The Annals of Statistics
1985, Vol. 13, No. 3, 845–874

INVITED PAPER

TESTING FOR INDEPENDENCE IN A TWO-WAY TABLE: NEW INTERPRETATIONS OF THE CHI-SQUARE STATISTIC

BY PERSI DIACONIS AND BRADLEY EFRON

Stanford University

The classical chi-square test for independence in a two-way contingency table often rejects the independence hypothesis at an extremely small significance level, particularly when the sample size is large. This paper proposes some alternative distributions to independence, to help interpret the χ^2 statistic in such situations. The uniform alternative, in which every possible contingency table of the given dimension and sample size receives equal probability, leads to the *volume test*, as originally suggested in a regression context by H. Hotelling. Exponential family theory is used to generate a class of intermediate alternatives between independence and uniformity, leading to a random effects model for contingency tables.

1. Introduction. The chi-square test for independence in a two-way contingency table is an important accomplishment of early twentieth-century statistics. Tables 1 and 2 show the test in action, in both cases rejecting the hypothesis of independence. The main disadvantage of the chi-square test is apparent: when the independence hypothesis is strongly rejected, the actual significance level obtained by χ^2 conveys almost no additional information. For example, $\chi^2_{12} = 568.57$ is much more significant than $\chi^2_9 = 138.29$, but it will turn out that Table 2 lies much nearer to independence than does Table 1.

The objection here is really a general complaint against pure tests of significance. Significance tests are easy to use because we need only consider the null hypothesis family of distributions, in this situation the independence distributions for two-way tables; *but*, if the test strongly rejects the null hypothesis, the statistician receives little guidance as to what distribution actually generated the data.

This paper proposes some alternative distributions to independence, to help interpret χ^2 in situations like those of Tables 1 and 2. For example, Section 2 considers the distribution of χ^2 under the uniform distribution, in which every possible contingency table of the given dimension and sample size receives equal probability. For dimension 4 × 4 with sample size $n = 592$ there are $\binom{607}{15} \doteq 3.59 \cdot 10^{29}$ such tables. We will show that about 10% of these tables have $\chi^2 \leq 138.29$. In other words, the value of χ^2 observed in Table 1 is not overwhelmingly unusual assuming the uniform distribution. Why we might be interested in the uniform distribution is discussed in Sections 2 and 5, but in an

Received October 1983; revised September 1984.

AMS 1980 *subject classifications*. 62F05, 62G10

Key words and phrases. Chi-square test for independence, overdispersion, volume tests, random effects for exponential families.

TABLE 1

Eye color versus hair color for $n = 592$ subjects, Snee (1974). The standard chi-square test for independence gives $\chi^2 = 138.29$ on 9 degrees of freedom, strongly rejecting the hypothesis of independence. There are $3.59 \cdot 10^{29}$ 4×4 tables of sample size 592. Among these about 10% have $\chi^2 \leq 138.29$. In this case we cannot reject the alternative hypothesis that the observed table was selected uniformly from the set of all possible tables.

Eye Color	Hair Color				Total
	Black	Brunette	Red	Blond	
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

TABLE 2

Number of children in family versus yearly income, for $n = 25263$ Swedish families, Cramér (1946). The chi-square test strongly rejects the hypothesis of independence, $\chi^2 = 568.57$ on 12 degrees of freedom. Among all possible 5×4 tables with $n = 25263$ only a very small proportion, about $2.1 \cdot 10^{-7}$, have $\chi^2 \leq 568.57$. In this case we also strongly reject the alternative hypothesis that the observed table was selected uniformly from the set of all possible tables. Sections 4 and 5 discuss models intermediate between independence and uniformity, allowing us to interpret χ^2 when neither hypothesis is tenable.

Number of Children	Yearly income, Units of 1000 Kroner				Total
	0–1	1–2	2–3	3+	
0	2161	3577	2184	1636	9558
1	2755	5081	2222	1052	11110
2	936	1753	640	306	3635
3	225	419	96	38	778
≥ 4	39	98	31	14	182
Total	6116	10928	5173	3046	25263

obvious sense it is an antagonistic alternative to the independence hypothesis, for which the marginal probabilities of the table determine all the interior probabilities. Hotelling (1939) used exactly the same idea to generate a class of tests for nonlinear regression problems.

The procedure just described will be called a *volume test*: in the simplex of 16 dimensions, each point of which corresponds to a 4×4 table of proportions or probabilities, the value 10% is essentially the ratio of volumes between the set of points having $\chi^2 \leq 138.29$ and the entire simplex. The set of perfectly independent tables is a six-dimensional curved surface inside the simplex. The usual chi-square test says that Table 1 lies too far away from this surface to have been generated by chance multinomial variation from a probability table lying on the surface. The volume test says that Table 1 does not lie particularly near the surface, under sampling from the uniform distribution.

The situation is different for Table 2. Here the observed table lies too far away from the surface of independence, in terms of multinomial variation, but also it lies too *near* the surface to have been chosen uniformly. If we could look at the simplex in 20 dimensions describing all 5 × 4 tables, we would see that Table 2 lies very near the seven-dimensional curved surface of independence. As a matter of fact, the set of tables lying nearer the surface than Table 2, in terms of distance as measured by the χ^2-statistic, is only about $2.1 \cdot 10^{-7}$ the volume of the entire simplex.

The decisive rejection of both independence and uniformity for Table 2 leaves us with little information still about what distribution actually generated the data. Sections 4 and 5 discuss a class of intermediate models. Roughly speaking, the class is a one-parameter exponential family passing through the independence and uniform distributions, and having χ^2 as its sufficient statistic.

The natural parameter of this family can be interpreted as "effective sample size," say ν. We imagine that Table 2 has observed proportions as indicated, for example, 2163/25263 in the upper left category, but that the sample size has been reduced from $n = 25263$ to some smaller number ν. Smaller sample size allows the observed table to lie further from the surface of independence under the hypothesis of independence. We will see that for Table 2 a 90% confidence interval for ν, consonant with the independence hypothesis, is

$$\nu \in [232, 935]. \tag{1.1}$$

Sections 4 and 5 show that these considerations relate to a *random effects* model for contingency tables.

All of the significance levels and confidence intervals suggested in this paper are functions of χ^2 (or of its close cousin, the Kullback-Leibler distance), mostly very simple functions which can be calculated on a hand calculator. The goal is to extend the usefulness of χ^2, not to dissect the table using more elaborate structural models. Needless to say, this is not an argument against structural models, which often can give deeper insights into the data; χ^2 is an effective device for preliminary data analysis, particularly when the statistician has many two-way tables under review. This paper tries to refine its powers of explanation. The literature contains other such refinements, for example, the *mean square* contingency, χ^2/n, see Cramér (1946).

None of the statistical ideas presented here are new. Hotelling's seminal paper of 1939 has already been mentioned. The basic defect of pure tests of significance, that the results may depend more on sample size than on the true state of nature, was forcefully pointed out by Berkson in 1938. Bayesian solutions have been proved by Jeffries, Savage, Lindley, Hald and many others (see Shafer, 1982). What we call the "volume test," following Hotelling's original terminology, was considered in the context of two-way tables by Good (1976), and at some points, which will be indicated as they occur, we will be closely following Good's line of thought. The interesting series of articles by Good (1976, 1983), and Good and Crook (1980) are particularly relevant to our Sections 2, 3, and 7.

Components of variance approaches to categorial data, as used in Sections 4 and 5, date back to Lexis in the nineteenth century. These are nicely described in Chapter 3 of Heyde and Seneta (1977). Recent Bayesian work on the analysis of contingency tables gives a class of random effects models. This work is surveyed in Chapter 12 of Bishop, Fienberg and Holland (1975). More recent work is in papers by Dickey (1983), Laird (1978) or Leonard (1977). Section 9 gives a more thorough discussion of related literature.

2. Volume tests for independence. This section motivates and describes volume tests for independence in a two-way contingency table. A simple formula is given which approximates the significance level of the volume test, for example, 10% in Table 1, as a function of the usual χ^2-statistic. A more careful discussion, including proofs and details, appears in Sections 6, 7, and 8.

Let I be the number of rows and J be the number of columns in the contingency table. The table of observed proportions $\mathbf{p}$ has ijth entry

$$p_{ij} = m_{ij}/n, \quad i = 1, 2, \cdots, I, \quad j = 1, 2, \cdots, J, \tag{2.1}$$

where m_{ij} is the number of observed counts in row i, column j, and n is the total sample size. For example, $I = 5$, $J = 4$ in Table 2, and $p_{3,2} = 419/25263$.

The set of all possible $I \times J$ probability tables π is the simplex in IJ dimensions,

$$\mathscr{S}_{IJ} = \{\pi\colon \pi_{ij} \geq 0, \textstyle\sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1\}. \tag{2.2}$$

The observed table of proportions $\mathbf{p}$, obtained by multinomial sampling from some true table of probabilities π, must lie in the lattice subset of $\mathscr{S}_{IJ}$ having coordinates which are nonnegative integer multiples of $1/n$,

$$\mathscr{S}_{IJ}^{(n)} = \{\mathbf{p} = \mathbf{m}/n\colon m_{ij} \text{ integer} \geq 0, \textstyle\sum_{i=1}^{I} \sum_{j=1}^{J} m_{ij}/n = 1\}. \tag{2.3}$$

There are

$$N_{IJ}^{(n)} = \binom{n + IJ - 1}{IJ - 1} \tag{2.4}$$

distinct lattice points in $\mathscr{S}_{IJ}^{(n)}$.

Figure 1 gives a schematic representation of the volume test: (1) The *independence surface* $\mathscr{I}_{I,J}$, which is the set of probability tables π having $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j (the plus indicating summation over the missing subscript), is a manifold of dimension $I + J - 2$ curving through the $(IJ - 1)$-dimensional flat space $\mathscr{S}_{IJ}$. The hypothesis of independence, called H_1 in this paper, is the hypothesis $\pi \in \mathscr{I}_{I,J}$. (2) The χ^2-statistic is n times the Mahalanobis squared distance between the observed table $\mathbf{p}$ and the point $\hat{\pi}$ on $\mathscr{I}_{I,J}$ nearest $\mathbf{p}$, "nearest" meaning maximizing the likelihood. The inner product for the Mahalanobis distance is determined by the covariance matrix of the multinomial distribution having $\pi = \hat{\pi}$. (3) The set of tables having chi-square statistic equal or less than the observed value of χ^2 is an elliptical tube $\mathscr{E}(\chi^2)$ surrounding $\mathscr{I}_{I,J}$. (4) The achieved significance level $\varepsilon(\chi^2)$ for the volume test is the ratio of the number of lattice points inside $\mathscr{E}(\chi^2)$ to the total number of lattice points $N_{IJ}^{(n)}$. Roughly speaking, $\varepsilon(\chi^2)$ is the ratio of volumes of $\mathscr{E}(\chi^2)$ to $\mathscr{S}_{IJ}$. A careful description of the geometry will be given later. Section 2.7 of Bishop, Fienberg and Holland (1975) provides detailed pictures of $\mathscr{S}_{IJ}$ and $\mathscr{I}_{I,J}$ for the case $I = J = 2$.

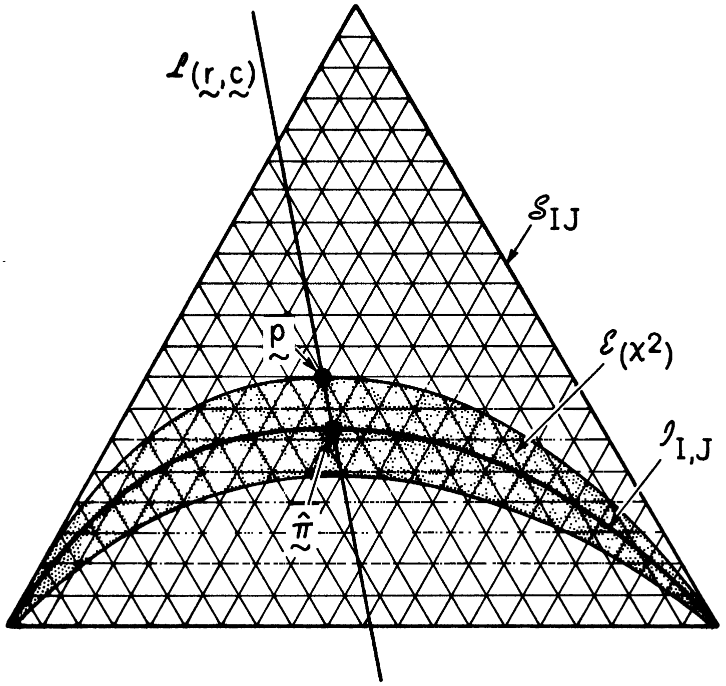
$\mathcal{L}(\underset{\sim}{r},\underset{\sim}{c})$
$\mathcal{S}_{IJ}$
$\mathcal{E}(\chi^2)$
$\mathcal{J}_{I,J}$
$\underset{\sim}{p}$
$\underset{\sim}{\hat{\pi}}$

FIG. 1. *Schematic representation of the volume test. The independence surface $\mathcal{J}_{I,J}$ is a curved manifold in the simplex $\mathcal{S}_{IJ}$. The shaded region $\mathcal{E}(\chi^2)$ represents those tables having chi-square statistic for independence equal or smaller than the observed value of χ^2. The achieved significance level for the volume test is the number of lattice points inside $\mathcal{E}(\chi^2)$ divided by the total number of lattice points. An obvious approximation for this is the ratio of volumes of $\mathcal{E}(\chi^2)$ to $\mathcal{S}_{IJ}$.*

As a first approximation for $\varepsilon(\chi^2)$, we develop the following expression:

$$\varepsilon(\chi^2) \doteq (\pi\chi^2/n)^{D/2}(c_{I,J}/\textstyle\prod_{h=1}^{IJ-1}(1+h/n)) \tag{2.5}$$

where

$$D = (I-1)(J-1) \tag{2.6}$$

and

$$c_{I,J} = \frac{\Gamma(IJ)\Gamma((J+1)/2)^I\Gamma((I+1)/2)^J}{(D/2)!\Gamma(I(J+1)/2)\Gamma(J(I+1)/2)}. \tag{2.7}$$

(The notation $x!$ stands for $\Gamma(x+1)$ even if x is not an integer.) A small tabulation of $c_{I,J}$ appears in Table 3. Stirling's formula can be used to approximate the product in the denominator of (2.5), leading to the slightly handier expression

$$\varepsilon(\chi^2) \doteq (\pi\chi^2/n)^{D/2}\, c_{I,J}\exp\{H-(n+H+{\textstyle\frac{1}{2}})\log(1+H/n)\},\quad H = IJ-1. \tag{2.5$'$}$$

TABLE 3
Values of the constant $c_{I,J}$ ($= c_{J,I}$) appearing in (2.5).

I	J						
	2	3	4	5	6	7	8
2	1.044	1.197	1.435	1.767	2.214	2.809	3.596
3		1.400	1.657	1.986	2.405	2.936	3.609
4			1.850	2.054	2.284	2.548	2.851
5				2.073	2.075	2.073	2.070
6					1.853	1.641	1.448
7						1.282	0.994
8							0.625

Formula (2.5) is based on the approximation

$$\text{Number of points in } \mathscr{E}(\chi^2) \doteq \{\text{Volume of } \mathscr{E}(\chi^2)\}\{\text{Density of lattice points in } \mathscr{S}_{IJ}^{(n)}\}. \tag{2.8}$$

Since the lattice is perfectly regular, density is unambiguously defined as the ratio of number of points in a cube to the volume of the cube as volume goes to infinity. It is shown in Section 6 that the density is

$$n^{IJ-1}/\sqrt{IJ}. \tag{2.9}$$

The following theorem is proved in Section 6:

THEOREM 1. *The* $(IJ - 1)$*-dimensional volume of* $\mathscr{E}(\chi^2)$ *equals*

$$(\pi\chi^2/n)^{D/2}c_{I,J}[\sqrt{IJ}/\Gamma(IJ)]. \tag{2.10}$$

Formula (2.5) is just (2.10) × (2.9)/(2.4). Because of edge effects, described in Section 7, (2.5) differs somewhat from the obvious approximation [volume $\mathscr{E}(\chi^2)$]/[volume $\mathscr{S}_{IJ}$] for $\varepsilon(\chi^2)$.

It turns out, for reasons that will become apparent in Section 3, that approximation (2.5) is a rough but useful upper bound for $\varepsilon(\chi^2)$. Applied to Table 1, for example, it gives $\varepsilon(\chi^2) \doteq .37$ compared to the actual value .10. The corresponding numbers for Table 2 are $2.6 \cdot 10^{-7}$ compared to $2.1 \cdot 10^{-7}$. Tables 4 and 5 give some additional numerical comparisons.

Why are we interested in $\varepsilon(\chi^2)$, the significance level of the volume test? As a simple alternative to the hypothesis of independence, consider H_0, the Bayesian hypothesis that the table π of true probabilities is chosen according to a uniform or flat Dirichlet distribution on IJ categories,

$$H_0\colon \pi \sim D_{IJ}(\mathbf{1}_{IJ}), \tag{2.11}$$

$\mathbf{1}_{IJ}$ being the IJ-dimensional vector of 1's. That is, π is chosen uniformly in the plane set $\pi_{ij} \geq 0$, $\sum\sum \pi_{ij} \leq 1$. Assume that given π, then $\mathbf{p}$ is obtained by multinomial sampling,

$$\mathbf{p} \mid \pi \sim \text{Mult}_{IJ}(n, \pi)/n, \tag{2.12}$$

where the notation indicates n draws from an IJ-category multinomial distribution with true probability vector π, rescaled by factor $1/n$.

As Good points out in Section 6 of [1976], (2.11) and (2.12) imply that marginally $\mathbf{p}$ has a uniform distribution over $\mathscr{S}_{IJ}^{(n)}$, with probability mass function say

$$H_0\colon g_0(\mathbf{p}) = 1/N_{IJ}^{(n)}, \quad \mathbf{p} \in \mathscr{S}_{IJ}^{(n)}. \tag{2.13}$$

If we now want to test the simple hypothesis H_0, (2.11) or (2.13), versus the composite hypothesis H_1 of independence, using χ^2 as the test statistic, then $\varepsilon(\chi^2)$ is the achieved significance level of the test. The choice of χ^2 as a test statistic is not totally arbitrary: in the statistically simpler context of Section 3, where we condition on the row and column margins of the observed table, the problem becomes one of testing simple versus simple hypotheses, and χ^2 is equivalent to the likelihood ratio statistic, at least to a first order of approximation.

The flat prior (2.11) is a useful alternative to H_1, the hypothesis of independence. Good and Crook (1980) consider several other alternatives. Under H_1, the marginal probabilities of the true table π completely determine the interior of π, by multiplication. Under H_0, the margins of π say very little about the interior: given the margins, the interior of π is uniformly distributed over all possible tables consistent with those margins (see Formula (7.3)). A different argument for the uniform appears in Section 5. It is seen to be the approximate end point of a one-parameter exponential family starting at H_1, and having χ^2 as sufficient statistic.

How far separated are the hypotheses H_0 and H_1? Is it easy or difficult to distinguish between them by means of a significance test based on χ^2? This depends on n, I, and J in a way which can be understood using (2.5). *For two-by-two tables, $I = J = 2$, Table 4 shows that it is difficult to distinguish H_1 from H_0*

TABLE 4

$I = J = 2$: probability that the usual chi-square test accepts the hypothesis H_1 of independence, even though the table is chosen according to the uniform distribution H_0. For example, the .05 significance level of the chi-square test is 3.841, and with sample size 640, formula (2.5′) gives $\varepsilon(3.841) \doteq .14$ as the probability of accepting H_1 if H_0 is true. For $n \le 160$ it is quite difficult to distinguish H_1 from H_0. A Monte Carlo check showed that these values of $\varepsilon(\chi^2)$ are accurate to within .01.

	Level of Usual Chi-Square Test					
	.5	**.25**	**.1**	**.05**	**.025**	**.01**
$\boldsymbol{n = 40}$	.17	.29	.42	.50	.57	.65
$\boldsymbol{n = 80}$	.13	.22	.32	.38	.43	.49
$\boldsymbol{n = 160}$	.10	.16	.23	.28	.32	.36
$\boldsymbol{n = 320}$	.07	.12	.17	.20	.23	.26
$\boldsymbol{n = 640}$	.05	.08	.12	.14	.16	.19

TABLE 5

$I = J = 4$: probability that the usual chi-square test accepts the hypothesis H_1 of independence, even though the table is chosen according to the uniform distribution H_0. The hypotheses H_1 and H_0 are far separated for $I = J = 4$, even for sample sizes as small as $n = 80$. Formula (2.5′) gives a good qualitative description of the situation but, as mentioned in the text, the actual values of $\varepsilon(\chi^2)$ (parentheses denote values obtained by Monte Carlo) can be substantially smaller. This effect is discussed in Sections 3 and 8.

	Level of Usual Chi-Square Test					
	.5	**.25**	**.1**	**.05**	**.025**	**.01**
$n = 60$	.007	.029	.089	.169	.287	.514
		(.029)	(.063)	(.115)	(.153)	(.190)
$n = 80$	.003	.012	.040	.072	.121	.218
	(.003)	(.010)	(.032)	(.058)	(.070)	(.109)
$n = 160$	.000	.001	.003	.006	.011	.019
			(.003)	(.005)	(.007)	(.014)
$n = 320$	.000	.000	.000	.000	.001	.001
$n = 640$	.000	.000	.000	.000	.000	.000

even for sample sizes as large as $n = 640$. For example, the usual chi-square test, level $\alpha = .05$, accepts H_1 for $\chi^2 \le 3.841$; and if H_0 is true this happens with substantial probability, $\varepsilon(3.841) \doteq .14$ even for $n = 640$.

The situation is different for $I = J = 4$. Table 5 shows that in this case H_1 and H_0 are far separated, even for sample sizes as small as $n = 80$. Sections 4 and 5 describe a family of alternatives which interpolate between H_1 and H_0.

3. Conditional volume tests. The problem of testing for independence in a two-way table becomes easier, from an inferential point of view, if we condition our inferences on the observed margins of the table. This section develops the conditional volume test. Proofs and details are deferred to Sections 6, 7, and 8. We begin with a more careful description of the chi-square test.

The row and column marginal proportions for the observed table $\mathbf{p}$ will be denoted

$$r_i = p_{i+} = \textstyle\sum_{j=1}^{J} p_{ij}, \quad c_j = p_{+j} = \textstyle\sum_{i=1}^{I} p_{ij}, \tag{3.1}$$

and likewise

$$\rho_i = \pi_{i+},, \quad \gamma_j = \pi_{+j} \tag{3.2}$$

for the marginal probabilities of the true table π; we also write $\mathbf{r} = (r_1, r_2, \cdots, r_I)'$, $\mathbf{c} = (c_1, c_2, \cdots, c_J)'$, $\rho = (\rho_1, \rho_2, \cdots, \rho_I)'$, and $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_J)'$.

Having observed $\mathbf{p} \sim \mathrm{Mult}_{IJ}(n, \pi)/n$, the maximum likelihood estimate of π assuming the hypothesis of independence H_1: $\pi_{ij} = \rho_i\gamma_j$ is

$$\hat{\pi}_{ij} = r_i c_j, \quad i = 1, 2, \cdots, I, \quad j = 1, 2, \cdots, J. \tag{3.3}$$

The chi-square test of H_1, significance level α, rejects for values of

$$\chi^2 = nS, \quad (S = \textstyle\sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - \hat{\pi}_{ij})^2/\hat{\pi}_{ij}) \tag{3.4}$$

larger than the upper $100 \cdot \alpha$ percentile point of a standard χ_D^2-distribution,

$D = (I - 1)(J - 1)$. Zero values of $\hat{\pi}_{ij}$ correspond to zero values of p_{ij}, and contribute nothing to S.

It is sometimes notationally convenient to think of $\mathbf{p}$ as a vector in IJ-dimensional Euclidian space $\mathscr{R}^{IJ}$ with its elements ordered lexicographically, $\mathbf{p} = (p_{11}, p_{12}, \cdots, p_{1J}, p_{21}, \cdots, p_{2J}, \cdots, p_{I1}, \cdots, p_{IJ})'$, and likewise for π and $\hat{\pi}$. Let $\hat{\Sigma}^{-1}$ be the $IJ \times IJ$ diagonal matrix with ijth diagonal element $1/\hat{\pi}_{ij}$. Then, we see that

$$S = (\mathbf{p} - \hat{\pi})' \, \hat{\Sigma}^{-1}(p - \hat{\pi}), \tag{3.5}$$

the squared Mahalanobis distance between $\mathbf{p}$ and $\hat{\pi}$, with inner product matrix $\hat{\Sigma}^{-1}$. The advantage of considering S, rather than the equivalent statistic χ^2, is that S has a clear geometric interpretation, not depending upon the sample size n.

The table $\hat{\pi}$ has the same margins $\mathbf{r}$ and $\mathbf{c}$ as the observed table $\mathbf{p}$. This means that $\mathbf{p} - \hat{\pi}$ has all margins zero. In other words $\mathbf{p} - \hat{\pi}$, thought of as an IJ-dimensional vector, lies in a certain D-dimensional linear subspace of $\mathscr{R}^{IJ}$, say $\mathscr{L}$, described explicitly in Section 6. The orientation of $\mathscr{L}$ in $\mathscr{R}^{IJ}$ is fixed, and does not depend on $\mathbf{p}$ or $\hat{\pi}$ in any way.

Define

$$\mathscr{V}(\mathbf{r}, \mathbf{c}) = \left\{\pi \colon \pi_{i+} = r_i,\ \pi_{+j} = c_j,\ \pi_{ij} \geq 0,\ \begin{matrix} i = 1, 2, \cdots, I \\ j = 1, 2, \cdots, J \end{matrix}\right\}, \tag{3.6}$$

the set of probability tables having the same margins as the observed table. Also let $\mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c})$ be the lattice of vectors in $\mathscr{V}(\mathbf{r}, \mathbf{c})$ having coordinates which are nonnegative integer multiples of $1/n$. If $\mathscr{L}(\mathbf{r}, \mathbf{c})$ is the parallel translate of the D-dimensional space $\mathscr{L}$ going through the point $\hat{\pi}$, then, using definitions (2.2) and (2.3), $\mathscr{V}(\mathbf{r}, \mathbf{c})$ equals $\mathscr{S}_{IJ} \cap \mathscr{L}(\mathbf{r}, \mathbf{c})$, and $\mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c}) = \mathscr{S}_{IJ}^{(n)} \cap \mathscr{L}(\mathbf{r}, \mathbf{c})$.

The process of slicing $\mathscr{S}_{IJ}$ with $\mathscr{L}(\mathbf{r}, \mathbf{c})$ is indicated in Figure 1. Figure 2 shows the slice $\mathscr{V}(\mathbf{r}, \mathbf{c})$ and also $\mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c})$. (In fact, Figure 2 depicts the case $I = 3$, $J = 2$, $n = 60$, $\mathbf{r} = (.2, .3, .5)'$, and $\mathbf{c} = (.4, .6)'$, with observed table $p = (.1, .1, .2, .1, .1, .4)'$.) The slice intersects just one point of the independence surface $\mathscr{I}_{I,J}$, namely $\hat{\pi}$.

The big advantage of conditioning on $\mathbf{r}$ and $\mathbf{c}$ is that the hypothesis of independence becomes simple: under H_1, points $\mathbf{p}$ in $\mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c})$ have what Good calls the Fisher-Yates distribution,

$$H_1\colon g_1(\mathbf{p} \mid \mathbf{r}, \mathbf{c}) = \binom{n}{n\mathbf{p}} \Big/ \binom{n}{n\mathbf{r}}\binom{n}{n\mathbf{c}}, \quad \mathbf{p} \in \mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c}). \tag{3.7}$$

Here

$$\binom{n}{n\mathbf{p}} = \frac{n!}{\prod_{i=1}^{I} \prod_{j=1}^{J} (np_{ij})!}, \quad \binom{n}{n\mathbf{r}} = \frac{n!}{\prod_{i=1}^{I} (nr_i)!},$$

etc. In the case $I = J = 2$, (3.7) is the hypergeometric distribution.

According to (3.4), (3.5), the set of points π in $\mathscr{L}(\mathbf{r}, \mathbf{c})$ having Mahalanobis squared distance from $\hat{\pi}$ no greater than the observed value S is a D-dimensional

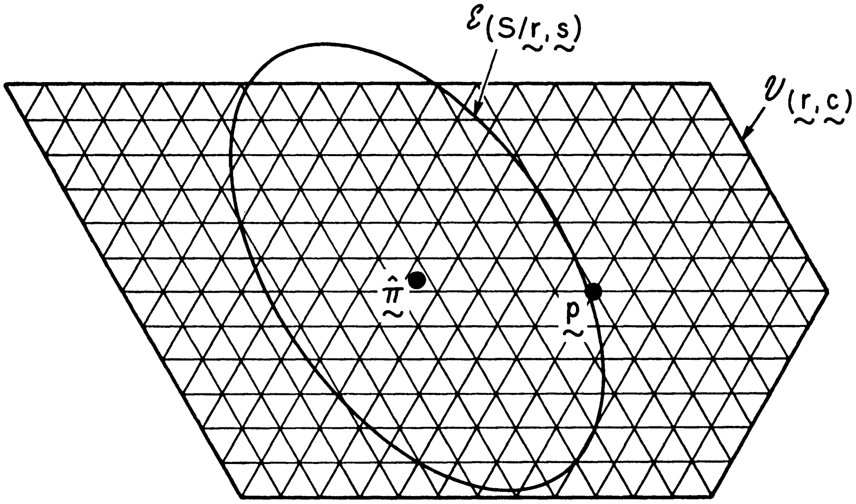
$\mathscr{E}(S/\underset{\sim}{r},\underset{\sim}{s})$
$\mathscr{V}(\underset{\sim}{r},\underset{\sim}{c})$
$\underset{\sim}{\hat{\pi}}$
$\underset{\sim}{p}$

FIG. 2. *The "slice" $\mathscr{V}(\mathbf{r}, \mathbf{c})$ is the set of all tables π having the same margins $\mathbf{r}$ and $\mathbf{c}$ as the observed table. In terms of Figure 1 it is obtained by slicing $\mathscr{S}_{IJ}$ through the point $\hat{\pi}$ with a D-dimensional hyperplane. The ellipsoid $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ consists of those vectors in the same hyperplane having Mahalanobis squared distance from $\hat{\pi}$ no greater than the observed value S. The achieved significance level for the conditional volume test is the number of lattice points inside $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ divided by the number of lattice points in $\mathscr{V}(\mathbf{r}, \mathbf{c})$. An obvious aproximation for this is the ratio of volumes of $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ to $\mathscr{V}(\mathbf{r}, \mathbf{c})$.*

ellipsoid,

$$(3.8) \qquad \mathscr{E}(S \mid \mathbf{r}, \mathbf{c}) = \{\pi\colon (\pi - \hat{\pi})'\hat{\Sigma}^{-1}(\pi - \hat{\pi}) \le S,\ \pi \in \mathscr{S}(\mathbf{r}, \mathbf{c})\}.$$

The hypothesis of uniformity H_0, (2.13), is also uniform when conditioned on $(\mathbf{r}, \mathbf{c})$,

$$(3.9) \qquad \begin{gathered} H_0\colon g_0(\mathbf{p} \mid \mathbf{r}, \mathbf{c}) = 1/N^{(n)}(\mathbf{r}, \mathbf{c}), \\ (N^{(n)}(\mathbf{r}, \mathbf{c}) = \text{number of lattice points in } \mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c}).) \end{gathered}$$

By definition, the achieved significance level of the conditional volume test is the ratio

$$(3.10) \qquad \varepsilon(S \mid \mathbf{r}, \mathbf{c}) = \frac{\text{number of points of } \mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c}) \text{ inside } \mathscr{E}(S \mid \mathbf{r}, \mathbf{c})}{N^{(n)}(\mathbf{r}, \mathbf{c})}$$

This is just the significance level achieved by $\mathbf{p}$ assuming distribution (3.9), where significance is defined by smallness of the statistic S. The choice of S to measure significance is not arbitrary. To a first approximation, the likelihood ratio statistic $g_0(\mathbf{p} \mid \mathbf{r}, \mathbf{c})/g_1(\mathbf{p} \mid \mathbf{r}, \mathbf{c})$ is a monotonic function of S, see Section 5.

Notice that $\mathbf{r}$, $\mathbf{c}$, and S completely determine $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ and $\mathscr{V}(\mathbf{r}, \mathbf{c})$, with no dependence on the sample size n. Essentially $\varepsilon(S \mid \mathbf{r}, \mathbf{c})$ *does not depend on n*, except for minor effects relating to the granularity of the lattice points. (Section 8 briefly discusses the granularity question.) Large values of n do not necessarily produce extremely small significance levels for the volume test, as is usually the case with the standard chi-square test.

The numerator of (3.10) can be approximated by the volume-times-density argument (2.8). The following two theorems are proved in Section 6:

THEOREM 2. *The D-dimensional volume of $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ equals*

$$(\pi S)^{D/2} c_{I,J}(\mathbf{r}, \mathbf{c}) \tag{3.11}$$

where

$$c_{I,J}(\mathbf{r}, \mathbf{c}) = (1/(D/2)!)(I \textstyle\prod_{i=1}^{I} r_i)^{(J-1)/2}(J \prod_{j=1}^{J} c_j)^{(I-1)/2}. \tag{3.12}$$

(Notice that $\pi S = \pi\chi^2/n$, so (3.11) is closely related to (2.10).)

THEOREM 3. *The density of lattice points in $\mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c})$ is*

$$n^D / I^{(J-1)/2} J^{(I-1)/2} \tag{3.13}$$

points per unit of D-dimensional volume.

REMARK. In Diaconis and Efron (1983), Theorem 3 is used to calculate the generalized variance of the Fisher-Yates distribution (3.7).

Computing the denominator of (3.10) is a well-known unsolved combinatorial problem. Good (1976), (1983) and Good and Crook (1977) give a nice review of the available results. We will approximate $N^{(n)}(\mathbf{r}, \mathbf{c})$ by the volume-times-density argument, using (3.13) and the following approximation for the volume of $\mathscr{V}(\mathbf{r}, \mathbf{c})$,

$$\hat{V} = \left(1 + \frac{IJ}{2n}\right)^D \frac{I^{(J-1)/2} J^{(I-1)/2} \Gamma(J k_{\bar{r}})}{\Gamma(J)^I \Gamma(k_{\bar{r}})^J} \left(\textstyle\prod_{i=1}^{I} \bar{r}_i\right)^{J-1} \left(\prod_{j=1}^{J} \bar{c}_j\right)^{k_{\bar{r}}-1}, \tag{3.14}$$

where

$$\bar{\mathbf{r}} = (1 - w)\frac{\mathbf{1}_I}{I} + w\mathbf{r}, \quad \bar{\mathbf{c}} = (1 - w)\frac{\mathbf{1}_J}{J} + w\mathbf{c} \quad \left(w = \frac{1}{1 + IJ/2n}\right) \tag{3.15}$$

and

$$k_{\bar{r}} = (J + 1)/(J \|\bar{\mathbf{r}}\|^2) - 1/J. \tag{3.16}$$

The notation $\mathbf{1}_I$ indicates the vector of ones in I dimensions. (Formula (3.14) includes a correction for edge effects which deliberately overestimates the volume of $\mathscr{V}(\mathbf{r}, \mathbf{c})$, for reasons discussed in Section 7.)

Section 7 discusses approximation (3.14) and also an exact formula for the volume. The approximation is quite satisfactory for the cases at hand. Table 2, for instance, has $\hat{V} \doteq 5.9 \cdot 10^{-17}$ while the actual volume, obtained by laborious Monte Carlo calculation, is $5.7\ (\pm .2) \cdot 10^{-17}$.

The volume-times-density approximation here, using 5.7 ± 2.10^{-17} for volume, and (3.13) with $I = 5$, $J = 4$, $D = 12$, $n = 25263$ for density gives $2.14\ (\pm .1) \cdot 10^{34}$ for the number of arrays with the same margins as Table 2. Formula (B2.24) of Good (1976) gives $2.63 \cdot 10^{34}$ for this number. Good (1976, 6.6) also offers an improved, though somewhat ad hoc, approximation $1.91 \cdot 10^{34}$. In applying (3.14), the choice of which factor is called "row" rather than "column" can be interchanged, giving a different numerical approximation. If one of the margins includes very small proportions, for example "Number of children" in Table 2, then this factor should be chosen to be the rows.

Assuming that $\hat{V}$ is approximating the true volume correctly, (3.14) × (3.13) gives an excellent approximation to $N^{(n)}(\mathbf{r}, \mathbf{c})$. For instance with $I = J = 3$, $n = 30$, and $\mathbf{r} = \mathbf{c} = (1/3, 1/3, 1/3)'$, the approximation is $N^{(n)}(\mathbf{r}, \mathbf{c}) \doteq 2080$ using (3.14) × (3.13), and 2186 using true volume × (3.13), compared to the actual value 2211.

As a first approximation to the significance level of the conditional volume test, we have

$$\varepsilon(S \mid \mathbf{r}, \mathbf{c}) \doteq \frac{(3.11) \times (3.13)}{(3.14) \times (3.13)} = (\pi S)^{D/2} \frac{c_{I,J}(\mathbf{r}, \mathbf{c})}{\hat{V}}. \tag{3.17}$$

The density (3.13) has cancelled out, so, as in the unconditional case, the approximate significance level is just the ratio of volumes of the ellipsoid to the slice, as shown in Figure 2.

Formula (3.13) looks unnecessary, but it turns out to be a crucial fact in proving (3.11).

Formula (3.17), like (2.5), is usually an overestimate of $\varepsilon(S \mid \mathbf{r}, \mathbf{c})$. The reason is that (3.11) applies to the entire D-dimensional volume of $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$, including the part of $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ protruding outside of $\mathscr{V}(\mathbf{r}, \mathbf{c})$. For Table 1, (3.17) gave $\varepsilon(S \mid \mathbf{r}, \mathbf{c}) \doteq .41$ compared to an actual value of .09, computed by Monte Carlo. The corresponding numbers for Table 2 were $\varepsilon(S \mid \mathbf{r}, \mathbf{c}) \doteq 4.3 \cdot 10^{-5}$ compared to an actual value of $1.2 \cdot 10^{-5}$. Corrections for "protrusion" are discussed in Section 8.

Notice that for Table 2, (3.17) gives a quite different answer from (2.5), $\varepsilon(S \mid \mathbf{r}, \mathbf{c}) \doteq 4.3 \cdot 10^{-5}$ compared to $\varepsilon(\chi^2) \doteq 2.6 \cdot 10^{-7}$. The reason for this difference has to do with conditional versus unconditioned inference. The hypothesis of uniformity H_0 implies that the margins $\mathbf{r}$ and $\mathbf{c}$ will be roughly uniform, $\mathbf{r}$ equalling about $(1/I, 1/I, \cdots, 1/I)$ and $\mathbf{c}$ equalling about $(1/J, \cdots, 1/J)$. In Table 2 the $\mathbf{r}$ margin is markedly nonuniform. The unconditional volume test interprets this as evidence in favor of H_1. By conditioning on $(\mathbf{r}, \mathbf{c})$, the conditional volume test guarantees that the margins furnish no evidence for either H_0 or H_1. Good and Crook (1980) discuss the question of marginal information in 2 × 2 tables. They do not observe a large difference between conditional and unconditional inferences. This is because they use a variant of the uniform prior (2.11) which eliminates most of the supposed information in the margins.

Partial Conditioning. There is an intermediate position between using the unconditional volume test of Section 2 and the fully conditional test we have been discussing here. Conditioning on just one set of margins, say $\mathbf{r}$, instead of both $\mathbf{r}$ and $\mathbf{c}$, leads to a partially conditioned achieved significance level $\varepsilon(S \mid \mathbf{r})$ for the volume test. Without going into details, an approximate formula for

$\varepsilon(S \mid \mathbf{r})$, analogous to (3.17), is

$$(3.18) \quad \varepsilon(S \mid \mathbf{r}) \doteq (\pi S)^{D/2} \frac{\Gamma(J)^I \Gamma((I+1)/2)^J (\prod_{i=1}^I r_i)^{(J-1)/2}}{(D/2)!\,\Gamma(J((I+1)/2))(1 + IJ/2n)^{I(J-1)} (\prod_{i=1}^I \bar{r}_i)^{J-1}},$$

with $\bar{r}_i$ as defined in (3.15).

Applied to Table 2, formula (3.18) gives $\varepsilon(S \mid \mathbf{r}) = 4.4 \cdot 10^{-5}$ and $\varepsilon(S \mid \mathbf{c}) = 2.6 \cdot 10^{-7}$. The former nearly equals $\varepsilon(S \mid \mathbf{r}, \mathbf{c}) \doteq 4.3 \cdot 10^{-5}$, from (3.17), while the latter equals the unconditioned value obtained from (2.5). This difference has to do with the evidence, or lack of evidence, in the margins of the table, as discussed above.

4. Intermediate models. The independence model H_1 predicts small observed values for the statistic $S = \chi^2/n$. The uniform model H_0 was introduced in order to provide a reference distribution for larger observed values of S. This strategy worked reasonably well for Table 1, but for Table 2 the observed S was much too small to have come from H_0, as well as much too large to have come from H_1. This section introduces a class of intermediate models, which allows us to interpret intermediate valus of S. A more careful development, based on exponential family theory, appears in Section 5.

What does "interpret intermediate values of S" mean? We have in mind an analogy with the usual random effects model for a one-way layout. In the simplest form of the one-way layout, the statistician observes a normally distributed random vector

$$(4.1) \qquad \mathbf{x} \sim N_D(\boldsymbol{\beta}, \mathbf{I}/n)$$

with $\boldsymbol{\beta}$ unknown, and wishes to test the hypothesis

$$(4.2) \qquad H_1 \colon \boldsymbol{\beta} = \mathbf{0}.$$

The covariance matrix in (4.1) is expressed as $\mathbf{I}/n$ in accordance with the usual formulation where $\mathbf{x}$ is the average of n independent vectors $y_1, y_2, \cdots, y_n \sim N_D(\boldsymbol{\beta}, \mathbf{I})$.

The random effects model adds the partial Bayesian assumption

$$(4.3) \qquad \boldsymbol{\beta} \sim N_D(\mathbf{0}, \sigma_\beta^2 \mathbf{I})$$

to (4.1), partial referring to the unknown parameter σ_β^2, which is usually estimated non-Bayesianly. This results in the marginal distribution for $\mathbf{x}$

$$(4.4) \qquad \mathbf{x} \sim N_D(\mathbf{0}, (\sigma_\beta^2 + 1/n)\mathbf{I}).$$

The null hypothesis (4.2) is equivalent to

$$(4.5) \qquad H_1 \colon \sigma_\beta^2 = 0.$$

The advantage of considering (4.4), (4.5) rather than (4.1), (4.2) is that for the former the alternative to H_1 is a simple one-parameter family of distributions, while for the latter it is a D-parameter family. Analysis of (4.4), (4.5) is based on the sufficient statistic

$$S = \| \mathbf{x} \|^2 \sim (\sigma_\beta^2 + 1/n)\chi_D^2. \tag{4.6}$$

The observed value of S is used to test H_1 and, if H_1 is rejected, to form a confidence interval for σ_β^2,

The random effects model can be restated in a way which will be useful when we pursue the analogy for two-way tables. Define

$$\theta = 1/(1 + n\sigma_\beta^2) \tag{4.7}$$

so (4.6) becomes

$$S \sim \chi_D^2/n\theta. \tag{4.8}$$

Under H_1 we have $\sigma_\beta^2 = 0$, $\theta = 1$, and $S \sim \chi_D^2/n$. If $\sigma_\beta^2 > 0$ then $\theta < 1$, and $S \sim \chi_D^2/n\theta$ tends to be bigger than under H_1. If we define

$$\nu = n\theta, \tag{4.9}$$

called the *effective sample size*, then the random effects model can be thought of as the one-parameter family of distributions for S, $S \sim \chi_D^2/\nu$, $0 < \nu \leq n$. Definitions (4.7), (4.9) can be expressed as

$$\sigma_\beta^2 = 1/\nu - 1/n, \tag{4.10}$$

leading directly to $\mathbf{x} \sim N_D(\mathbf{0}, \mathbf{I}/\nu)$ in (4.4) and $S \sim \chi_D^2/\nu$ in (4.6).

Values of S which are too large under H_1 are easy to interpret within the random effects model. For example, with $D = 10$ and $n = 100$, observing $S = .60$ is much too large under H_1, but quite plausible for $\nu = 16.67$, since then $S \sim \chi_{10}^2/\nu = .60(\chi_{10}^2/10)$. (Notice that the MLE for ν is $\hat{\nu} = D/S$, so in this case $\hat{\nu} = 16.67$.) The central 90% interval for a χ_{10}^2 variate, [3.94, 18.31], gives central 90% confidence interval [6.57, 30.52], for ν, based on observing $S = \chi_{10}^2/\nu = .60$. A central rather than one-sided interval for ν is appropriate here because both upper and lower limits for ν are important, in (4.10) for example.

We return to the context of two-way tables, considered conditionally on the margins $(\mathbf{r}, \mathbf{c})$ as in Section 3. Let S have its original definition (3.4), then nS has a limiting χ_D^2 distribution under H_1, say

$$H_1\colon S \mid (\mathbf{r}, \mathbf{c}) \to \chi_D^2/n. \tag{4.11}$$

As a first step toward intermediate models for the two-way table situation, we assume that to every value of θ, $0 < \theta \leq 1$, there corresponds an hypothesis H_θ such that $n\theta S$ has a limiting χ_D^2 distribution under H_θ,

$$H_\theta\colon S \mid (\mathbf{r}, \mathbf{c}) \to \chi_D^2/n\theta, \quad 0 < \theta \leq 1. \tag{4.12}$$

These distributions are intermediate between H_1 and H_0.

If we ignore the fact that (4.12) is an approximation, we can formally apply the random effects model (4.8) to the analysis of two-way tables. Then $\nu = n\theta$ has MLE $\hat{\nu} = D/S$ and $1 - 2\alpha$ central confidence interval $[\chi_D^{2(\alpha)}/S, \chi_D^{2(1-\alpha)}/S]$, where $\chi_D^{2(\alpha)}$ is the $100 \cdot \alpha$ percentile point for a standard χ_D^2 distribution. The more careful approximations of Section 5 show that effective sample size is an apt name for $\nu = n\theta$ in the two-way table context.

Consider Table 1. The MLE for ν is $\hat{\nu} = 38.5$, with central 90% confidence interval

$$\nu \in [14.2, 72.4]. \tag{4.13}$$

The interpretation of these results can be pictured in terms of Figure 2: under the independence hypothesis H_1, the observed distance of $\mathbf{p}$ from $\hat{\pi}$ for Table 1 is most typical of sample sizes around 38; the observed distance would not be unreasonable for sample sizes in range (4.13).

We can get a more familiar interpretation of these results by referring back to the normal situation (4.1)–(4.10). For values of π near $\hat{\pi}$, the multinomial distribution $\mathbf{p} \sim \text{Mult}_{IJ}(n, \pi)/n$, conditioned on $(\mathbf{r}, \mathbf{c})$, has an approximate normal distribution

$$\mathbf{p} \mid \pi, (\mathbf{r}, \mathbf{c}) \sim N_D(\pi, \hat{\Sigma}/n), \tag{4.14}$$

where $(1/n)\hat{\Sigma}$ is the covariance matrix of the Fisher-Yates distribution (3.7). Here "N_D" indicates that the distribution is confined to the D-dimensional space $\mathscr{L}(\mathbf{r}, \mathbf{c})$ containing $\mathscr{V}(\mathbf{r}, \mathbf{c})$. A rough analogy of (4.3) for two-way tables is the partial Bayesian model

$$\pi \mid (\mathbf{r}, \mathbf{c}) \sim N_D(\hat{\pi}, \sigma_\beta^2 \hat{\Sigma}). \tag{4.15}$$

Notice that (4.14), (4.15) combine with (4.10) to give

$$\mathbf{p} \mid (\mathbf{r}, \mathbf{c}) \sim N_D(\hat{\pi}, \hat{\Sigma}/\nu). \tag{4.16}$$

This is just the normal approximation for the Fisher-Yates distribution (3.7), with sample size reduced from n to ν.

The confidence interval (4.13) for ν transforms into a confidence interval for σ_β^2 via relationship (4.10),

$$\sigma_\beta^2 \in [.0121, .0686], \tag{4.17}$$

and likewise for the MLE, $\hat{\sigma}_\beta^2 = .0243$. How big are these random effects? A measure comparing the amount of random effects variation in (4.14), $\sigma_\beta^2 \hat{\Sigma}$, relative to the size of the slice $\mathscr{V}(r, c)$, is

$$\sigma_{\text{rel}} = \sigma_\beta(|\hat{\Sigma}|^{1/2}/\hat{V})^{1/D}. \tag{4.18}$$

This is just the average standard deviation of (4.13) along a single dimension of the D-dimensional space $\mathscr{L}(\mathbf{r}, \mathbf{c})$, compared to the side of a D-cube having the same volume as $\mathscr{V}(\mathbf{r}, \mathbf{c})$.

It is shown in Section 6 that

$$|\hat{\Sigma}| = I^{J-1} J^{I-1} \left(\prod_{i=1}^{I} r_i\right)^{J-1} \left(\prod_{j=1}^{J} c_j\right)^{I-1}. \tag{4.19}$$

This and (3.14) give $(|\hat{\Sigma}|^{1/2}/\hat{V})^{1/D} = 1.64$ for Table 1. The previous results translate into $\hat{\sigma}_{\text{rel}} = .26$, with 90% confidence interval

$$\sigma_{\text{rel}} \in [.18, .43]. \tag{4.20}$$

We see that for Table 1, the random effects must be quite substantial.

The same analysis applied to Table 2 shows that although the random effects cannot be zero, or else the usual chi-square test would not have rejected independence, they are in fact very small. The MLE of the effective sample size is $\hat{\nu} = 533$, with central 90% confidence interval

$$\nu \in [232, 935]. \tag{4.21}$$

This gives $\hat{\sigma}_{\text{rel}} = .0051$, confidence interval

$$\sigma_{\text{rel}} \in [.003, .012]. \tag{4.22}$$

The maximum likelihood estimates and confidence intervals for ν, and for σ_{rel}, depend on the observed table $\mathbf{p}$ only through the statistic S. (Even the sample size n is not used.) The inferences they provide relate only to a corresponding feature of the true table π, its gross overall distance from the independence surface. Of course, a more incisive analysis of how π deviates from independence can be based on other features of $\mathbf{p}$. This is the point of log-linear modelling, correspondence analysis, and other structural models. The methods suggested in this paper are well-suited to quick preliminary analyses of two-way tables, but are not intended to replace a careful structural investigation.

5. Random effects for exponential families.

Most of the results of the last section, in particular the components of variance calculations beginning at (4.14), are familiar ideas in the theory of overdispersion of binomial proportions. There is an immense literature on overdispersion, going back a century to Lexis. A good review appears in Chapter 3 of Heyde and Seneta (1977).

This section is devoted to a more exact analogue of the normal-theory random effects model (4.1)–(4.10), applying to general multiparameter exponential families. Our goal is to justify approximation (4.12), which was used to interpret intermediate values of S for two-way tables. The idea of effective sample size turns out to play a basic role in this development.

Suppose then that $\mathbf{x}$ is the observed sufficient vector for a D-parameter exponential family $\mathscr{G}$, and that β is the expected value of $\mathbf{x}$. In our previous context, $\mathbf{x} = \mathbf{p}$ and $\beta = \pi$. The vector β indexes the family $\mathscr{G}$, and we can write the density function for a typical member of $\mathscr{G}$ as

$$g_\beta^{(n)}(\mathbf{x}) = \exp(n[\alpha'\mathbf{x} - \psi(\beta)]), \tag{5.1}$$

where $\alpha = \alpha(\beta)$ is the natural (or canonical) parameter vector and $\psi(\beta)$ is a normalizing constant. The constant n is the sample size, assuming as usual that $\mathbf{x}$ is actually the average of n original observed vectors $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n \sim_{\text{i.i.d.}} g_\beta^{(1)}$.

The sufficient vector $\mathbf{x} = \sum_{k=1}^n \mathbf{y}_k/n$ takes its values in some sample space $\mathscr{X}^{(n)}$. Its expectation vector β takes values in a space $\mathscr{B}$, both $\mathscr{B}$ and $\mathscr{X}^{(n)}$ being subsets of $\mathscr{R}^D$. In the context of Section 3, where we condition on $(\mathbf{r}, \mathbf{c})$, we have

$\mathscr{B} = \mathscr{V}(\mathbf{r}, \mathbf{c})$ and $\mathscr{X}^{(n)} = \mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c})$. We will assume $\mathscr{X}^{(n)} \subset \mathscr{B}$ to simplify the discussion.

Now suppose that there is some member of $\mathscr{B}$, say β_1, of particular interest. We wish to test the hypothesis

$$H_1\colon \beta = \beta_1, \tag{5.2}$$

versus the general alternative $\beta \in \mathscr{B}$. In Section 3, $\beta_1 = \hat{\pi}$. A well-known result of Hoeffding (1965) shows that the maximum likelihood ratio test statistic is

$$\sup_{\beta \in \mathscr{B}} g_{\beta}^{(n)}(\mathbf{x})/g_{\beta_1}^{(n)}(\mathbf{x}) = g_{\mathbf{x}}^{(n)}(\mathbf{x})/g_{\beta_1}^{(n)}(\mathbf{x}) = \exp(nT(\mathbf{x}, \beta_1)/2), \tag{5.3}$$

where T is twice the Kullback-Leibler information for one original observation,

$$T(\beta_2, \beta_1) = 2E_{\beta_2}\log g_{\beta_2}^{(1)}(\mathbf{y})/g_{\beta_1}^{(1)}(\mathbf{y}). \tag{5.4}$$

For the normal model (4.1), (4.2), $T(\mathbf{x}, \mathbf{0}) = \| \mathbf{x} \|^2$. Efron (1978) gives a review of exponential family theory, including Hoeffding's result and the technical details omitted from our brief discussion here.

By Wilks' theorem, $2 \log[g_{\mathbf{x}}^{(n)}(\mathbf{x})/g_{\beta_1}^{(n)}(\mathbf{x})]$ will be asymptotically distributed as χ_D^2 under H_1, so we have

$$H_1\colon nT(\mathbf{x}, \beta_1) \longrightarrow \chi_D^2. \tag{5.5}$$

We can test H_1 at approximate level α by rejecting for $nT(\mathbf{x}, \beta_1) > \chi_D^{2(1-\alpha)}$. Often, as we have seen, this test will reject at extremely small α levels, in which case it is helpful to have an interpretive theory like the normal random effects model. The crucial step in that theory is (4.4), which replaces the full D-parameter family (4.1) with a one-parameter exponential family of alternatives, having $T(\mathbf{x}, \beta_1)$ as sufficient statistic. This same program will now be carried out for general exponential families.

Let H_θ be the hypothesis that $\mathbf{x}$ is distributed according to the following density $f_\theta(\mathbf{x})$,

$$H_\theta\colon f_\theta(\mathbf{x}) = g_{\beta_1}^{(n)}(\mathbf{x})\exp[n(1 - \theta)T(\mathbf{x}, \beta_1)/2]\phi(\theta), \quad 0 < \theta \leq 1. \tag{5.6}$$

Here $\phi(\theta)$ is a normalizing constant. The one-parameter exponential family $\mathscr{F} = \{f_\theta, 0 < \theta \leq 1\}$ has sufficient statistic $T(\mathbf{x}, \beta_1)$. As shown at (5.13), (5.14) below, $\phi(\theta)$ can be approximated by

$$\phi(\theta) \doteq \theta^{D/2}. \tag{5.7}$$

Notice that for $\theta = 1$, $f_1(\mathbf{x}) = g_{\beta_1}^{(n)}(\mathbf{x})$, so H_1 can also be

$$H_1\colon \theta = 1. \tag{5.8}$$

There is another way to express $\mathscr{F}$. Suppose that $n\theta = \nu$ is an integer. By Hoeffding's result (5.3), $g_{\beta_1}^{(n)}(\mathbf{x})\exp[\mathrm{nT}(\mathbf{x}, \beta_1)/2] = g_{\mathbf{x}}^{(n)}(\mathbf{x})$, so

$$f_\theta(\mathbf{x}) = g_{\mathbf{x}}^{(n)}(\mathbf{x})\exp[-n\theta T(\mathbf{x}, \beta_1)/2]\phi(\theta). \tag{5.9}$$

Multiplying and dividing (5.9) by $g_{\mathbf{x}}^{(n\theta)}(\mathbf{x})/g_{\mathbf{x}}^{(n)}(\mathbf{x})$, and applying (5.3) again, gives

$$f_\theta(\mathbf{x}) = g_{\beta_1}^{(n\theta)}(\mathbf{x})(g_{\mathbf{x}}^{(n)}(\mathbf{x})/g_{\mathbf{x}}^{(n\theta)}(\mathbf{x}))\phi(\theta). \tag{5.10}$$

The structure of $\mathscr{F}$ can be understood in terms of (5.10), which we now state more carefully. The densities (5.1) produce the probability distributions of $\mathscr{G}$ by integration with respect to some common carrier measure on $\mathscr{X}^{(n)}$, say $G^{(n)}$, and $f_\theta(\mathbf{x})$ is also a density with respect to $G^{(n)}$. Let H_θ be the hypothesis that $\mathbf{x}$ is generated from f_θ. Then (5.10) says that the probability content of an infinitesimal region $d\mathbf{x}$ around $\mathbf{x}$, under H_θ, is

$$f_\theta(\mathbf{x})G^{(n)}(d\mathbf{x}) = \mathbf{g}_{\beta_1}^{(n\theta)}(\mathbf{x})G^{(n\theta)}(d\mathbf{x})\left(\frac{g_{\mathbf{x}}^{(n)}(\mathbf{x})G^{(n)}(d\mathbf{x})}{g_{\mathbf{x}}^{(n\theta)}(\mathbf{x})G^{(n\theta)}(d\mathbf{x})}\right)\phi(\theta). \tag{5.11}$$

For large values of n, the central limit theorem says that the distribution of $\mathbf{x}$ under $\mathbf{g}_\beta^{(n)}$ will approach a multivariate normal distribution, say

$$g_\beta^{(n)}\colon \sqrt{n}(\mathbf{x} - \beta) \to N_D(\mathbf{0}, \Sigma_\beta). \tag{5.12}$$

Applying the central limit theorem locally, as in Stone (1965), with $\beta = \mathbf{x}$, gives

$$\lim_{n\to\infty}(g_{\mathbf{x}}^{(n)}(\mathbf{x})G^{(n)}(d\mathbf{x})/g_{\mathbf{x}}^{(n\theta)}(\mathbf{x})G^{(n\theta)}(d\mathbf{x})) = 1/\theta^{D/2}. \tag{5.13}$$

(If $G^{(n)}$ is discrete, then $d\mathbf{x}$ is not allowed to become small too quickly in (5.13); taking $d\mathbf{x}$ a cube of side $O(n^{-2/3})$ is allowable.) The convergence in (5.13) tends to be rapid, because we are applying the CLT at the center of the limiting distribution, $\beta = \mathbf{x}$. Edgeworth calculations show that the bracketed term in (5.13) equals $\theta^{-D/2}[1 + 0_p(1/n)]$.

If the bracketed term in (5.11) is approximated by $\theta^{-D/2}$, then

$$1 = \int_{\mathscr{X}^{(n)}} f_\theta(\mathbf{x})G^{(n)}(d\mathbf{x}) \doteq \theta^{-D/2}\phi(\theta)\int_{\mathscr{X}^{(n\theta)}} \mathbf{g}_{\beta_1}^{(n)}(\mathbf{x})G^{(n\theta)}(d\mathbf{x}) = \theta^{-D/2}\phi(\theta), \tag{5.14}$$

and we see that $\phi(\theta) \doteq \theta^{D/2}$, as stated in (5.7). More importantly, (5.11) then gives $f_\theta(x)G^{(n)}(d\mathbf{x}) \doteq g_{\beta_1}^{(n\theta)}(d\mathbf{x})G^{(n\theta)}(d\mathbf{x})$, verifying that f_θ is approximately the same as $g_{\beta_1}^{(n\theta)}$: the hypothesis H_θ amounts to setting $\beta = \beta_1$, but reducing the sample size from n to $\nu = n\theta$. This interpretation of $\mathscr{F}$ is exactly correct for the normal random effects model, and gives excellent approximations in general exponential families.

As a simple example consider the case where x is real, $D = 1$, and $\mathscr{G}$ is the normalized binomial family $x \sim \mathrm{Bi}(n, \beta)/n$; that is, β is a probability, and x is an observed proportion, with expectation β. Take $\beta_1 = .5$ and $n = 20$. In this case $T(x, \beta_1) = 2[x\log(x/.5) + (1 - x)\log((1 - x)/.5)]$. For $\theta = .5$, expression (5.6), with $\phi(\theta)$ approximated by $\theta^{1/2}$, gives the following values for $f_\theta(x)$:

	x:	0	$1/20$	$2/20$	$3/20$	$4/20$	$5/20$	$6/20$	$7/20$	$8/20$	$9/20$	$10/20$
	f_θ:	.001	.002	.005	.012	.022	.039	.060	.083	.014	.119	.125
(5.15)	$Bi(10, .5)/10$:	.001		.010		.044		.117		.205		.246
	(f_1):			(.000)	(.001)	(.005)	(.015)	(.037)	(.074)	(.120)	(.160)	(.176)

with $f_\theta(11/20) = f_\theta(9/20)$, etc. Notice that f_θ is more dispersed about the central value

$x = {}^{10}\!/_{20}$ than is $f_1 \sim \mathrm{Bi}(20, .5)/20$. Making allowance for the change in support, f_θ closely matches the distribution $\mathrm{Bi}(10, .5)/10$, in accord with the effective sample size $\nu = 10$. The sum $\sum_x f_\theta(x) = 1.015$ from (5.15) shows the accuracy of the approximation $\phi(\theta) \doteq \theta^{1/2}$.

The family $\mathscr{G}$ of particular interest in this paper is that of Section 3, $g_\pi(\mathbf{p} \mid \mathbf{r}, \mathbf{c})$, the conditional distribution of $\mathbf{p} \sim \mathrm{Mult}_{IJ}(n, \pi)/n$ given $(\mathbf{r}, \mathbf{c})$. For this example we will work backwards, from the reduced sample size interpretation of $\mathscr{F}$ to the form (5.6). With the sample size reduced from n to $n\theta$, (3.7) becomes

$$(5.16) \qquad H_\theta\colon f_\theta(\mathbf{p} \mid \mathbf{r}, \mathbf{c}) \doteq \theta^D \left(\binom{n\theta}{n\theta\mathbf{p}} \Big/ \binom{n\theta}{n\theta\mathbf{r}} \binom{n\theta}{n\theta\mathbf{c}} \right) \quad (\mathbf{p} \in \mathscr{Y}^{(n)}(r, c)).$$

The factor θ^D would not appear in (5.16) if we restricted $\mathbf{p}$ to be in $\mathscr{Y}^{(n\theta)}(\mathbf{r}, \mathbf{c})$, since then we would have only changed the sample size in (3.7) from n to $\nu = n\theta$. However the density of lattice points in $\mathscr{Y}^{(n)}(\mathbf{r}, \mathbf{c})$ is θ^{-D} times the density in $\mathscr{Y}^{(n\theta)}(\mathbf{r}, \mathbf{c})$, formula (3.13), so the factor θ^D is necessary to make (5.16) sum, approximately, to one.

Stirling's approximation $x! \doteq \sqrt{2\pi x} x^x e^{-x}$, applied to the formula for $f_\theta(\mathbf{p} \mid \mathbf{r}, \mathbf{c})$ in (5.16), gives, without any further approximation,

$$(5.17) \qquad H_\theta\colon f_\theta(\mathbf{p} \mid \mathbf{r}, \mathbf{c}) = g_1(\mathbf{p} \mid \mathbf{r}, \mathbf{c}) \exp[n(1 - \theta) T(p, \hat{\pi})/2] \theta^{D/2},$$

where

$$(5.18) \qquad T(\mathbf{p}, \hat{\pi}) = 2 \textstyle\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \log(p_{ij}/\hat{\pi}_{ij}).$$

This is form (5.6), (5.7) for $\mathscr{F}$, except for one additional approximation: (5.18) is twice the Kullback-Leibler distance for the unconditional multinomial family, rather than for the multinomial family conditioned on $(\mathbf{r}, \mathbf{c})$. Starting from (5.16), which doesn't involve these last two approximations, standard asymptotic calculations give (4.12), the key result in Section 4.

Going back to general exponential families, consider the approximation to (5.6) obtained using (5.7),

$$(5.19) \qquad f_\theta(\mathbf{x}) \doteq g_{\beta_1}^{(n)}(\mathbf{x}) \exp[n(1 - \theta) T(\mathbf{x}, \beta_1)/2] \theta^{D/2}.$$

Differentiation w.r.t. θ gives the approximate maximum likelihood estimates

$$(5.20) \qquad \hat{\theta} \doteq D/nT, \quad \hat{\nu} = n\hat{\theta} \doteq D/T.$$

This is an improved version of the formula $\hat{\nu} = D/S$ used in Section 4. A standard Taylor series expansion shows that

$$(5.21) \quad T(\mathbf{x}, \beta_1) = S(\mathbf{x}, \beta_1) + O(\| \mathbf{x} - \beta_1 \|^3) \quad (S = (\mathbf{x} - \beta_1)' \Sigma_{\beta_1}^{-1} (\mathbf{x} - \beta_1))$$

so S approximates T for $\mathbf{x}$ near β_1. In the context of Section 3, $\Sigma_{\beta_1}^{-1} = \hat{\Sigma}^{-1}$, so for the two-way table situation S has the same meaning in (5.21) as in (3.5).

In Table 2 $T = .0211$ compared to $S = .0225$, so (5.20) gives $\hat{\nu} = 569$, compared to the estimate $D/S = 533$ of Section 4. In Table 1, $T = .2470$ compared to $S = .2336$. Again D/T gives about the same estimate of ν as D/S.

Applied in the context of Section 3, (5.19) also leads to a somewhat more exact confidence interval for ν,

$$\nu \in [\chi_D^{2(\alpha)}/T,\ \chi_D^{2(1-\alpha)}/T] \tag{5.22}$$

instead of the interval $[\chi_D^{2(\alpha)}/S, \chi_D^{2(1-\alpha)}/S)$ used in Section 4. (See formula (5.24).) However the numerical results are usually not much different, since S closely approximates T.

The interpretation of H_θ as being asymptotically equivalent to $g_{\beta_1}^{(n\theta)}$, (5.11)–(5.14), combines with (5.12) to give an important distributional result: with θ fixed, $\mathbf{x}$ has the limiting normal distribution, as $n \to \infty$,

$$H_\theta\colon \sqrt{n}(\mathbf{x} - \beta_1) \to N_D(\mathbf{0}, \Sigma_{\beta_1}/\theta). \tag{5.23}$$

In other words, *asymptotically θ acts as a simple scaling factor for the distribution of* $\mathbf{x}$. From (5.21) we then see that T and S are similarly scaled,

$$H_\theta\colon nT(\mathbf{x}, \beta_1) \to \chi_D^2/\theta, \quad nS(\mathbf{x}, \beta_1) \to \chi_D^2/\theta, \tag{5.24}$$

as in (4.12).

We could have gotten the asymptotic scaling property (5.23) for other definitions of $\mathscr{F}$, for example, replacing T by S in (5.6). However, definition (5.6) has another property which helps justify calling $\mathscr{F}$ a random effects model for $\mathscr{G}$; the maximum likelihood ratio test statistic for H_1 versus the full family, considered as a function of $\mathbf{x}$, *has the same contours of equal value in either $\mathscr{F}$ or $\mathscr{G}$*, namely the contours of equal value of $T(\mathbf{x}, \beta_1)$. In this sense, two values of $\mathbf{x}$ which provide equal evidence against H_1 in the family $\mathscr{G}$, also provide equal evidence against H_1 in the family $\mathscr{F}$.

It is easy to verify this property. First of all notice that $\hat{\theta}$, the MLE of θ obtained from (5.6), depends on $\mathbf{x}$ only through the statistic $T(\mathbf{x}, \beta_1)$, say $\hat{\theta} = \hat{\theta}(T)$, since $T(\mathbf{x}, \beta_1)$ is sufficient for θ in (5.6). The maximum likelihood ratio test statistic in $\mathscr{F}$ is

$$\sup_{\theta\in(0,1]} f_\theta(\mathbf{x})/f_1(\mathbf{x}) = \exp[n(1 - \hat{\theta})T(\mathbf{x}, \beta_1)/2]\phi(\hat{\theta}). \tag{5.25}$$

Comparing (5.25) with (5.3) gives

$$\sup_{\theta\in(0,1]} f_\theta(\mathbf{x})/f_1(\mathbf{x}) = [\sup_{\beta\in\mathscr{B}} g_\beta^{(n)}(\mathbf{x})/g_{\beta_1}^{(n)}(\mathbf{x})]^{1-\hat{\theta}(T)}\phi(\hat{\theta}(T)), \tag{5.26}$$

and shows that the maximum likelihood ratio statistic is a function only of T, in both $\mathscr{F}$ and $\mathscr{G}$.

How does the uniform distribution $g_0(\mathbf{p} \mid \mathbf{r}, \mathbf{c})$ considered in Section 3 relate to the family $\mathscr{F}$? For values of $\mathbf{p}$ near $\hat{\pi}$, and large values of n, it turns out that $g_0(\mathbf{p} \mid \mathbf{r}, \mathbf{c})$ is the limit as $\theta \to 0$ of $f_\theta(\mathbf{p} \mid \mathbf{r}, \mathbf{c})$. These results are easy to verify from (5.9), (5.12), which give

$$\begin{aligned} \lim_{\theta\to 0}(f_\theta(\mathbf{x})G^{(n)}(d\mathbf{x})/\phi(\theta)) &= g_{\mathbf{x}}^{(n)}(\mathbf{x})G^{(n)}(d\mathbf{x}) \\ &\doteq (n/2\pi)^{D/2}(d\mathbf{x}/\lvert \Sigma_{\mathbf{x}} \rvert^{1/2}). \end{aligned} \tag{5.27}$$

For values of $\mathbf{x}$ near β_1 (e.g., $\mathbf{p}$ near $\hat{\pi}$), we see that $\lim_{\theta\to 0} f_\theta(\mathbf{x})$ is asymptotically constant in $\mathbf{x}$.

6. Volume and density calculations. We will now prove Theorems 1, 2, and 3 of Sections 2 and 3. These give the volume of $\mathscr{E}(\chi^2)$ in Figure 1 and $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ in Figure 2, and the density of lattice points in $\mathscr{S}_{IJ}^{(n)}$ and in $\mathscr{V}^{(n)}(\mathbf{r}, \mathbf{c})$.

First we need a precise description of $\mathscr{L}$, the D-dimensional linear subspace of those tables, like $p - \hat{\pi}$, having all margins 0. We use the lexicographic notation of Section 3, as in (3.5): Let Λ_I and Λ_J be $I \times (I-1)$ and $J \times (J-1)$ matrices such that

$$(6.1) \qquad (\mathbf{1}_I/\sqrt{I}, \Lambda_I) \quad \text{and} \quad (\mathbf{1}_J/\sqrt{J}, \Lambda_J)$$

are orthogonal matrices, with dimensions $I \times I$ and $J \times J$, respectively. Let Γ_J be the $IJ \times (J-1)$ matrix

$$(6.2) \qquad \Gamma_J' = (1/\sqrt{I})(\Lambda_J', \Lambda_J', \cdots, \Lambda_J'),$$

and let Γ_I be the $IJ \times (I-1)$ matrix

$$(6.3) \qquad \Gamma_I' = (1/\sqrt{J})(\lambda_1'\mathbf{1}_J', \lambda_2'\mathbf{1}_J', \cdots, \lambda_I'\mathbf{1}_J'),$$

where $\lambda_1, \cdots, \lambda_I$ are the rows of Λ_I. Finally, let Γ_1 be the $IJ \times (I+J-1)$ matrix

$$(6.4) \qquad \Gamma_1' = \begin{pmatrix} \mathbf{1}_{IJ}'/\sqrt{IJ} \\ \Gamma_I' \\ \Gamma_J' \end{pmatrix}.$$

LEMMA 1. *The matrix Γ_1' has orthonormal rows, and $\mathscr{L}$, the space of tables having all margins* 0, *is given by*

$$(6.5) \qquad \mathscr{L} = \{\mathbf{v}: \Gamma_1'\mathbf{v} = \mathbf{0}\}.$$

PROOF. Orthonormality follows directly from the orthogonality of the matrices (6.1). For any vector $\mathbf{v}$ in $\mathscr{R}^{IJ}$, $\mathbf{1}'/\sqrt{IJ} \cdot \mathbf{v} = \mathbf{v}_{++}/\sqrt{IJ}$; furthermore letting $\mathbf{p} \equiv \mathbf{v}/\mathbf{v}_{++}$, we have, using notation (3.1),

$$(6.6) \qquad \mathbf{q}_I \equiv \Gamma_I'\mathbf{p} = \Lambda_I'\mathbf{r}/\sqrt{J} \quad \text{and} \quad \mathbf{q}_J \equiv \Gamma_J'\mathbf{p} = \Lambda_J'\mathbf{c}/\sqrt{I}.$$

Therefore if $\mathbf{p}$ and $\hat{\pi}$ are probability tables having the same margins, then $\Gamma_1'(\mathbf{p} - \hat{\pi}) = 0$, and conversely. □

We can complete the $(I+J-1) \times IJ$ orthonormal matrix Γ_1' with a $D \times IJ$ orthonormal matrix Γ_2', such that $\Gamma = (\Gamma_1, \Gamma_2)$ is $IJ \times IJ$ orthogonal. The orthogonal transformation

$$(6.7) \qquad \mathbf{q} \equiv \begin{pmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \end{pmatrix} \equiv \begin{pmatrix} \Gamma_1' \\ \Gamma_2' \end{pmatrix} \mathbf{p}$$

rotates $\mathscr{S}_{IJ}$ in such a way that the $(I+J-1)$-dimensional vector $\mathbf{q}_1$ contains all the marginal information about the table $\mathbf{p}$ (including the fact that $p_{++} = 1$).

This transformation is often used when $I = J = 2$, where it has the form

$$\text{(6.8)} \qquad \mathbf{q} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}\mathbf{p},$$

$\mathbf{q}_2$ in this case being just the last coordinate of $\mathbf{q}$.

For a probability table $\mathbf{p}$, the first $I - 1$ coordinates of $\mathbf{r}$, say $\tilde{\mathbf{r}}$, determine the last coordinate $r_I = 1 - \sum_{i=1}^{I-1} r_i$, and likewise $\tilde{\mathbf{c}} = (c_1, \cdots, c_{J-1})'$ determines c_J. Formula (6.6) defines a linear mapping from $(\tilde{\mathbf{r}}, \tilde{\mathbf{c}})$ to $(\mathbf{q}_I, \mathbf{q}_J)$, and we need to know its Jacobian.

LEMMA 2. *The linear mapping $\tilde{\mathbf{r}} \to \mathbf{q}_I$ has Jacobian $d\tilde{\mathbf{r}}/d\mathbf{q}_I = J^{(I-1)/2}/\sqrt{I}$, and likewise $d\tilde{\mathbf{c}}/d\mathbf{q}_J = I^{(J-1)/2}/\sqrt{J}$, so that $(\tilde{\mathbf{r}}, \tilde{\mathbf{c}}) \to (\mathbf{q}_I, \mathbf{q}_J)$ has Jacobian*

$$\text{(6.9)} \qquad d(\tilde{\mathbf{r}}, \tilde{\mathbf{c}})/d(\mathbf{q}_I, \mathbf{q}_J) = I^{(J-2)/2}J^{(I-2)/2}.$$

PROOF. Let $\tilde{\mathbf{q}}_I = \sqrt{J}\mathbf{q}_I = \Lambda_I'\mathbf{r}$. Notice that as $\tilde{\mathbf{r}}$ ranges over the space $\tilde{\mathscr{S}}_I = \{r_1, r_2, \cdots, r_{I-1} \geq 0, \sum_1^{I-1} \mathbf{r}_i \leq 1\}$, $\tilde{\mathbf{q}}_I$ ranges over $\Lambda_I'\mathscr{S}_I$, a rotated and translated version of $\mathscr{S}_I$. However $\tilde{\mathscr{S}}_I$ has $(I - 1)$-dimensional volume $1/\Gamma(I)$, while $\mathscr{S}_I$ and also $\Lambda_I'\mathscr{S}_I$ have volume $\sqrt{I}/\Gamma(I)$. (See Hotelling, 1961, formula 7.6.) Therefore $d\tilde{\mathbf{q}}_I/d\tilde{\mathbf{r}} = \sqrt{I}$, implying $d\mathbf{q}_I/d\tilde{\mathbf{r}} = \sqrt{I}/J^{(I-1)/2}$. □

The density results (2.9) and Theorem 3 are obtained by Jacobian calculations. Consider for example the vector $\tilde{\mathbf{p}} = (p_1, \cdots, p_{I,J-1})'$, which is the table of observed proportions $\mathbf{p}$ without its last coordinate:

$$\tilde{\mathbf{p}} \in \tilde{\mathscr{S}}_{IJ}^{(n)} = \{\tilde{\mathbf{p}} = \tilde{\mathbf{m}}/n\colon m_{ij}\text{integer} \geq 0, \textstyle\sum\sum_{(i,j)\neq(I,J)} m_{ij}/n \leq 1\}.$$

Notice that $\tilde{\mathscr{S}}_{IJ}^{(n)}$ is a lattice of points in $(IJ - 1)$-dimensional space, with the points set in a regular cubic array having density n^{IJ-1}. The mapping $\tilde{\mathbf{p}} \to \mathbf{p}$ takes $\tilde{\mathscr{S}}_{IJ}$ into $\mathscr{S}_{IJ}$ and $\tilde{\mathscr{S}}_{IJ}^{(n)}$ into $\mathscr{S}_{IJ}^{(n)}$. The Jacobian $d\mathbf{p}/d\tilde{\mathbf{p}} = \sqrt{IJ}$ used in the proof of Lemma 2, which is the factor by which volumes multiply in mapping $\tilde{\mathbf{p}}$ to $\mathbf{p}$, shows that $\mathscr{S}_{IJ}^{(n)}$ must have lattice density $n^{IJ-1}/\sqrt{IJ}$, formula (2.9).

We can now prove Theorem 3. The vector $(\tilde{\mathbf{r}}, \tilde{\mathbf{c}})$ takes its values in a lattice of points in $(I + J - 2)$-dimensional space, with the points set in a regular cubic array of density n^{I+J-2}. By Lemma 2, the lattice points for $(\mathbf{q}_I, \mathbf{q}_J)$ have density $\text{dens}(\mathbf{q}_I, \mathbf{q}_J) = n^{(I+J-2)}I^{(J-2)/2}J^{(I-2)/2}$. The $(IJ - 1)$-dimensional vector $(\mathbf{q}_I, \mathbf{q}_J, \mathbf{q}_2)$, with $\mathbf{q}_2$ as in (6.7), has $\text{dens}(\mathbf{q}_I, \mathbf{q}_J, \mathbf{q}_2) = n^{IJ-1}/\sqrt{IJ}$, as in (2.9), since $\mathbf{q}$ is an orthogonal rotation of $\mathbf{p}$. However, if we let $\text{dens}(\mathbf{q}_2)$ represent the density of lattice points $\mathbf{q}_2$ with $(\mathbf{q}_I, \mathbf{q}_J)$ held fixed, then $\text{dens}(\mathbf{q}_I, \mathbf{q}_J, \mathbf{q}_2) = \text{dens}(\mathbf{q}_I, \mathbf{q}_J)\text{dens}(\mathbf{q}_2)$ by the orthogonality of the coordinates $(\mathbf{q}_I, \mathbf{q}_J)$ to $\mathbf{q}_2$, so

$$\text{(6.10)} \qquad \text{dens}(\boldsymbol{q}_2) = \frac{\text{dens}(\mathbf{q}_I, \mathbf{q}_J, \mathbf{q}_2)}{\text{dens}(\mathbf{q}_I, \mathbf{q}_J)} = \frac{n^D}{I^{(J-1)/2}J^{(I-1)/2}},$$

verifying Theorem 3. □

Theorem 2 can be verified by straightforward matrix manipulations, but it is much easier to use a trick based on the multivariate normal distribution: if $\mathbf{x} \sim N_D(\mathbf{0}, \Sigma)$ then

$$\text{volume}\{\mathbf{x}'\Sigma^{-1}\mathbf{x} \le \chi^2\} = (\chi^2/2)^{D/2}/[(D/2)!g^{\mathbf{x}}(\mathbf{0})], \tag{6.11}$$

where $g^{\mathbf{x}}(\mathbf{0}) = (2\pi)^{-D/2}|\Sigma|^{-1/2}$ is the density of $\mathbf{x}$ at $\mathbf{0}$. (Substituting this last expression for $g^{\mathbf{x}}(\mathbf{0})$ shows that (6.11) is just the usual formula $[\pi^{D/2}/(D/2)!]|\Sigma|^{1/2}\chi^D$ for the volume of an ellipsoid.) Formula (6.1) holds even in singular situations, where $\mathbf{x}$ has more than D coordinates but Σ is of rank D, provided that $g^{\mathbf{x}}(\cdot)$ is interpreted as the density of $\mathbf{x}$ in the D-dimensional space supporting the distribution of $\mathbf{x}$. "Volume" then refers to D-dimensional volume in the support space, and Σ^{-1} can be any pseudo-inverse of Σ.

We can use (6.11) to evaluate the volume of $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$, by letting $\mathbf{x} = \sqrt{n}(\mathbf{p} - \hat{\pi})$. Definition (3.8) of $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ does not involve n, so we can take n to be as large as we wish. Asymptotically, $\mathbf{p}$ has a conditional normal distribution in $\mathscr{V}(\mathbf{r}, \mathbf{c})$ under H_1, as in (4.15), $\sqrt{n}(\mathbf{p} - \hat{\pi}) \mid (\mathbf{r}, \mathbf{c}) \to N_D(\mathbf{0}, \hat{\Sigma})$. The pseudo-inverse of $\hat{\Sigma}$ is the matrix $\hat{\Sigma}^{-1}$ in (3.5), see Bishop, Fienberg and Holland (1975), Section 14.9.2. It remains only to compute the quantity $g^{\mathbf{x}}(\mathbf{0})$.

Consider the Fisher-Yates density (3.7), with $(\mathbf{r}, \mathbf{c})$ fixed, $\mathbf{p} = \hat{\pi}$, and n tending toward infinity. Stirling's formula gives

$$\lim_{n\to\infty} n^{D/2} g_1(\hat{\pi} \mid \mathbf{r}, \mathbf{c}) = (1/2\pi)^{D/2}[(\textstyle\prod_{i=1}^{I} r_i)^{J-1}(\prod_{j=1}^{J} c_j)^{I-1}]^{-1/2}. \tag{6.12}$$

The limiting density at the center of the Fisher-Yates distribution is $g_1(\hat{\pi} \mid \mathbf{r}, \mathbf{c})$ times the lattice density (6.10),

$$g^{\mathbf{x}}(\mathbf{0}) \to (n/2\pi)^{D/2}[(I \textstyle\prod r_i)^{J-1}(J \prod c_j)^{I-1}]^{-1/2.} \tag{6.13}$$

Substituting $g^{\mathbf{x}}(\mathbf{0})$ into (6.11) gives (3.11). □

Theorem 1 is an easy consequence of Theorem 2. We make the orthogonal transformation (6.7), and notice that for $\mathbf{q}_1$ fixed, equivalently for $(\mathbf{q}_I, \mathbf{q}_J)$ fixed, the corresponding slice of $\mathscr{E}(\chi^2)$ is the ellipsoid $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$, with volume (3.11). The change of coordinates $(\mathbf{q}_I, \mathbf{q}_J) \to (\tilde{\mathbf{r}}, \tilde{\mathbf{c}})$ allows us to evaluate $\int \mathscr{E}(S \mid \mathbf{r}, \mathbf{c})\, d(\mathbf{q}_I, \mathbf{q}_J)$ using a standard Dirichlet integral and Jacobian (6.9). This gives (2.10). This proof uses the fact that the ellipsoids $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ are in parallel subspaces for all $(\mathbf{r}, \mathbf{c})$, so that the integration is easy to perform.

7. The volume of the slice $\mathscr{V}(\mathbf{r}, \mathbf{c})$. The volume of the slice $\mathscr{V}(\mathbf{r}, \mathbf{c})$ played an important role in the conditional volume test of Section 3. Here, we will motivate approximation (3.14) and give an exact formula for the case $\min(I, J) \le 3$.

LEMMA 3. *The D-dimensional volume of $\mathscr{V}(\mathbf{r}, \mathbf{c})$ is*

$$V(\mathbf{r}, \mathbf{c}) = \frac{I^{(J-1)/2}J^{(I-1)/2}}{\Gamma(J)^I}\,(\textstyle\prod_{i=1}^{I} r_i)^{J-1} g(\tilde{\mathbf{c}} \mid \mathbf{r}), \tag{7.1}$$

where $g(\tilde{\mathbf{c}} \mid \mathbf{r})$ is the density function evaluated at $\tilde{\mathbf{c}} = (c_1, \cdots, c_{J-1})$ of a mixture, *according to weights r_i, of I independent Dirichlet vectors,*

$$\sum_{i=1}^{I} r_i D_J(1, 1, \cdots, 1). \tag{7.2}$$

PROOF. Suppose that the vector $\mathbf{p} \in \mathscr{S}_{IJ}$ is drawn according to the flat Dirichlet distribution $D_{IJ}(\mathbf{1}_{IJ})$, which is to say that the first $IJ - 1$ coordinates of $\mathbf{p}$ have the distribution described immediately after (2.11). The orthogonal rotation (6.7) makes $\mathbf{q}$ uniformly distributed over a rotated simplex. Using the notation of Section 6, we see that $\mathbf{q}_2$ given $(\mathbf{q}_I, \mathbf{q}_J)$ has a uniform distribution over a rotated version of $\mathscr{V}(\mathbf{r}, \mathbf{c})$, and therefore the conditioned density of $\mathbf{q}_2$ must be

$$g(\mathbf{q}_2 \mid (\mathbf{q}_I, \mathbf{q}_J)) = 1/V(\mathbf{r}, \mathbf{c}). \tag{7.3}$$

The density of the uniformly distributed vector $(\mathbf{q}_I, \mathbf{q}_J, \mathbf{q}_2)$ is $\Gamma(IJ)/\sqrt{IJ}$, one over the volume of $\mathscr{S}_{IJ}$, so

$$\frac{1}{V(\mathbf{r}, \mathbf{c})} = g(\mathbf{q}_2 \mid (\mathbf{q}_I, \mathbf{q}_J)) = \frac{g(\mathbf{q}_I, \mathbf{q}_J, \mathbf{q}_2)}{g(\mathbf{q}_I, \mathbf{q}_J)} = \frac{\Gamma(IJ)}{\sqrt{IJ}} \frac{1}{g(\mathbf{q}_I, \mathbf{q}_J)}. \tag{7.4}$$

According to Lemma 2, $g(\mathbf{q}_I, \mathbf{q}_J) = I^{(J-2)/2} J^{(I-2)/2} g(\tilde{\mathbf{r}}, \tilde{\mathbf{c}})$. Also $g(\tilde{\mathbf{r}}, \tilde{\mathbf{c}}) = g(\tilde{\mathbf{r}}) g(\tilde{\mathbf{c}} \mid \tilde{\mathbf{r}}) = [\Gamma(IJ)/\Gamma(J)^I](\prod r_i)^{J-1} g(\tilde{\mathbf{c}} \mid \tilde{\mathbf{r}})$, the last equality following from the fact that $\mathbf{p} \sim D_{IJ}(\mathbf{1}_{IJ})$ implies $\mathbf{r} \sim D_I(J, J, \cdots, J)$. Combining this with (7.4) gives (7.1). The interpretation (7.2) of $g(\tilde{\mathbf{c}} \mid \tilde{\mathbf{r}})$ is a well-known property of the distribution $\mathbf{p} \sim D_{IJ}(\mathbf{1}_{IJ})$, see Wilks (1962) result 7.7.5. □

The density $g(\tilde{\mathbf{c}} \mid \tilde{\mathbf{r}})$ is difficult to calculate exactly. It seems reasonable to approximate $g(\tilde{\mathbf{c}} \mid \tilde{\mathbf{r}})$ with a symmetric Dirichlet distribution since unconditionally $\mathbf{c} \sim D_J(I, I, \cdots, I)$. The approximation $\mathbf{c} \mid \mathbf{r} \sim D_J(k_r, \cdots, k_r)$ where

$$k_r = (J + 1)/J \,\|\, \mathbf{r} \,\|^2 - 1/J, \tag{7.5}$$

has the same mean vector and covariance matrix as the mixture $\sum r_i D_J(1, \cdots, 1)$ in (7.2). Substituting in (7.1) the density of $D_J(k_r, \cdots, k_r)$ at $\tilde{\mathbf{c}}$ for $g(\tilde{\mathbf{c}} \mid \tilde{\mathbf{r}})$ gives the approximation

$$V(\mathbf{r}, \mathbf{c}) \doteq \frac{I^{(J-1)/2} J^{(I-1)/2} \Gamma(Jk_r)}{\Gamma(J)^I \Gamma(k_r)^J} \left(\prod_{i=1}^{I} r_i\right)^{J-1} \left(\prod_{j=1}^{J} c_j\right)^{k_r - 1}. \tag{7.6}$$

Some numerical evidence on the accuracy of (7.6) appears in Table 6. The worst results are for the 2 × 2 case, where we do not need to use an approximation since there is a simple exact formula for (7.1), see (7.8). Overall, the accuracy is quite satisfactory. (Note: Good's, 1976, (6.5) approximation is closely related to our (7.6). The difference amounts to using the unconditional distribution of $\mathbf{c}$ rather than the moment-matching distribution $D_J(k_r, \cdots, k_r)$ to approximate $g(\tilde{\mathbf{c}} \mid \mathbf{r})$ in (7.1).)

Approximation (3.14) incorporates one further improvement on (7.6) having to do with *edge effects* occurring at the boundaries of $\mathscr{V}\mathbf{r}, \mathbf{c})$. As a simple example of edge effects, consider a line segment of length $L = N\Delta$ having $N + 1$ points

TABLE 6
Accuracy of approximation (7.6) *for* $V(\mathbf{r}, \mathbf{c})$.

r	**c**	**formula (7.6)**	**true** $V(\mathbf{r}, \mathbf{c})$
(.5, .5)	(.5, .5)	.849	1.000
(.1, .9)	(.5, .5)	.214	.200
(.1, .9)	(.1, .9)	.153	.200
(.05, .95)	(.1, .9)	.0884	.100
(.2, .3, .5)	(.4, .6)	.1900	.1905
(⅓, ⅓, ⅓)	(⅓, ⅓, ⅓)	.0132	.0139
(⅙, ⅙, ⋯, ⅙)	(⅙, ⅙, ⋯, ⅙)	$2.59 \cdot 10^{-28}$	$2.58 \cdot 10^{-28}$
	Table 1	$7.2 \cdot 10^{-9}$	$6.5 \pm .3 \cdot 10^{-9}$
	Table 2	$5.9 \cdot 10^{-17}$	$5.7 \pm .2 \cdot 10^{-17}$

placed regularly along it at equal intervals Δ. The density of points is $1/\Delta$, so length-times-density is $L \cdot (1/\Delta) = N$, rather than the correct number of points $N + 1$. The reason, of course, is that the line segment has a lattice point at each end, so its effective length as far as the length-times-density argument is concerned is obtained by adding half an interval, $\Delta/2$, at each end: then $(L + \Delta) \cdot (1/\Delta) = N + 1$, the correct number of points.

The same type of adjustment substantially improves the volume-times-density approximation for the number of lattice points in $\mathscr{V}(\mathbf{r}, \mathbf{c})$. Let $\mathscr{V}_+(\mathbf{r}, \mathbf{c})$ be $\mathscr{V}(\mathbf{r}, \mathbf{c})$ as defined in (3.6) except with the constraint $\pi_{ij} \geq 0$ replaced by $\pi_{ij} \geq -1/(2n)$, for all i and j. This is the equivalent of adding $\Delta/2$ at each end of the line segment. It is easy to show that

$$\text{volume } \mathscr{V}_+(\mathbf{r}, \mathbf{c}) = (1 + IJ/2n)^D \cdot \text{volume } \mathscr{V}(\bar{\mathbf{r}}, \bar{\mathbf{c}}), \tag{7.7}$$

with $\bar{\mathbf{r}}$, $\bar{\mathbf{c}}$ defined as in (3.15). Formula (3.14) is obtained from (7.6) via (7.7). Formulas (2.5) and (3.18) incorporate similar corrections.

From formulas like A1.4 in Good (1976), it can be shown that correcting for edge effects in this way produces the right second-order asymptotic expansion for $N^{(n)}(\mathbf{r}, \mathbf{c})$. The correction is often quite substantial, being about 20% in Table 1, for instance.

Exact formulas for $V(\mathbf{r}, \mathbf{c})$ are possible if $\min(I, J) = 2$ or 3. The approach is through the exact evaluation of $g(\tilde{\mathbf{c}} \mid \mathbf{r})$ in (7.1). For $J = 2$ (or equivalently, by interchanging rows and columns, for $I = 2$), $\tilde{\mathbf{c}} = c_1$, and an argument based on Laplace transforms gives

$$\begin{aligned} g(\mathbf{c}_1 \mid \mathbf{r}) = \frac{1}{\Gamma(I)\prod_{i=1}^{I} r_i} [c_1^{I-1} - \textstyle\sum_{|A|=1} (c_1 - r_A)_+^{I-1} \\ + \textstyle\sum_{|A|=2} (c_1 - r_A)_+^{I-1} - \cdots (-1)^I (c_1 - 1)_+^{I-1}]. \end{aligned} \tag{7.8}$$

The symbol A represents any subset of $\{1, 2, \cdots, I\}$, $|A|$ is the number of elements in A, $r_A = \sum_{i \in A} r_i$, and $(c_1 - r_A)_+ = \max(0, c_1 - r_A)$.

Higher dimensions make the Laplace transform argument more difficult to apply. For $J = 3$, extensive computations give

$$g(\tilde{\mathbf{c}} \mid \mathbf{r}) = \frac{2^I}{\prod_{i=1}^I r_\iota^2} \Bigg\{ (c_1 c_2)^{I-1} + \textstyle\sum_{A,B} (-1)^{I+b-a} \sum_{\iota=0}^{I-b} \binom{I-a+i-1}{i}$$

$$\cdot \frac{(c_1 + c_2 - r_A - r_B)_+^{I-b-i-1} (r_B - c_2)_+^{I+b+\iota-1}}{\Gamma(I-b-i)\Gamma(I+b+i)} \tag{7.9}$$

$$+ \textstyle\sum_{A,B} (-1)^b \sum_{i=0}^{I-a-1} \binom{I-b+i-1}{i} \frac{(c_1 + c_2 - r_A - r_B)_+^{I-a-i-1} (c_1 - r_A)_+^{I+a+i-1}}{\Gamma(I-a-i)\Gamma(I+a+i)} \Bigg\}.$$

The symbol $\sum_{A,B}$ indicates summation over all disjoint subsets A and B of $\{1, 2, \cdots, I\}$; $r_A = \sum_{i \in A} r_i$, $r_B \sum_{i \in B} r_i$: and $\binom{x}{0} = 1$ for any value of x.

8. Granularity and protrusion effects. The volume-times-density argument used to calculate the number of points in the numerator of the volume test significance level ignores two things: that we are dealing with a discrete lattice of points rather than with a continuous uniform distribution (granularity); and that the ellipsoid in Figure 2 or the elliptical tube in Figure 1 does not stay within the boundaries of the simplex (protrusion). This section gives a brief discussion of those two effects, in the conditional testing framework of Section 3.

Granularity by itself has little effect on our calculations. Suppose that the ellipsoid $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ of Figure 2 lies entirely within $\mathscr{V}(\mathbf{r}, \mathbf{c})$, so that protrusion can be ignored. A great deal of mathematical effort has gone into verifying the accuracy of the volume-times-density approximation; see, for example, Leveque (1971) Section 1, page 24; and Kendall and Moran (1963) Section 5. Under reasonable conditions, the asymptotic error of the approximation will be $o_p(1/\sqrt{n})$, which is to say smaller than typical statistical variation.

From another perspective, granularity produces no error at all. Suppose that in Figure 2 we make the continuity correction of spreading the mass $1/N^{(n)}(\mathbf{r}, \mathbf{c})$ at each lattice point $\mathbf{p}$ uniformly over the small hexagon of planer points nearest to $\mathbf{p}$: $\{\boldsymbol{\pi}: \min_{\mathbf{p}^* \in \mathscr{V}^{(n)}(\mathbf{r},\mathbf{c})} \| \mathbf{p}^* - \boldsymbol{\pi} \| = \| \mathbf{p} - \boldsymbol{\pi} \| \}$. Then by definition the volume-times-density argument gives exactly the continuity-corrected probability content of $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$. (The correction for edge effects in the denominator of (3.17), formula (7.7), can be motivated by this same argument.)

Protrusion is the most serious source of error in (3.17), often causing considerable overestimates of the actual volume test significance level. This was unimportant in Table 2, where (3.17) itself gave an extremely small number, but was troublesome in Table 1. The following result, which will not be derived here, improves on Theorem 2 and gives a diagnostic for the existence of large outsideness effects.

THEOREM 4. *For a given value of i and j, the D-dimensional volume of the portion of the ellipsoid $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ having $p_{ij} > -\frac{1}{2}n$ equals*

$$[\text{volume } \mathscr{E}(S \mid \mathbf{r}, \mathbf{c})] \cdot \omega_D \left[\frac{1}{S} \frac{(r_i c_j + (1/2n))^2}{r_i(1 - r_i) c_j (1 - c_j)} \right], \tag{8.1}$$

where

$$\omega_D[t] = \frac{\Gamma(k/2 + 1)}{\sqrt{\pi}\,\Gamma(k/2 + \frac{1}{2})} \int_{-\pi/2}^{\sin^{-1}(t)} (\cos\theta)^D \, dt \tag{8.2}$$

for $t \leq 1$, $\omega[t] = 1$ *for* $t > 1$.

The factor ω_D in (8.1) reduces (3.11) to allow for the amount of $\mathscr{E}(S \mid \mathbf{r}, \mathbf{c})$ outside *one* of the boundaries of $\mathscr{V}(\mathbf{r}, \mathbf{c})$. The reason for stating the boundary condition as $p_{ij} > -\frac{1}{2}n$ rather than as $p_{ij} > 0$ has to do with edge effects, as in (7.7). Table 7 shows ω_D for every choice of (i, j) in Tables 1 and 2, correctly indicating the large outsideness effects. The smallest of these factors times (3.17) still tends to be an overestimate of $\varepsilon(S \mid \mathbf{r}, \mathbf{c})$: .25 compared to the true value .09 in Table 1; $2.5 \cdot 10^{-5}$ compared to true value $1.2 \cdot 10^{-5}$ in Table 2.

The actual values of $\varepsilon(S \mid \mathbf{r}, \mathbf{c})$ for Tables 1 and 2 were calculated by an inexpensive Monte Carlo method:

(1) Choose a reduced sample size $\nu = n\theta$ near the MLE value $\hat{\nu} = D/S$ (e.g., $\nu = 40$ for Table 1).

(2) Draw Monte Carlo samples $\mathbf{p}^*$ from the Fisher-Yates distribution (3.7), with $\mathbf{r}$ and $\mathbf{c}$ as in the observed table, but with $n = \nu$. This can be done efficiently by sampling without replacement, in the usual hypergeometric manner. Fractional values of $\nu\mathbf{r}$ or $\nu\mathbf{c}$ can be handled by a variety of ad hoc devices.

(3) Calculate the Monte Carlo expectation of $\text{IND}[S^* \leq S]/g_1(\mathbf{p}^* \mid \mathbf{r}, \mathbf{c})$ where IND is the indicator function, $S^* = (\mathbf{p}^* - \boldsymbol{\pi})'\hat{\Sigma}^{-1}(\mathbf{p}^* - \boldsymbol{\pi})$, $\hat{\boldsymbol{\pi}}$ and $\hat{\Sigma}$ are fixed at their observed values, and $g_1(\mathbf{p}^* \mid \mathbf{r}, \mathbf{c})$ is the Fisher-Yates density (3.7), with $n = \nu$.

Except for edge effects, this method gives an unbiased estimate of the numerator of (3.10), with $n = \nu$. (Between 250 and 500 Monte Carlo repetitions were sufficient to give 10% accuracy for Tables 1 and 2.) The corresponding denominator $N^{(\nu)}(\mathbf{r}, \mathbf{c})$ was approximated by (3.13) × (3.14), $n = \nu$. A simple geometric analysis of the edge effects indicates that ε_ν, (3.10) with $n = \nu$, satisfies $\varepsilon_\nu \doteq a - b/\nu$, $b > 0$. Trying the Monte Carlo analysis for different values of ν

TABLE 7

Edge effect factors ω_D for each choice of (i, j) in Table 1 (left) and Table 2 (right); dashes indicate $\omega_D = 1.00$.

i	j				i	j			
	1	2	3	4		1	2	3	4
1	.98	—	.87	—	**1**	—	—	—	—
2	.97	—	.87	.99	**2**	—	—	—	—
3	.72	—	.64	.76	**3**	—	—	—	—
4	.65	.95	.60	.68	**4**	.96	—	.91	.76
					5	.95	.81	.62	.57

verified this for Tables 1 and 2. For example in Table 1 we computed

$$(8.3)\qquad \frac{\nu = \;\; 40 \;\; 60 \;\; 80}{\varepsilon_\nu = .036\ .051\ .069}\,.$$

The value $\varepsilon(S \mid \mathbf{r}, \mathbf{c}) = .093$ was obtained by extrapolating in $1/\nu$ to the actual sample size $\nu = 592$.

9. Review of the literature. Connections between goodness of fit, small p-values and large sample sizes have been made many times before. Berkson (1938) raised the issue by noting that with small sample sizes we can often find models to give a satisfactory fit. With large samples, no model fits.

Hodges and Lehmann (1954) suggest a cure for the problem: test to see if the data are compatible with a model that is close to the null hypothesis. In testing for independence, they regard an observed table of counts as a point in the simplex and accept independence if the distance between the observed point and the surface of independence is smaller than a cutoff c. They do not suggest an explicit way to choose c. To tie their test to the usual 5% level, they suggest that if the distance is larger than c, then a 5% test of "distance $= c$" be performed. In a sense, our paper suggests ways to choose and interpret values of c.

Martin-Löf (1974) also suggests supplementing the usual test with a quantitative measure of the size of the discrepancy between the statistical model and observed data. The aim is to see if the discrepancy, although highly significant, is so small that the model must be considered as providing a satisfactory approximation of the data.

In testing for independence in a two-way table, Martin-Löf's discrepancy is (approximately) the chi-square statistic divided by n times the sum of the entropies of the marginal distributions of the table. Martin-Löf gives several justifications for this choice: the discrepancy can be interpreted as the relative decrease in the number of bits needed to specify the table, given the margins and the value of the usual test, as compared with the number of bits necessary to specify the margins only. An alternative interpretation comes from considering an exponential family through the test statistic parametrized in such a way that the null hypothesis corresponds to some of the parameters being zero. The discrepancy is then (approximately) minus the relative entropy distance between the maximum likelihood member of the family and the maximum likelihood member of the family subject to the null hypothesis being true.

Martin-Löf calculates a number of examples in an effort to calibrate a discrepancy scale. One of the examples is our Table 2. He finds that the discrepancy falls nicely between a good fit and a bad fit on the scale he suggests. Our analysis states the same conclusion in terms of the relative sample size. Hald (1971) treats the testing problem decision-theoretically with both prior and loss function specified. He suggests that an "indifference-zone" can be introduced around the null hypothesis by having zero, or small, loss there. He investigates the asymptotic properties of such tests in one-dimensional problems. Hald is mainly concerned with the widely perceived intuitive feeling that significance levels should sometimes decrease with increasing sample size. His analysis shows when this feeling is correct for Bayesian tests.

Our results can be put into a rough decision theoretic framework by specifying the following program to go along with standard tests: First, carry out the standard test. If it rejects, test to see if an alternative far from the null is compatible with the data. If yes, reject the null. If no, carry out the components of variance approach suggested. If the reduced sample size is not much smaller than the observed sample size, report that the data are compatible with the null hypothesis. If not, report the reduced sample as an indication of how far the data are from the null hypothesis.

Work on robustness of tests also bears on our problem. Chapter 10 of Huber (1981) contains an up-to-date review. Along these lines, Ylvisaker (1977) introduces a notion of resistance which measures the smallest proportion of the data that can be altered to force the level of the test statistic to its observed value. If a believably small proportion of misclassifications can inflate the test statistic to its observed value, this gives grounds for accepting the null hypothesis. Neither of these authors explicitly discusses tests for independence.

Bayesian statisticians have carried out work related to the subject of our paper. Lindley (1964) shows how the usual χ^2 test can be given a Bayesian interpretation. Gûnel and Dickey (1974), Bishop, Fienberg and Holland (1975) carry through Bayesian analysis when Dirichlet priors have been put on the parameters involved. The most extensive contributions to the subject come from I. J. Good. The papers most closely connected to ours are Good (1976) and Crook and Good (1980). Points of contact are described in Sections 2, 3 and 7. A current survey of his many related notes and papers on this subject is in Good (1983). While our volume test arises from the Bayesian assumption of a uniform prior on the simplex, we do not know if the other reference distributions we suggest arise (approximately) from nonuniform distributions, such as the mixtures of symmetric Dirichlets suggested by Good.

Hotelling (1939) proposed what we are calling the volume test for a class of testing problems in nonlinear regression. His paper generated considerable activity in the mathematical literature, but except for scattered applications, as in Efron (1971), its statistical content seems to have gone largely ignored.

10. Acknowledgement. The authors thank I. J. Good, Richard Melrose, David Rogosa, Mehrdad Shahshahani, and Lawrence Shepp for many helpful discussions during the preparation of this paper.

REFERENCES

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Statist. Assoc.* **33** 526–542.

Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis.* M.I.T. Press, Cambridge, MA.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton, NJ.

Crook, J. F. and Good, I. J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, part II. *Ann. Statist.* **8** 1198–1218.

DIACONIS, P. and EFRON, B. (1983). Generalized variance of the multinomial and Fisher-Yates distributions. *Stanford University Technical Report No. 208.*

DICKEY, J. M. (1983). Multiple hypergeometric functions: probabilistic interpretations and statistical uses. *J. Amer. Statist. Assoc.* **78** 628–639.

EFRON, B. (1971). Does an observed sequence of numbers follow a simple rule? (Another look at Bode's law). *J. Amer. Statist. Assoc.* **66** 552–559.

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376.

GOOD, I. J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4** 1159–1189.

GOOD, I. J. (1983). The robustness of a hierarchical model for multinomials and contingency tables. In *Scientific Inference Data Analysis, and Robustness.* (G. Box, T. Leonard, C. F. Wu, eds.) Academic, New York.

GOOD, I. J. and CROOK, J. F. (1977). The enumeration of arrays and a generalization related to contingency tables. *Discrete Math.* **19** 23–45.

GOOD, I. J. and CROOK, J. F. (1980). The information in the margins of a 2 × 2 contingency table. Unpublished report.

GÛNEL, E. and DICKEY, J. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61**, 545–557.

HALD, A. (1971). The size of Bayes and minimax tests as functions of the sample size and the loss ratio. *Skand. Aktuar. Tidskr.* **00** 53–73.

HEYDE, C. and SENETA, E. (1977). *I. J. Bienaymé: Statistical Theory Anticipated.* Springer-Verlag, New York.

HODGES, J. L. and LEHMANN, E. L. (1954). Testing the approximate validity of statistical hypotheses. *J. Roy. Statist. Soc. Ser. B* **16** 261–268.

HOEFFDING, W. (1965). Asymptotically optimum tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–400.

HOTELLING, H. (1939). Tubes and spheres in n-spaces, and a class of statistical problems. *Amer. J. Math.* **61** 440–460.

HOTELLING, H. (1961). The behavior of some standard statistical tests under nonstandard conditions. *Proc. Fourth Berk. Symp.* **I** 319–360.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

KENDALL, M. and MORAN, P. (1963). *Geometrical Probability.* Hafner, New York.

LAIRD, N. M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65** 581–590.

LEONARD, T. (1977). A Bayesian approach to some multinomial estimation and pretesting problems. *J. Amer. Statist. Assoc.* **72** 869–874.

LEVEQUE, W. (1971). *Reviews in Number Theory,* 4. Amer. Math. Soc., Providence, RI.

LINDLEY, D. V. (1964). The Bayesian analysis of contingency tables. *Ann. Math. Statist.* **35** 1622–1643.

MARTIN-LÖF, P. (1974). The notation of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. *Scand. J. Statist.* **1** 3–18.

SHAFER, G. (1982). Lindley's paradox (with discussion). *J. Amer. Statist. Assoc.* **77** 325–351.

SNEE, R. (1974). Graphical display of two-way contingency tables. *Amer. Statist.* **38** 9–12.

STONE, C. (1965). A local limit theorem for nonlattice multi-dimensional distribution functions. *Ann. Math. Statist.* **36** 546–551.

WILKS, S. (1962). *Mathematical Statistics.* Wiley, New York.

YLVISAKER, D. (1977). Test resistance. *J. Amer. Statist. Assoc.* **72** 551–556.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305