REJOINDER

P. DIACONIS AND B. EFRON

Stanford University

The basic elements of our paper, volume testing and components of variance, are staples of good practice in the application of ordinary linear models. Table A gives a simple example. Twenty dice have been seized from a fictitious Las Vegas casino, on the suspicion that the casino has subtly overweighted the occurrence of "1" and "6." The police roll each die 50,000 times, obtaining the results shown. Is the casino guilty?

Let m_i be the number of occurrences observed for the ith die. The total $m = \sum_{i=1}^{20} m_i$ equals 335,294. If the dice are fair, m is nearly normally distributed with mean 333,333.3 and standard deviation $[10^6(\frac{1}{3})(\frac{2}{3})]^{1/2} = 471.4$. The obvious normal test statistic $z = [335{,}294 - 333{,}333.3]/471.4$ equals 4.16, significance level $2 \cdot 10^{-5}$, overwhelming proof against the hypothesis of fairness. The multidimensional test statistic

$$\chi^2 = \sum_{i=1}^{20} (m_i - 16{,}666.7)^2/[50{,}000(\tfrac{1}{3})(\tfrac{2}{3})] = 687.9$$

is also overwhelmingly significant when compared to the null distribution χ^2_{20}. On the other hand, the t-statistic

$$t = (\bar{m} - 16{,}666.7)/[\Sigma(m_i - \bar{m})^2/(19 \cdot 20)]^{1/2} = 0.70$$

is not at all significant, with attained level only .246 compared to a t_{19} distribution. Which test should we believe?

The correct answer is the t-test. The dice are unfair, as the χ^2 test shows, but not in the systematic manner of which the casino was accused. The trouble with the z-test, which is just the t-test applied to all 1,000,000 rolls, is that the sample size in this problem is really 20 and not 1,000,000. This is clear from a components of variance analysis of the m_i, which indicates that the component due to variation between dice is about 34 times larger than the binomial variation from 50,000 rolls.

All of this is standard statistical practice. The point of bringing it up here is that the t-test, which leads to the correct conclusion, is in fact a volume test. The centered, renormalized data vector $\mathbf{u} = (u_1, u_2, \cdots, u_n)$, where $u_i = (m_i - 16{,}666.7)/[\sum_{j=1}^{20} (m_j - 16{,}666.7)^2]^{1/2}$, is a point on $\mathcal{SP}_{20}$, the unit sphere in 20 dimensions. The set of vectors $\mathbf{U}$ lying as close or closer than $\mathbf{u}$ to the point $\mathbf{e} = (1, 1, \cdots, 1)/\sqrt{20}$ is a spherical cap on $\mathcal{SP}_{20}$, centered at $\mathbf{e}$. The significance

TABLE A

Number of occurrences of either 1 *or* 6 *in* 50,000 *rolls each of twenty dice; total number of occurrences is* 335,294 *out of* 1,000,000 *total rolls. Were the dice weighted?*

17755	16734	16769	16359	16661	16285	16309
17479	16529	16486	15668	16292	17511	17020
15929	16829	17665	17981	16677	16356	

level .246 for the t-test is the ratio of the 19-dimensional volumes of the cap compared to the entire surface of the sphere. See Efron (1969) for a picture, and Hotelling (1961) for an extensive discussion of the t-test as a volume test. Both authors show that the volume interpretation gives the t-test validity under more general assumptions than normality. Indeed, the geometric interpretation allows direct interpretation. The t-test example undoubtedly motivated Hotelling's interest in volume testing.

Several of the commentaries question the relevance of the volume test in the two-way table situation. In fact H_0, the hypothesis of uniformity leading to the volume test, is not particularly interesting in its own right. We have used H_0 as a *hypothesis of disinterest*, which is very much in the spirit of standard null hypothesis. (That is why we called it H_0). In Table 4 for example, the fact that, with sample size 160, the distributions of χ^2 under H_1 and H_0 are hopelessly overlapped shows the futility of standard χ^2 testing in this situation; we can't even distinguish H_1 from the uninteresting hypothesis H_0.

Why did we choose H_0, the uniform distribution, to represent the hypothesis of disinterest, rather than any other broadly dispersed distribution over $\mathscr{S}_{IJ}(n)$? There are four basic reasons: (1) Mathematical tractability. The significance level calculations for the volume test can be done reasonably accurately, in some ways more accurately than those for the usual χ^2 test. (Notice that our calculations are not asymptotic in the usual sense, and in particular make no use of the central limit theorem. The central limit theorem does appear in the reduced sample size considerations of Section 5, but only in a saddlepoint form, where it is applied at the center of the approximated distribution.) The uniform distribution on the sphere $\mathscr{S}\mathscr{P}_{20}$, which underlies the t-test, has many justifications, but it would certainly not be much used if it weren't mathematically tractable.

(2) Consonance with the usual χ^2 test. The usual χ^2 test measures how *far* the observed table $\mathbf{p}$ is from the independence surface, compared to a χ^2_D distribution. The volume test measures how *close* $\mathbf{p}$ is to the independence surface, compared to the uniform distribution. It is nice to have "close" and "far" defined in the same way, according to the natural Mahalanobis distance, and moreover for the curves of constant distance, e.g., the ellipsoid in Figure 2, to be isopleths of approximately constant density for both distributions. This last property, which is also approximately true of the exponential family considered in Sections 4 and 5, makes the Mahalanobis distance a sufficient statistic.

(3) Components of variance arguments. The exponential family of Sections 4 and 5 is a components of variance analogue applicable to two-way tables. As $\theta \to 0$ in that family—that is, as the component of variance $\sigma^2_\beta \to \infty$—the exponential family approaches the uniform distribution, at least for tables $\mathbf{p}$ near the independence surface. In this sense H_0 lies at one end of a one-parameter components of variance family, with the independence hypothesis H_1 at the other end.

(4) The Bayesian argument (2.11), (2.14).

Other of the commentaries raise a more fundamental objection to our paper: perhaps the simple type of analysis we propose, which does not examine the structure of the table, but only its Mahalanobis distance from the independence

surface, is misguided; *only* a structural analysis can yield interesting answers. The same objection can be raised against components of variance (random effects) analyses of standard linear models. In the casino data, for example, might we not do better with a careful linear model analysis which took into account the size, weight, and color of the individual dice?

The fact is that components of variance analyses are useful in standard applications precisely because they avoid the specification of detailed structural models. Structural analyses can be difficult and confusing in their own right, and when possible it is nice to answer simple questions, e.g., is the casino guilty or innocent, in a simple way.

The charms of simplicity become particularly obvious when one is faced with a large amount of work. In Table 2, for example, we can suppose that many more questions were asked of the Swedish families besides income and number of children. Perhaps each family answered a census form with 50 main questions. These days the Swedish statistician would probably receive a fat computer printout including all $\binom{50}{2} = 1225$ two-way tables.

It is wise to take a preliminary look at the data in this way before launching an ambitious structural analysis, but it is impossible to examine 1225 tables carefully. A simple summary statistic for each table, like the standard significance level of χ^2, helps sort them into categories for deeper investigation. Our paper proposes two other helpful summaries, the volume test significance level and the effective sample size, or equivalently $\hat{\sigma}_{\text{rel}}$, (4.18).

Even when the statistician intends to do a structural analysis of a particular table, it is nice to know how much structure there is to analyze. Our nonstructural analyses of Tables 1 and 2 show that Table 1 is far more nonindependent than Table 2, $\hat{\sigma}_{\text{rel}} = .26$ versus $\hat{\sigma}_{\text{rel}} = .0051$, so that there is a lot more to explain in the first case. (Here is a more familiar way to say the same thing: let $\hat{\boldsymbol{\pi}}_0$ be the table of constant probabilities, all equal to $1/16$ in Table 1. Then the Kullback-Leibler distance from the observed table to $\hat{\boldsymbol{\pi}}_0$ decomposes as $I(\mathbf{p}, \hat{\boldsymbol{\pi}}_0) = I(\mathbf{p}, \hat{\boldsymbol{\pi}}) + I(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\pi}}_0)$. The numerical values of the decomposition are $.383 = .124 + .254$ in Table 1, compared to $.570 = .011 + .559$ in Table 2. In other words, $I(\mathbf{p}, \hat{\boldsymbol{\pi}})/I(\mathbf{p}, \hat{\boldsymbol{\pi}}_0)$, the proportion of the observed table not explained by independence, is 32% in Table 1 and only 2% in Table 2.)

In fact, one might prefer to start a structural analysis of Table 1 in terms of departures from perfect dependence, rather than departures from perfect independence.

In summary, our paper tries to extend the benefits of standard random effects modelling to the two-way table situation. We are by no means the first to make such an effort, although we have employed somewhat different mathematical tools than our many predecessors. With these points in mind, the commentaries comprise a valuable catalog of other approaches and viewpoints. For the most part they avoid the sneer genre of statistical commentary ("our learned colleagues seem to have forgotten the correct formula for the normal distribution"), and concentrate on constructive alternative analyses. Here is a brief guide, critique, and response.

Professor Pierce. Writing from basically the same point of view as our paper,

Pierce raises an insightful question: perhaps the Mahalanobis distance, as illustrated in Figure 2, is the wrong metric for measuring overall discrepancies from independence. He perfers the metric which would be appropriate if the individual cell counts m_{ij} were independent binomials or Poissons, so that $\text{var}(m_{ij})$ was proportional to $\hat{\pi}_{ij}$ rather than to the ijth diagonal of the Fisher-Yates matrix.

Pierce is on strong ground in not automatically trusting a fully invariant approach, which essentially treats all linear combinations of the m_{ij} as being of equal interest, to properly model the original cell counts m_{ij}. The original counts are defined in terms of natural names, like "blue" and "two children," and perhaps this fact should be given more weight in the analysis. On the other hand, (i) the counts m_{ij} are *not* independent binomials, especially when conditioned on the marginals; (ii) the overdispersion differences shown in Pierce's Table 1 become a lot less dramatic when expressed in terms of standard deviation instead of variance; (iii) the differences become smaller still when I and J are bigger, in particular for $\min(I, J) > 2$.

Professors Breslow and Moore. Also writing from a components of variance viewpoint, Breslow and Moore start off in the same direction as Pierce, but wind up preferring a different model of overdispersion for the cell counts, based on Wedderburn-Nelder-McCullagh quasi-likelihood calculations. This is a nice approach, having both the virtues and defects of Pierce's method. There is no reason that their analysis, in terms of the parameter τ, should agree with ours in terms of σ_{rel}, since the two parameters measure overdispersion on different scales. Nevertheless it is disturbing that the Breslow-Moore method does not cleanly separate two situations which seem as different as Tables 1 and 2.

With Professor Nelder's considerable help, we were able to see the close connection of quasi-likelihood theory with the results in Section 5, as suggested by Breslow and Moore. The connection is not so much with the original Nelder-Wedderburn theory, as with the "extended quasi-likelihood" of Nelder and Pregibon, described briefly on page 212 of the McCullagh-Nelder monograph.

Consider (5.6), with dimension $D = 1$, sample size $n = 1$, so

$$\log f_\theta(x) = -\theta I(x, \beta_1) + \log \phi(\theta) + \log g_x(x).$$

Here $I(x, \beta_1) = T(x, \beta_1)/2$ is the Kullback-Leibler distance, and we have used (5.3). Now make the following approximations: $\log g_x(x) \doteq \frac{1}{2} \log[2\pi \text{ var}_x(X)]$, (central limit theorem) and $\log \phi(\theta) \doteq \frac{1}{2} \log \theta$, (5.7). Then the expression above becomes

$$\log f_\theta(x) = -\theta I(x, \beta_1) + \tfrac{1}{2} \log \theta - \tfrac{1}{2} \log[2\pi \text{ var}_x(X)],$$

which is exactly the extended quasi-likelihood family (11.2) of McCullagh-Nelder.

This close relationship may make our approach more palatable to those familiar with generalized linear models. It may also be of some interpretive value in using the Nelder-Pregibon theory. For example, if a logistic regression is fit by extended quasi-likelihood with $\hat{\theta} = \frac{1}{2}$, then the fitting is essentially the same as ordinary logistic regression, except with the sample size at each covariate value reduced by factor $\frac{1}{2}$.

We are grateful to Professors Dickey, Fienberg, Leonard, and Good for exploring the connections between our paper and a variety of Bayesian analyses. In

thinking about such things, we have found the following simple example instructive. Consider n tosses of a coin. The (5%) volume test for $p = \frac{1}{2}$ rejects if the number of heads does not fall among the central $n/20$ points.

Let us begin by comparing this to the usual (5%) frequentist test: reject if the number of heads does not fall among the central $2\sqrt{n}$ points. Here, asymptotics can be misleading, suggesting asymptotically huge acceptance regions for the volume test, $O(n)$ compared to $O(\sqrt{n})$. In fact, for $n = 99$, the volume test rejects outside the interval [48, 52] while the standard test rejects outside [40, 60]. The volume test acceptance region exceeds the frequentist region only for $n > 1600$(!).

One can give a naive Bayesian interpretation to the volume test by imagining a Bayesian with uniform prior. Then the central $n/20$ points have prior mass $\frac{1}{20}$. So this Bayesian is reasonably certain that the number of heads will fall outside the central interval.

Following Jeffreys, a modern Bayesian might formulate the testing problem as follows: put mass $P(0)$ on H_0: $\pi = \frac{1}{2}$, and mass $P(1)$ on H_1: $\pi \sim$ Uniform. Let a loss function be specified by $L(0, 0) = L(1, 1) = 0$, $L(0, 1), L(1, 0) > 0$; the Bayes rule rejects H_0 for

$$P(1 \mid x)/P(0 \mid x) > L(0, 1)/L(1, 0),$$

where $P(i \mid x) = P(i)P(x \mid i)/[P(0)P(x \mid 0) + P(1)P(x \mid 1)]$. When $P(0) = P(1)$ and $L(0, 1) = L(1, 0)$ the rejection region is

$$\frac{1}{n} > \binom{n}{j} \frac{1}{2^n}.$$

Asymptotically, this is not very different from the frequentist test, rejecting outside an interval of length proportional to $\sqrt{n(\log n)^2}$, (when $n = 99$ the Bayes test rejects outside [39, 61]).

When would a Bayesian use the volume test for the simple problem? In the framework above, it turns out that the volume test is approximately the Bayes test when either $P(0) \gg P(1)$ or $L(0, 1) \gg L(1, 0)$.

While the above asymptotics are suggestive, they also underscore the fact that there need not be a reasonable correspondence between observed significance levels and posterior probabilities. Berger and Sellke (1985) contains some striking further examples. The volume test is not intended as a substitute for a full Bayesian analysis, any more than as a substitute for a structural analysis.

Of course, a Bayesian analysis must also confront the same problems: with a huge sample size, small deviations from a null hypothesis will result in huge posterior odds. Can we meaningfully distinguish between varities in such situations? One could contemplate a Bayesian analysis which tests for "closeness." So far as we know, such tests have not been worked out.

Here are some more specific responses.

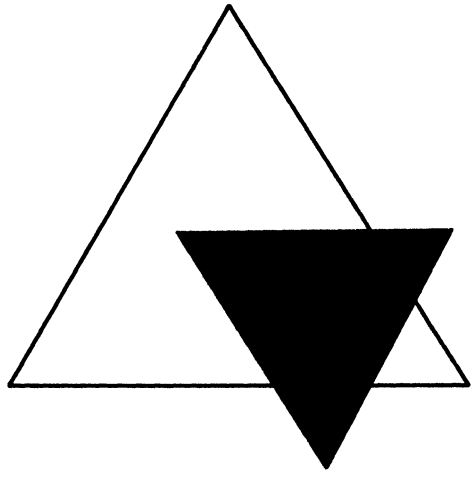
Professor Dickey. When the null hypothesis is not simple, one must put a prior on the surface of independence as well. We are grateful to Professor Dickey for demonstrating the difficulties this can cause, and presenting his way through the maze. We particularly appreciate his numerical results for the two tables. They underscore the fundamental problem: just like the classical analysis, the Bayesian

analysis rejects H_0, more strongly for Table 2. This misses the point that Table 2 is far closer to the null hypothesis than Table 1.

Professor Fienberg raises the following reasonable objection: Bayesian models of overdispersion, like Professor Leonard's, begin wtih a one-parameter family of prior distributions ξ_θ for π, then generate the observed table $\mathbf{p}$ in two steps, $\xi_\theta \to \pi \to \mathbf{p}$. The last step is by ordinary multinomial sampling as in (2.12). Our one-parameter family (5.6) goes directly from θ to the marginal density $f_\theta(\mathbf{p})$, without saying what the family of priors ξ_θ is.

In fact, we have not found priors ξ_θ which yield as margins the family $f_\theta(\mathbf{p})$, (5.6) (except for the uniform density $f_0(\mathbf{p})$ which comes from (2.11)). Instead we have followed a quasi-Bayesian strategy for interpreting $f_\theta(\mathbf{p})$: the effective sample size, originally defined as $\nu = n\theta$, is converted to a variance component σ_β^2 via definition (4.10), with Bayesian interpretation (4.15). This leads to statements like $\hat{\sigma}_\beta^2 = .0243$ or $\hat{\sigma}_{\text{rel}} = .26$, which estimate the dispersion of ξ_θ about the central point $\hat{\pi}$, without ever saying what ξ_θ is. As Breslow and Moore point out, this is quite similar in spirit to a quasi-likelihood approach, which makes use of convenient exponential family properties without displaying the exponential family itself.

We also think the mathematical results have application outside our analysis. Indeed, the set of points in the IJ simplex with prescribed margins is called a transportation polytype in the operations research literature. Bolker (1972, 1976) shows that this set is the intersection of I standard $J-1$ simplices, reciprocally oriented in J-dimensional space. For example, the set of 2×3 tables with prescribed margins is the intersection of 2 reciprocally oriented equilateral triangles. Such a region can have 3, 4, 5 or 6 vertices. Compare the drawing below with Figure 2.



For 2×4 tables we get the intersection of 2 tetrahedra. We have thus given a formula for the volume of such objects for $2 \times I$ and $3 \times I$.

We are currently working with Karel Reisz from Stanford's Operations Research Department trying to use OR techniques to find volumes for other cases.

Our approach involves computing a simplicial decomposition of the set of tables with given margins, and using standard formulas for the volume of a simplex. Even for 5×7 tables, a full simplicial decomposition is prohibitively time consuming. We are working at a "greedy" algorithm, approximating the region with large simplices.

Professor Leonard's 1977 paper introduces another components of variance model, which is similar but not identical to our model (5.6). In situation (5.15), where $x \sim \text{Bi}(20, \beta)/20$, Leonard induces overdisperison of x about .5 with the beta prior density $\beta^8(1-\beta)^8$. (The exponent 8 gives effective sample size $n/2 = 10$ in his equation (2.3).) The marginal density for x is numerically close to but not identical with our f_θ in (5.15).

There are, in fact, many technically different one-parameter models of overdispersion which lead to the asymptotic scaling property (4.12). There are four justifications for the particular choice (5.6):

(i) It exactly preserves the contours of equal likelihood ratio, as stated following (5.24).
(ii) It can be defined for general exponential families, not just the multinomial.
(iii) It is simple to calculate with, as in (5.7), using the central limit theorem approximations only in saddlepoint form, as in (5.10).
(iv) The effective sample size interpretation of θ, $\nu = n\theta$, can be interpreted in terms of the corresponding phenomena for standard linear models (4.1)–(4.10).

Probably none of these advantages make a great deal of difference in applications. What does, or at least could, make a difference is Pierce's point: models that scale as in (4.12) agree with the Fisher-Yates dispersion matrix, and not with the familiar models of cell-by-cell binomial variation.

Professor Leonard's Bayesian analysis of the Marine Corps data is pleasing and informative. The table has $n = 5698$, $\chi^2 = 1018.28$ (not 456.93 as stated), and $D = 77$. The unconditional volume test significance level is very small $1.3 \cdot 10^{-7}$, estimated effective sample size $\hat{\nu} = 431$, with 90% confidence interval $\nu \in [323, 551]$. This indicates departures from independence smaller than Table 1 but bigger than Table 2, $\sigma_{\text{rel}} \in [.061, .082]$. In other words, we might expect to find some interesting deviations from independence, but not of enormous magnitude, which is what Leonard's analysis nicely reveals. (Incidentally, 50% of the χ^2 statistic comes from school G, with 40% from the entry "169" alone. Why doesn't its residual from independence look significant in Leonard's last table?).

Professor Good. Good and Crook had to work hard and cleverly to produce priors which agree with Fisher's conditional inference. They were aiming for a general-purpose test. Examples like Table 2 suggest that such universal priors are a lot to ask for. The papers by Good, and Crook, are of considerable interest to both Bayesian and non-Bayesian readers.

Regarding the asymmetry of our approximation (7.6), we have tried a symmetrized version: the square root of the product of the two approximations. The version recommended seemed (very slightly) more accurate in the examples we tried. We recommend choosing the direction that involves a mixture (as in (7.2))

with the smoother of $\{r_i\}$, $\{c_j\}$. This allows an approximation of as nice a function as possible.

Professor Sundberg. We agree that Martin-Löf's redundancy statistic is a useful overall measure of departure from independence, and warmly recommend his 1974 paper to the reader. Redundancy's definition in terms of Shannon information leads to nice mathematical properties, but also to difficulties of interpretation for statisticians. For Tables 1 and 2 the redundancy measure equals .046 and .005, respectively, interpreted as the relative decrease in the number of binary units needed to specify the entries in the table when we take into account the regularities in the exact test.

Martin-Löf also offers a rough scale (his Table 5). On this scale, Table 1 falls between a very bad fit and bad fit while Table 2 falls between a bad fit and good fit. We agree with this, but only as a very rough summary.

For a 2×2 table, the slice $\mathscr{Y}(\mathbf{r}, \mathbf{c})$ in Figure 2 is a straight line segment. In this case σ_{rel}, (4.18), is just the random effects component of standard deviation divided by the length of the segment. If $\sigma_{\text{rel}} = .07$, for instance, then the true table π deviates from the point $\hat{\pi}$ on the segment by an expected root-mean-square amount 7% of its maximum range of deviation. Sundberg objects to σ_{rel} as unconvincing, but it certainly has geometry on its side.

Sundberg's objection to the effective sample size ν is based on an incorrect extension of ν to three-way tables. The normal theory motivation for ν, (4.1)–(4.10), does not depend on the dimension D, and neither would its proper extension to three-way tables. Such an extension is always possible according to the theory of Section 5, but, to answer Professor Fienberg, we have not attempted it. In any case, $\hat{\sigma}_{\text{rel}}$ is preferred to $\hat{\nu}$ as a descriptive statistic because of its easier geometric interpretation.

(We are grateful to Professor Sundberg for catching an error in our original manuscript.)

Professor Goodman gives the best argument for structural models, a successful analysis for each of our two tables. How good is the fit obtained? Consider his model H′ for Table 1. In terms of our Figure 2, the Mahalanobis squared distance from $\mathbf{p}$ to $\hat{\pi}$ in the q-dimensional space, $\mathscr{Y}(\mathbf{r}, \mathbf{c})$ is $138.29/n$, $n = 592$. If $\hat{\pi}'$ represents the table fitted under the two-parameter model H′, its Mahalanobis squared distance from $\hat{\pi}$ is $(138.29 - 10.48)/n$.

If H′ is a fixed two-dimensional linear subspace of $\mathscr{Y}(\mathbf{r}, \mathbf{c})$, passing through $\hat{\pi}$, then the correct test statistic for goodness of fit is

$$F = \frac{(138.29 - 10.48)/2}{10.48/7} = 42.7,$$

which is compared with a standard $F_{2,7}$ distribution. This comparison is a direct analogue of the t-test in the casino example. Having once decided that $\mathbf{p}$ is too far away from $\hat{\pi}$ to have arisen from multinomial sampling (2.12), it is necessary to compare the amount of explanation, $138.29 - 10.48$, with the amount to be explained, 138.29, rather than making theoretical comparisons based on (2.12). Another way to say the same thing is that the comparison can be made in terms of (2.12), but with n replaced by the effective sample size ν.

In the case at hand, $F = 42.7$ is highly significant compared to the .05 critical value $F_{2,7}^{(.05)} = 4.7$.

Professor Plackett. Most of the calculations in our paper were done on a hand calculator too! The methods we propose are not computationally difficult or sophisticated. Plackett's rough and ready structural analysis of Table 2 is fine, but what about Table 1, or Leonard's Marine Corps data, or any other table that might arise? Plackett's argument works equally well against any systematic method of analysis, not just our paper.

Drs. McCullagh and Pregibon. We are sorry that McCullagh and Pregibon, who have a lot to say about this problem, have chosen to be so contentious here (which is especially strange given the close connection with their own work, see the Breslow-Moore remarks above). Besides some legitimate differences of opinion, of the type discussed earlier, there are several outright misstatements of fact: $\hat{v}$ does *not* depend on the sample size; sample size does *not* necessarily decrease when there are many tables to consider (think of the Swedish statistician); the most powerful test of uniformity does *not* depend only on the marginal totals; our calculations are *not* highly asymptotic. This type of careless commentary can be accepted in praise, but is unforgiveable in criticism.

In summary, the usual chi-square test for $I \times J$ tables seems easy to misuse and misinterpret. Over the past 10 years a number of now highly developed approaches have evolved. These often have considerable overlap, and no one (including the present authors) seems to have them all in focus. We have learned a lot from the commentaries. In particular, we leave the discussion with a healthy respect for the simple direct interpretation of Hotelling's volume test.

Finally, our thanks to the Editor for his considerable efforts in assembling this discussion.

REFERENCES

BERGER, J. and SELLKE, T. (1985). Testing a point null hypothesis: the irreconcilability of significance levels and evidence. Technical Report #84-27, Dept. of Statistics, Purdue University.

BOLKER, E. (1972). Transportation polytopes. *Jour. Combin. Theory Ser. B* **13** 251–262.

BOLKER, E. (1976). Simplical geometry and transportation polytopes. *Trans. Amer. Math. Soc.* **217** 121–142.

EFRON, B. (1969). Student's t-test under symmetry conditions. *J. Amer. Statist. Assoc.* **64** 1278–1302.

HOTELLING, H. (1961). The behavior of some standard statistical tests under non-standard conditions. *Proc. Fourth Berkeley Symp.* **1** 319–360.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305