.	0.001	0.019	0.001	0.045	0.004				15.02
		conditional d.f.	0.003	0.153	0.008	0.022	0.002	0.016	0.034	0.141	
	50	unconditional d.f.	0.017	0.140	0.027	0.047	0.011				5.30
		conditional d.f.	0.023	0.162	0.033	0.044	0.020	0.033	0.051	0.050	
	100	unconditional d.f.	0.039	0.108	0.049	0.051	0.034				0.95
		conditional d.f.	0.039	0.109	0.049	0.051	0.032	0.042	0.044	0.041	
	150	unconditional d.f.	0.040	0.080	0.046	0.048	0.038				0.15
		conditional d.f.	0.040	0.080	0.046	0.048	0.038	0.046	0.046	0.052	
20	50	unconditional d.f.	0.007	0.138	0.012	0.049	0.006				11.02
		conditional d.f.	0.013	0.206	0.024	0.042	0.009	0.033	0.048	0.057	
	100	unconditional d.f.	0.028	0.146	0.042	0.048	0.026				3.04
		conditional d.f.	0.030	0.150	0.044	0.047	0.027	0.038	0.043	0.045	
	150	unconditional d.f.	0.042	0.113	0.052	0.052	0.033				0.83
		conditional d.f.	0.042	0.113	0.052	0.052	0.033	0.043	0.036	0.036	
(b) Inter-group homogeneity; non-uniform pattern (from Table 2)											
5	20	unconditional d.f.	0.010	0.049	0.014	0.069	0.011				2.83
		conditional d.f.	0.025	0.079	0.032	0.054	0.025	0.030	0.037	0.034	
	50	unconditional d.f.	0.032	0.068	0.037	0.072	0.033				1.10
		conditional d.f.	0.035	0.072	0.040	0.069	0.033	0.042	0.038	0.040	
	100	unconditional d.f.	0.044	0.075	0.048	0.075	0.040				0.33
		conditional d.f.	0.044	0.076	0.049	0.075	0.038	0.042	0.040	0.042	
	150	unconditional d.f.	0.045	0.067	0.048	0.073	0.046				0.09
		conditional d.f.	0.045	0.067	0.048	0.073	0.049	0.042	0.047	0.047	
10	20	unconditional d.f.	0.001	0.011	0.001	0.092	0.004				10.18
		conditional d.f.	0.011	0.084	0.017	0.060	0.010	0.020	0.041	0.070	
	50	unconditional d.f.	0.008	0.053	0.011	0.094	0.007				5.58
		conditional d.f.	0.022	0.097	0.027	0.083	0.018	0.028	0.041	0.044	
	100	unconditional d.f.	0.029	0.089	0.035	0.096	0.022				2.60
		conditional d.f.	0.035	0.101	0.043	0.093	0.028	0.035	0.043	0.046	
	150	unconditional d.f.	0.030	0.074	0.037	0.090	0.029				1.34
		conditional d.f.	0.032	0.078	0.039	0.089	0.033	0.036	0.054	0.059	
15	20	unconditional d.f.	0.001	0.002	0.001	0.087	0.013				17.68
		conditional d.f.	0.006	0.102	0.010	0.045	0.006	0.015	0.033	0.092	
	50	unconditional d.f.	0.002	0.044	0.004	0.091	0.007				10.49
		conditional d.f.	0.015	0.122	0.024	0.074	0.013	0.026	0.041	0.049	
	100	unconditional d.f.	0.015	0.077	0.020	0.095	0.014				5.65
		conditional d.f.	0.027	0.109	0.036	0.087	0.021	0.034	0.048	0.049	
	150	unconditional d.f.	0.026	0.082	0.030	0.086	0.019				3.48
		conditional d.f.	0.032	0.097	0.038	0.084	0.031	0.037	0.054	0.059	
20	50	unconditional d.f.	0.001	0.031	0.002	0.087	0.015				16.27
		conditional d.f.	0.014	0.148	0.024	0.064	0.010	0.028	0.048	0.057	
	100	unconditional d.f.	0.010	0.076	0.015	0.092	0.010				9.32
		conditional d.f.	0.024	0.138	0.035	0.083	0.017	0.032	0.043	0.045	

TABLE 3

Continued

R	N_+		X^2	G^2	C^2	M^2	Z_X	Z_D	*hg-X*	*hg-P*	Sparse
	150	unconditional d.f.	0.018	0.087	0.025	0.085	0.018				6.11
		conditional d.f.	0.026	0.115	0.035	0.080	0.023	0.039	0.036	0.036	

R, number of mutant categories; N_+, total sample size; X^2, Pearson goodness-of-fit statistic; G^2, likelihood ratio statistic; C^2, CRESSIE-READ statistic; M^2, Margolin-Light ANOVA statistic; Z_X, Studentized X^2 statistic; Z_D, Zelterman Studentized D statistic; *hg-X*, Monte Carlo hypergeometric test over 1700 pseudo-samples, using X^2 as measure of departure; *hg-P*, Monte Carlo hypergeometric test over 1700 pseudo-samples, using hypergeometric Probability as measure of departure; and "sparse" indicates average number of zero cells per table.

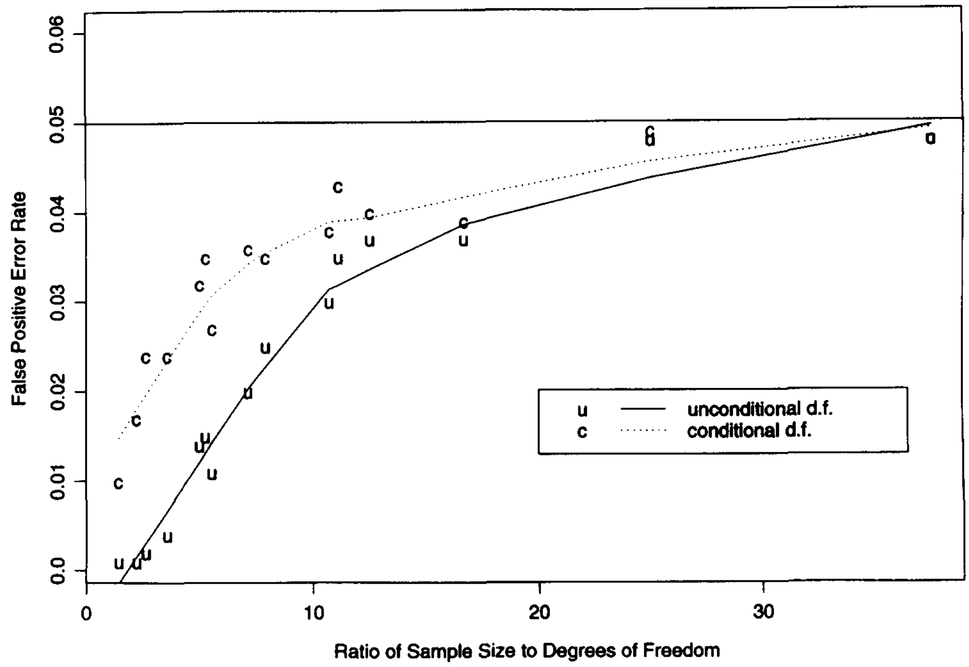

FIGURE 1.—Empirical false positive error rates from Table 3 for CRESSIE-READ C^2 statistic as a function of sample size-to-(unconditional) d.f. ratio, $N_+/(R-1)$. Comparison is made between unconditional and conditional d.f. calculations under non-uniform pattern of response probabilities (Table 2). Nominal α level is 0.05 (horizontal line: ——).

total number of cells, $2R$, suggest increasing sparseness.

The results from Table 3 are somewhat varied. Of greatest significance is the observation that, in general, tests performed using the conditional d.f., $I-1$, appear more stable than the unconditional tests. (We define "stable" to be false positive errors that are either close to the nominal α-level, or below it–"conservative"–if departing by more than 0.5% or so from α.) This stability is best seen by plotting the estimated false positive rate as a function of the ratio of sample size to unconditional d.f., $N_+/(R-1)$. The ratio is a summary measure of the potential sparseness in a given table or experiment. Small levels of $N_+/(\mathrm{R}-1)$ suggest greater potential sparseness, hence greater potential instability in a test statistic's performance. Figure 1 illustrates the effect with the CRESSIE-READ statistic, C^2, using the non-uniform patterns from Table 2. The estimated rates begin below the nominal level of $\alpha = 0.05$, and gradually move closer to it as $N_+/(R-1)$ increases. This increase is more pronounced, however, when employing the conditional d.f. Similar effects occur for most of the statistics studied in Table 3.

Among the results under conditional d.f., greatest stability in false positive error rates at levels of $N_+/(R-1)$ above about 10 is seen with Fisher's hypergeometric tests (*hg-P* and *hg-X*), the Pearson X^2 and CRESSIE-READ C^2 statistics, and the Studentized Z_D statistic. This is illustrated in Figure 2 for the non-uniform pattern of response probabilities. The figure also illustrates that small values of $N_+/(R-1)$ tend to drive false positive error rates below $\alpha = 0.05$ for most of these methods. As might be expected, however, the hypergeometric tests' false positive rates remain near the nominal 5%-level, although in selected cases with very sparse tables, they exhibit some instability. [The probability-computed form (*hg-P*) appears slightly less stable than the X^2-computed form (*hg-X*) in Table 3, hence only the latter is presented for comparison in Figure 2.]

Table 3 also indicates that the MARGOLIN-LIGHT M^2 statistic exhibits very strong stability under the uniform pattern (a), but this is not maintained when the within-group probabilities differ drastically, as in pattern (b). Thus the simple χ^2 approximation to the true, weighted sum of χ^2's representing the limiting distribution of M^2 may be a poor one in selected

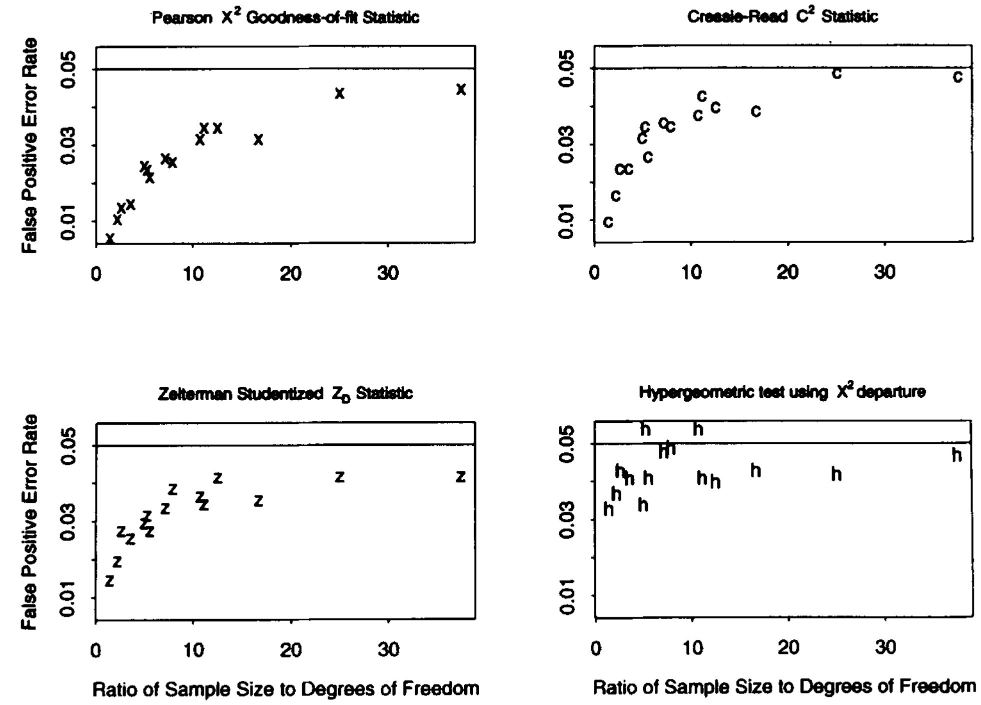

FIGURE 2.—Empirical false positive error rates from Table 3 for Monte Carlo hypergeometric test (*h*) using X^2 measure of departure, and CRESSIE-READ C^2, Person X^2, and Zelterman Z_D statistics as a function of sample size-to-(unconditional) d.f. ratio, $N_+/(R - 1)$. Rates are calculated for conditional d.f. and nonuniform pattern of response probabilities (Table 2). Nominal α level is 0.05 (horizonal line: ——).

settings. We do not recommend its use for testing mutational spectra, although development of the appropriate weighted sum of χ^2's for the true limiting distribution of M^2 may be a rich avenue of further research in this area; see, *e.g.*, TANECHI (1988). Similarly, the likelihood ratio statistic, G^2, exhibits generally poor performance at almost every level and setting studied. This is not surprising; many authors have noted that the G^2 statistic can exhibit extremely high false positive rates in small samples when referred to a χ^2 distribution; see CRESSIE and READ (1989). We will note an alternative to the χ^2 reference distribution when employing the G^2 statistic for some extensions discussed below.

Table 4 presents estimated powers under spectral heterogeneity for a limited subset of the statistics and cases presented in Table 3. Only tests that condition on the non-zero rows are presented, and among these only those tests that exhibited relatively stable false positive errors (from Table 3) are studied in detail. These restrictions led to consideration of the following statistics: X^2, C^2, Z_D, Z_X, and *hg-X*. The table illustrates superiority in power for the Monte Carlo hypergeometric (*hg-X*) test, although at values of $N_+/(R - 1)$ above about 7 the zero-row-conditional C^2 power divergence statistic exhibits performance comparable with the *hg-X* test.

All of the statistics considered herein perform better with increasing sample size and, in particular, with increasing $N_+/(R - 1)$. If $N_+/(R - 1)$ is greater than about 10, the C^2 statistic and the Z_D statistic perform adequately enough to recommend their use. (Use Z_D if 25% or more of the data table's cells exhibit zero counts.) If $N_+/(R - 1)$ is less than 10, and if appropriate computer resources are available, consider instead hypergeometric tests, such as *hg-X*. The C^2 and Z_D statistics are still reasonable choices for small $N_+/(R - 1)$, if, *e.g.*, digital computing resources are not available. They may not be as powerful as the hypergeometric tests in these cases, however.

If $N_+/(R - 1)$ is less than about 5, then an "exact" version of FISHER's test (MEHTA and PATEL 1983) can be considered as well.

EXAMPLE

To illustrate use of these test statistics, consider again the data from Table 1. Recall that these are counts of G:C → A:T transition mutants in $T = 2$ different strains of *E. coli* after treatment with ethyl methanesulfonate (BURNS *et al.* 1986). As the table indicates, although $R = 25$ different sites were displayed by the authors, only $I = 20$ sites exhibited one or more mutants for either strain. Conditioning the analysis on these non-zero rows yields $(20 - 1)(2 - 1) = 19$ d.f. The ratio of sample size to unconditional d.f., $N_+/(R - 1) = 135/25 = 5.6$, is low enough that use of the Monte Carlo version of FISHER's hypergeometric tests may be warranted, although for illustrative purposes we will also compute the C^2 and Z_D statistics.

The null hypothesis of homogeneity between the two spectra is H_0: $p_{i1} = p_{i2}$ (for all i). Table 5 gives the expected cell counts, $E_{ij} = \hat{p}_i^0 N_j$. Notice that a number of these values are below 1.0, suggesting a fair degree of sparseness for these data: 5/50 cells, or 10%, of the original table exhibit zero observed counts. Again, FISHER's hypergeometric test appears warranted. Applying the Monte Carlo form of the test with X^2 as the

TABLE 4

Estimated powers (conditional d.f.) under spectral heterogeneity; nominal α-level = 5%

R	N_+	$N_+/(R-1)$	X^2	C^2	Z_D	Z_X	*hg-X*
5	20	5	0.181	0.203	0.191	0.184	0.221
	50	12.5	0.596	0.611	0.590	0.591	0.605
	100	25	0.925	0.928	0.919	0.920	0.920
	150	37.5	0.992	0.992	0.990	0.991	0.989
10	20	2.2	0.106	0.149	0.165	0.091	0.225
	50	5.6	0.552	0.589	0.584	0.519	0.637
	100	11.1	0.939	0.944	0.935	0.927	0.944
	150	16.7	0.995	0.996	0.995	0.994	0.996
15	20	1.4	0.034	0.060	0.086	0.027	0.168
	50	3.6	0.286	0.348	0.343	0.269	0.404
	100	7.1	0.767	0.793	0.771	0.739	0.788
	150	10.7	0.953	0.958	0.953	0.950	0.968
20	50	2.6	0.140	0.203	0.196	0.105	0.266
	100	5.3	0.543	0.598	0.550	0.491	0.622
	150	7.9	0.833	0.857	0.817	0.801	0.838

See Table 3 for abbreviations and symbols.

TABLE 5

Observed and expected cell counts for Table 1

Non-zero row, i [position/sequence]	Group 1		Group 2	
	Observed, Y_{i1}	Expected, E_{i1}	Observed, Y_{i2}	Expected, E_{i2}
1 [42/CGT]	4	2.49	2	3.51
2 [56/CGC]	9	4.98	3	7.02
3 [57/TGC]	1	3.32	7	4.68
4 [75/AGA]	2	4.98	10	7.02
5 [80/TGA]	1	2.07	4	2.93
6 [84/GGT]	3	3.73	6	5.27
7 [90/GGA]	2	1.66	2	2.34
8 [92/CGG]	5	6.64	11	9.36
9 [93/CGC]	4	4.15	6	5.85
10 [95/CGT]	0	0.41	1	0.59
11 [104/TGG]	1	0.83	1	1.17
12 [120/AGA]	0	1.66	4	2.34
13 [140/AGT]	4	3.73	5	5.27
14 [174/GGG]	6	2.49	0	3.51
15 [185/GGC]	9	5.39	4	7.61
16 [186/TGC]	1	1.66	3	2.34
17 [188/TGT]	0	1.24	3	1.76
18 [191/TGT]	3	2.90	4	4.10
19 [201/GGC]	1	0.83	1	1.17
20 [206/TGT]	0	0.83	2	1.17
Total	56	55.99	79	79.01

measure of departure as described above (*hg-X*) yields an estimated P value of 0.007. Since the estimated P can change, depending on form of random number generator, seed values, etc., one should also calculate a confidence interval on the estimated P value; *e.g.*, at 99% confidence with 1700 Monte Carlo pseudo-samples, use $P \pm 2.58\sqrt{P(1-P)/1700}$. Here, such a 99% confidence interval on the Monte Carlo estimated P of 0.007 is $0.002 < P < 0.012$.

The values in Table 5 also yield the test statistics $C^2 = 36.18$ and $|Z_D| = |(-4.05) - (-20.86)|/\sqrt{28.85} = 3.13$. On 19 d.f., the χ^2 reference distribution for the C^2 statistic provides a P value of 0.010. The P value for Z_D based on a standard normal distribution is 0.002. Thus all three methods suggest significant departure from homogeneity between the two spectra.

EXTENSION: MULTIPLE COMPARISONS AMONG SITES AND SPECTRA

The tests for spectral homogeneity described above are global in nature: when significant, they suggest evidence for differences among the column variables (spectra) without specifying where such differences may lie. It is well known, however, that the mutability

TABLE 6

Mutational spectra of single base-pair substitutions in *S. cerevisiae*

Row, *i* [site]	Group 1: strain *RAD3*	Group 2: strain *rad3-102*
1 [18]	17	1
2 [27]	2	8
3 [29]	5	7
4 [64]	0	7
5 [83]	11	4
6 [88]	10	0
7 [90]	121	56
Total	166	83

Data from MONTELONE *et al.* (1992).

of DNA varies substantively across sites in a genome (FOSTER, EISENSTADT and CAIRNS 1982), and an important question left unanswered by the global tests is precisely where or between which spectra the differences, if any, exist. For example, can we identify potential mutational "hot spots" within a target gene, or can we form sub-tables of the full $R \times T$ table that indicate associated forms of heterogeneity in the mutant spectra?

Consider, *e.g.*, the data in Table 6, which are counts of single base pair mutations in $T = 2$ strains of the yeast *Saccharomyces cerevisiae at* $R = 7$ different sites (MONTELONE *et al.* 1992). For this 7 × 2 table, $C^2 = 39.18$ ($P < 0.001$ on 6 d.f.); very significant departure from homogeneity is evidenced. Of additional interest in this study, however, is the identification of those sites that differed significantly between the two strains. For example, MONTELONE *et al.* (1992) reported that the 3 × 2 *sub*-table consisting of only sites 29, 83, and 90 (*i.e.*, rows 3, 5, and 7 in Table 6) was the largest sub-table constructed from the full table that exhibited evidence of spectral homogeneity. This suggests that the other four sites drive the spectral heterogeneity seen in the full table. Statistically, one wishes to determine the α-level at which these differences are significant.

Clearly, repeated, unadjusted application of C^2, X^2, or any other statistical method to recursive sub-tables is a form of multiple comparison. As is well known, repeating comparisons at individual significance levels of, say, $\alpha = 0.05$ drives the experiment-wise false positive rate above 0.05, and, in some instances, well above it (WEIR 1990, pp. 109–110). To correct for such false positive error inflation, *simultaneous inference* procedures are required.

For the specific problem of testing homogeneity across multiple sub-tables of an $R \times T$ contingency table, one can apply a simultaneous method due to GABRIEL (1966). In the method's general form, one supposes that a particular sub-table with $A \leq I$ rows and $B \leq T$ columns is of interest. (Notice that the inference is conditional on the $I \leq R$ non-zero rows.) For simplicity of presentation, we will suppose that these are the first $i = 1, \ldots, A$ rows and the first $j = 1, \ldots, B$ columns of the full table. We also require $A \geq 2$ and $B \geq 2$. Of interest is whether the $A \times B$ sub-table exhibits homogeneity, *i.e.*, if

$$H_0^{AB}\colon p_{il} = p_{i2} = \ldots = p_{iB} \quad (\text{for all } i = 1, \ldots, A),$$

holds.

Obviously, if this particular sub-hypothesis were the only hypothesis of interest, one would simply apply the method(s) recommended above to test the AB-fold homogeneity. GABRIEL's (1966) approach applies, however, if one is simultaneously testing any or all of the possible $A \times B$ sub-tables constructed from the full $I \times T$ table. It employs the log-likelihood statistic, G^2, noted above: for any $A \times B$ sub-table, compute the associated sub-statistic, say G^2_{AB}, which can be written in closed form as

$$G^2_{AB} = 2\left\{\sum_{i=1}^{A}\sum_{j=1}^{B} Y_{ij}\log(Y_{ij}) - \sum_{i=1}^{A}\left[\sum_{j=1}^{B} Y_{ij}\right]\log\left(\sum_{j=1}^{B} Y_{ij}\right) - \sum_{j=1}^{B}\left[\sum_{i=1}^{A} Y_{ij}\right]\log\left(\sum_{i=1}^{A} Y_{ij}\right) + \left[\sum_{i=1}^{A}\sum_{j=1}^{B} Y_{ij}\right]\log\left(\sum_{i=1}^{A}\sum_{j=1}^{B} Y_{ij}\right)\right\}.$$

Reject H_0^{AB} when G^2_{AB} exceeds the upper-α quantile of a χ^2-distribution with $(I - 1)(T - 1)$ d.f. Notice that the d.f. are those employed from the original $I \times T$ table, *and that they do not change*, regardless of the number or form of sub-tables tested. GABRIEL shows that this approach restricts the experiment-wise false positive rate to no more than α as the smallest $N_j \rightarrow \infty$. In addition, the overall inference possesses a form of *coherence*: when any $A \times B$ sub-table is seen to exhibit heterogeneity based on G^2_{AB}, so must all larger sub-tables that contain the $A \times B$ heterogeneous sub-table.

These inferences hold for all possible $A \times B$ sub-tables due to their simultaneous construction. Thus, as GABRIEL (1966, p. 1082) notes, with this method "it is in no way necessary to decide *a priori*, *i.e.*, before seeing the data, what combinations [of sub-tables] are to be tested, and one may be guided by the data themselves in selecting what to test." (In most data-analytic settings, such *a posteriori*, data-driven comparisons are strictly forbidden, due to the false positive error inflation they engender.) The cost of this *a posteriori* luxury is that the procedure is inherently conservative. That is, since it protects against false positive errors across all possible sub-tables, the false positive rate for only a selected sub-collection of $A \times B$ tables will be generally less than α. This is countered, however, by the fact that in small samples, the G^2 statistic's false positive rate often exceeds α. Thus, there is the possibility that in practice, these two conflicting features will counterbalance one another.

Unfortunately, GABRIEL's method for testing row

TABLE 7

Empirical false positive error rates for likelihood ratio statistic G^2 at large sample sizes; conditional d.f. only, nominal α-level = 5%

R	N_+	$N_+/(R-1)$	False positive rate
(a) Inter-group homogeneity; uniform pattern from Table 3			
5	200	50.0	0.030
	250	62.5	0.028
10	200	22.2	0.040
	250	27.8	0.035
(b) Inter-group homogeneity; non-uniform pattern from Table 3			
5	200	50.0	0.033
	250	62.5	0.031
10	200	22.2	0.049
	250	27.8	0.048

deletions is tied to the G^2 form for the test statistic. Its desirable property of coherence, noted above, does not hold if the more stable C^2 or X^2 statistics are calculated instead.

Selected recommendations can be made regarding application of the G^2 statistic. For instance, for very large values of $N_+/(R-1)$, the G^2 statistic does behave in a stable fashion. To illustrate this, we recalculated the simulations from Table 3 at $N_+ = 200$, 250, and recorded the empirical false positive rates for the G^2 statistic using conditional d.f. at $\alpha = 0.05$ for $R = 5$, 10. This gave values for $N_+/(R-1)$ of at least 22.2. The resulting false positive rates are given in Table 7. As can be seen therein, G^2's empirical false positive rate for $N_+/(R-1) > 20$ does exhibit reasonable stability.

For cases where $N_+/(R-1) < 20$, a small-sample adjustment to the χ^2 quantile can be employed instead: rather than reject when G^2_{AB} exceeds the upper-α quantile of a χ^2-distribution with $(I-1)(T-1)$ d.f., reject when $G^2_{AB}/(I-1)(T-1)$ exceeds the upper-α quantile of an F-distribution with $(I-1)(T-1)$ and $N_+ - (I-1)(T-1) = N_+ - 1 + I + T - IT$ d.f. This has the effect of reducing the rate at which G^2 rejects each null hypothesis, providing greater stability in terms of closer-to-nominal false positive error rates. To illustrate this, we recomputed the simulated false positive rates for the G^2 statistic using the F-quantile, and compared them to the χ^2-based rates from Table 3. The results, using conditional d.f. and the non-uniform patterns in Table 2, are compared in Figure 3 as a function of $N_+/(R-1)$. The closer-to-nominal error rates under the F-quantile illustrate the improvement available.

Example using *S. cerevisiae* data: As an example of the GABRIEL method's use, consider again the *S. cerevisiae* data (MONTELONE *et al.* 1992) in Table 6. Recall that the overall hypothesis of 7 × 2 homogeneity was rejected by the C^2 statistic ($P < 0.001$). The G^2 statistic is also very significant: $G^2 = 44.62$ ($P < 0.001$; since $N_+ = 249$ and $N_+/(R-1) = 41.5$ are so large with these data, we are employing χ^2 as the reference distribution, rather than the F-distribution approximation noted above.) At $\alpha = 0.05$, the χ^2 table value with $(7-1)(2-1) = 6$ d.f. is 12.59. Thus, any $A \times 2$ sub-table whose associated G^2_{A2} statistic is larger than 12.59 also exhibits significant heterogeneity among its A rows according to the GABRIEL procedure. The 4 × 2 sub-table consisting of rows 1, 2, 4, and 6 suggested by the study's authors is one such collection: its G^2_{42} statistic is 40.84. Conversely, the complementary sub-table of rows 3, 5, and 7 generates a G^2_{32} statistic of 3.66, which is clearly non-significant. As the study's authors noted (MONTELONE *et al.* 1992), the spectral heterogeneity evidenced in the original table appears due to significant heterogeneity at the four rows, 1, 2, 4, and 6 (*i.e.*, sites 18, 27, 64, and 88) at $\alpha = 0.05$.

Example using *E. coli* data: Application of GABRIEL's method to the *E. coli* data from Table 5 produces a rather different set of inferences. Recall that the overall hypothesis of 20 × 2 homogeneity for these data was rejected by the C^2 statistic ($P = 0.010$). As expected, G^2 is also significant: $G^2 = 41.88$ ($P = 0.005$ for comparing $41.88/19 = 2.20$ against an F reference distribution with 19 and 116 d.f.).

With these data, $N_+/(R-1) = 5.6$ is low enough that an F-distribution adjustment to G^2 may help reduce potential inflation in false positive error; see Figure 3. If $\alpha = 0.05$, one employs the upper 0.05 table value from the $F_{19,116}$ distribution, which is 1.68. Thus rejection occurs if $G^2/19$ is greater than 1.68, or, alternatively, if $G^2 > 31.86$. GABRIEL's method requires that the G^2_{AB} statistic from any $A \times B$ sub-table must exceed this F-based value in order to be considered heterogeneous.

Notice then that in Table 5, by removing the row at $i = 14$ corresponding to position #174/sequence GGG, the G^2 statistic from the associated 19 × 2 table drops to 30.92. Employing the F-based quantile, the method suggests that the reduced table exhibits homogeneity, since $30.92 \not> 31.86$. No other row can be removed from the original table, and still reduce the corresponding 19 × 2 G^2 statistic to below 31.86. For example, if we force row 14 to remain in the table, the greatest reduction is achieved by removing row 2 in Table 5: this yields a 19 × 2 G^2 statistic of 35.77. This suggests that the row 14 data appear heterogeneous relative to the rest of the original table, but that we are unable to distinguish heterogeneity among the other 19 rows, at $\alpha = 0.05$.

Multiple comparison adjustments for collapsing contingency tables: Corrections for performing multiple comparisons with contingency tables are many and varied. For example, rather than delete rows or columns from the original $I \times T$ table to construct sub-tables for multiple comparison, one may wish to

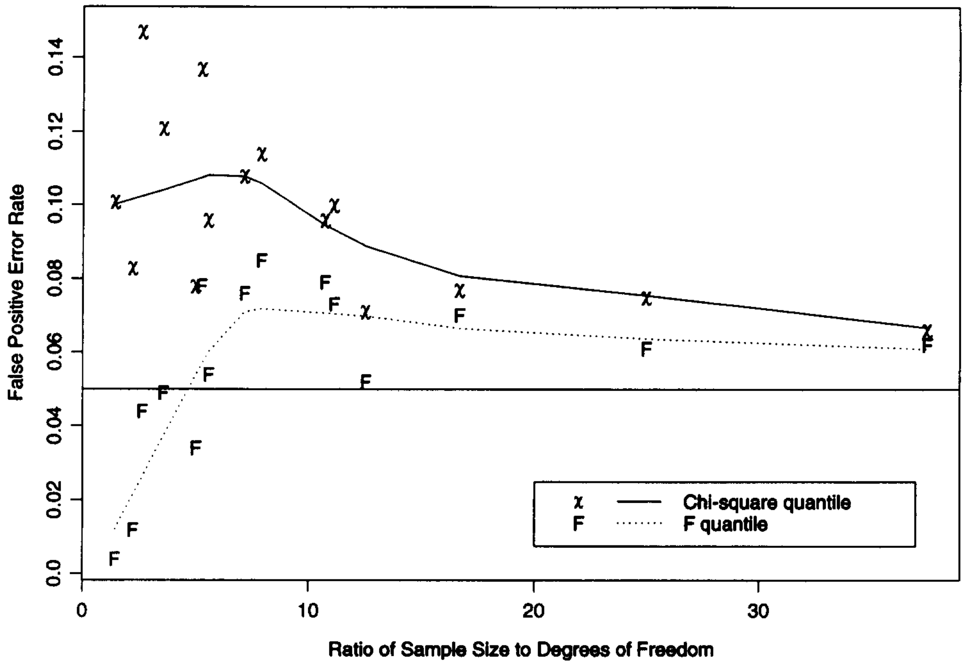

FIGURE 3.—Empirical false positive error rates for likelihood ratio G^2 statistic as a function of sample size-to-(unconditional) d.f. ratio, $N_+/(R-1)$. Rates are calculated for conditional d.f. and under non-uniform pattern of response probabilities (Table 2). Comparison is made between χ^2 (χ) and F-distribution (F) reference quantiles for G^2 rejection. Nominal α level is 0.05 (horizonal line: ——).

combine or collapse rows or columns and perform comparisons among various forms of collapsed sub-tables; see ADAMS and SKOPEK (1987) or WEIR (1990, pp. 110–111). In this case, it is the Pearson X^2 statistic, rather than the G^2 statistic, that should be employed to measure the departure from homogeneity in each sub-table. Such a simultaneous collapsing method for $I \times T$ tables has been described by GILULA (1986) and GILULA and KRIEGER (1989): begin by calculating the X^2 statistic for the original $I \times T$ table, denoted by X^2_{IT}. As above, reject the overall null hypothesis of $I \times T$ homogeneity if X^2_{IT} exceeds a χ^2 quantile with $(I-1)(T-1)$ d.f. Then, for any collapsed table of dimension $A \times B$, reject homogeneity within that table if the associated X^2 statistic, X^2_{AB}, shows a significant reduction in χ^2, *i.e.*, if $X^2_{IT} - X^2_{AB}$ exceeds an α-level χ^2 table value with $(I-1)(T-1)$ d.f. In the same manner as GABRIEL's simultaneous method for row deletions (above), the original d.f. from the full table are employed in the χ^2 table value to correct for multiple comparisons across all possible collapsings. Even if the experimenter chooses to inspect the data *a posteriori* and select rows or columns for collapsing based on this inspection, the experiment-wise false positive rate is held at α under this approach.

Of course, further research is needed to identify and study other methods that can identify specific spectral differences with good power, and yet correct for multiplicity and associated false positive error inflation. In this endeavor, simplicity and ease of application are additional, important criteria to keep in mind, since methods requiring prohibitive sample sizes or extensive computer resources face greater obstacles to their implementation.

DISCUSSION

The problem of assessing differences in mutational spectra is of growing interest, and experiments to detect such differences will multiply with further advances in microbiology and biotechnology. Statistical methods to facilitate these analyses are important, and our goal has been to identify such methods for data in categorical form, as illustrated in Tables 1 and 6. As we noted, many different forms exist for the test statistics. Based on the conditions and forms we have evaluated, we conclude that four important features require recognition when analyzing tables of mutational spectra: first, the simple chi-square approximation to the common X^2 or G^2 statistics is not reliable. Second, the natural extension of FISHER's exact test to $R \times T$ tables–a hypergeometric test described herein in its computer-intensive Monte Carlo implementation (AGRESTI, WACKERLY and BOYETT 1979; ADAMS and SKOPEK 1987; ROFF and BENTZEN 1989)–exhibits stable false positive error properties, and also good power to detect global differences among truly divergent spectra. We recommend the exact test's use for comparing mutational spectra in categorical form. When the "exact" formulation is computationally intractable, we found the Monte Carlo form with the X^2 statistic as the measure of departure to be a stable and fairly powerful test of spectral homogeneity. Third, when sample sizes are large enough, as measured when the ratio of total sample size to d.f. exceeds about 10, a modified goodness-of-fit statistic, the CRESSIE-READ C^2, also performs well, and may be recommended when computational resources do not allow for calculation of the exact test. Fourth, in sparse tables where the percentage of zero cells is greater than about 25%, a Studentized statistic, given above as Z_D, exhibits roughly similar characteristics to the C^2 statistic. We recommended its use in such cases.

In passing, we should note that other simulation results and analytical results exist that provide limited corroboration of the simulation values we achieve herein, particularly for the performance of the X^2 statistic. For example, direct comparison of our empirical false positive errors for X^2 (Table 3) can be made with the simulation results of ROSCOE and BYARS (1971, Table 4). Those authors reported an estimated error for X^2 of 0.0296 when $R = 5$ and $N_+ = 20$ (in our notation), and an error of 0.0508 when $R = 5$ and $N_+ = 100$ at the nominal $\alpha = 0.05$ level. These compare favorably with our own results of 0.033 and 0.053, respectively.

Further results on the small sample powers of the X^2 and C^2 statistics were reported by READ (1984, Table 1). Although not directly comparable with our parameter configurations, READ's calculations do show some similarities with selected powers that we achieved for X^2 and C^2. His "alternative 1" configuration is roughly similar to our heterogeneity model at $R = 5$, with his calculated powers for X^2 and C^2 given as 0.6997 and 0.6890, respectively. The roughly comparable setting from our Table 4 yields powers of 0.596 and 0.611. While not precisely equivalent, the values do suggest a reasonable level of similarity between the calculated and simulated results.

For the problem of identifying specific differences among selected sub-tables of the original $I \times T$ table, we have illustrated the use of a simultaneous inference procedure due to GABRIEL (1966). The method allows the data analyst to compare various collections of rows and columns from the full table, and even select these after examining the data, while still retaining α-level significance. A similar method that allows for collapsing rows or columns within the full data (GILULA 1986) was also noted. Both can be recommended for use.

This research was initiated while the first author was with the Statistics and Biomathematics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina. Thanks are due to numerous colleagues for their encouraging suggestions, including ALAN AGRESTI, DOUGLAS BELL, NEAL CARIELLO, JAN DRAKE, IRENE JONES, NORMAN KAPLAN, BARRY MARGOLIN, MICHAEL SHELBY, DAVID UMBACH and THOMAS ADAMS. Gracious assistance came from the Associate Editor, and the helpful suggestions of three anonymous reviewers are gratefully acknowledged. The computing and support services of STEVEN GARREN are also greatly appreciated.

LITERATURE CITED

ADAMS, W. T., and T. R. SKOPEK, 1987 Statistical test for the comparison of samples from mutational spectra. J. Mol. Biol. **194:** 391–396.

AGRESTI, A., 1990 *Categorical Data Analysis*. John Wiley, New York.

AGRESTI, A., D. WACKERLY and J. M. BOYETT, 1979 Exact conditional tests for cross-classifications: approximations of attained significance levels. Psychometrika **44:** 75–83.

BENIGNI, R., F. PALOMBO and E. DOGLIOTTI, 1992 Multivariate statistical analysis of mutational spectra of alkylating agents. Mutat. Res. **267:** 77–88.

BISHOP, Y. M. M., S. E. FIENBERG and P. W. HOLLAND, 1975 *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.

BOYETT, J. M., 1979 Random $R \times C$ tables with given row and column totals. Appl. Statist. **28:** 329–332.

BURNS, P. A., F. L. ALLEN and B. W. GLICKMAN, 1986 DNA sequence analysis of mutagenicity and site specificity of ethyl methanesulfonate in Uvr$^+$ and UvrB$^-$ strains of *Escherichia coli*. Genetics **113:** 811–819.

CARIELLO, N. F., P. KEOHAVONG, A. G. KAT and W. G. THILLY, 1990 Molecular analysis of complex human cell populations: mutational spectra of MNNG and ICR-191. Mutat. Res. **231:** 165–176.

CRESSIE, N. A. C., and T. R. C. READ, 1984 Multinomial goodness-of-fit tests. J. R. Statist. Soc. Ser. B **46:** 440–464.

CRESSIE, N. A. C., and T. R. C. READ, 1989 Pearson's X^2 and the loglikelihood ratio statistic G^2: a comparative review. Int. Statist. Rev. **57:** 19–43.

DEMARINI, D. M., A. ABUSHAKRA, R. GUPTA, L. J. HENDEE and J. G. LEVINE, 1992 Molecular analysis of mutations induced by the intercalating agent ellipticine at the *hisD3052* allele of *Salmonella typhimurium* TA98. Environ. Mol. Mutagen. **20:** 12–18.

DEVROYE, L., 1986 *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

DIACONIS, P., and B. EFRON, 1985 Testing for independence in a two-way table: new interpretations of the chi-square statistic. Ann. Statist. **13:** 845–874.

FISHER, R. A., 1935 The logic of inductive inference. J. R. Statist. Soc. Ser. A **98:** 39–54.

FOSTER, P. L., E. EISENSTADT and J. CAIRNS, 1982 Random components in mutagenesis. Nature **299:** 365–367.

GABRIEL, K. R., 1966 Simultaneous test procedures for multiple comparisons on categorical data. J. Am. Statist. Assoc. **61:** 1081–1096.

GILULA, Z., 1986 Grouping and association in contingency tables: an exploratory canonical correlation approach. J. Am. Statist. Assoc. **81:** 780–788.

GILULA, Z., and A. M. KRIEGER, 1989 Collapsed two-way contingency tables and the chi-square reduction principle. J. R. Statist. Soc. Ser. B **51:** 425–433.

KRAMER, M., and J. SCHMIDHAMMER, 1992 The chi-squared statistic in ethology: use and misuse. Anim. Behav. **44:** 833–841.

LAMBERT, I. B., A. J. E. GORDON, B. W. GLICKMAN and D. R. MCCALLA, 1992 The influence of local DNA sequence and DNA repair background on the mutational specificity of 1-nitroso-8-nitropyrene in *Escherichia coli*: inferences for mutagenic mechanisms. Genetics **132:** 911–927.

LEWONTIN, R. C., and J. FELSENSTEIN, 1965 The robustness of homogeneity tests in $2 \times N$ tables. Biometrics **21:** 19–33.

MARGOLIN, B. H., and R. J. LIGHT, 1974 An analysis of variance for categorical data, II: Small sample comparisons with chi-square and other competitors. J. Am. Statist. Assoc. **69:** 755–764.

MEHTA, C. R., and N. R. PATEL, 1983 A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Am. Statist. Assoc. **78:** 427–434.

MEHTA, C. R., and N. R. PATEL, 1986 FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. ACM Trans. Math. Software **12:** 154–161.

MONTELONE, B. A., L. A. GILBERTSON, R. NASSAR, C. GIROUX and R. E. MALONE, 1992 Analysis of the spectrum of mutations induced by the rad3-102 mutator allele of yeast. Mutat. Res. **267:** 55–66.

Read, T. R. C., 1984 Small-sample comparisons for the power divergence goodness-of-fit statistics. J. Am. Statist. Assoc. **79:** 929–935.

Roff, D. A., and P. Bentzen, 1989 The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol. Biol. Evol. **6:** 539–545.

Roscoe, J. T., and J. A. Byars, 1971 An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. J. Am. Statist. Assoc. **66:** 755–759.

Rudas, T., 1986 A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie-Read statistics. J. Statist. Comput. Simul. **24:** 107–120.

SAS Institute Inc., 1985 *SAS® User's Guide: Statistics, Version* 5 *Edition*. SAS Institute Inc., Cary, N.C.

Tanechi, N., 1988 Test of homogeneity of multinomial populations by an analysis of disperson, pp. 433–445 in *Statistical Theory and Data Analysis II*, edited by K. Matusita. Elsevier, Amsterdam.

Tindall, K. R., J. Stein and F. Hutchinson, 1988 Changes in DNA base sequence induced by gamma-ray mutagenesis of lambda phage and prophage. Genetics **188:** 551–560.

Weir, B. S., 1990 *Genetic Data Analysis*. Sinauer Associates, Sunderland, Mass.

Yates, F., 1984 Tests of significance for 2 × 2 contingency tables (with discussion). J. R. Statist. Soc. Ser. A **147:** 426–463.

Zelterman, D., 1987 Goodness-of-fit tests for large sparse multinomial distributions. J. Am. Statist. Assoc. **82:** 624–629.

Communicating editor: B. S. Weir