Rejoinder

Alan Agresti

I thank the discussants for their comments. I appreciate their compliments, as well as their criticisms and suggestions, which nicely supplement the presentation in my paper. It is reassuring to see that, despite what some might call the excessive length of my paper, alternative perspectives raise yet other noteworthy issues and cast additional light on this subject.

In my article, I mentioned a need for the development of exact methods for model checking, and I am pleased to see the contribution by Ed Bedrick and Joe Hill on this topic. They suggest a simple algorithm for generating the relevant reference set for logistic regression. Their algorithm would seem to generalize to loglinear models.

Regarding the problem noted by Bedrick and Hill of potential near degeneracy in using conditional methods, I am afraid I do not see any simple solutions (other than perhaps becoming a Bayesian). Degeneracy is the most extreme form of the severe discreteness that can occur with conditional methods. The severe discreteness is the primary weakness of this type of method and is, I believe, at the heart of the objection many statisticians have with methods such as Fisher's exact test. An approximate solution in logistic regression is to slightly collapse the data in order to produce a fuller conditional distribution of the data.

I thank Bedrick and Hill for clarifying and extending my remark about exact analysis for parameters when the link function in a generalized linear model is noncanonical. They show that some types of conditional inference may still be useful, both for model checking and inference about a parameter of interest. I would like to see them develop this discussion further for some interesting noncanonical models for categorical data. Likewise, I would like to see further discussion of their view treating Fisher's exact test only as a goodness-of-fit test. This is subtly distinct from the usual view of its also serving as a test comparing two independent binomials. Perhaps they can help to clarify this longstanding controversy, although I do not expect to see statisticians reach agreement about how to analyze 2×2 tables, at least not in my lifetime.

I think that both of Diane Duffy's ideas merit considerable attention. We statisticians commonly complain that users of statistics pay too much attention to p-values and statistical tests at the expense of more informative types of analysis. Perhaps we can at least convince them to perform sensitivity analyses, such as the ones Duffy recommends, so that they do not take their p-values too literally, feeling compelled to report them to several decimal places.

When $n_{1+} = n_{2+}$, deleting an observation from one row has the same impact on the magnitude of the one-sided p-value (but in the opposite direction) as deleting the same type of observation from the other row. Thus, in this balanced case, the observed p-value falls in the middle of the interval (P_L, P_U) for Duffy's first type of perturbation. For instance, for table $(10, 90/20, 80)$, the one-sided p-value is 0.0367, with $P_L = 0.0231$ and $P_U = 0.0503$. As shown by her example, this need not happen for the usual two-sided p-value, for which deletion of an observation sometimes has no effect on the p-value.

Duffy suggests studying whether algorithms for exact analyses can be adapted to aid Bayesian computations. She also proposes a conditional Bayesian analysis and suggests comparing it to ordinary Bayesian and frequentist procedures. The recent significant improvements in computational tools (e.g., Gibbs sampling) for Bayesian methods suggests that we should soon see Bayesian and hybrid methods more fully developed for multidimensional contingency tables. For sparse contingency tables, several problems exist for which unconditional frequentist approaches fail and for which a Bayesian approach would be a natural alternative to a conditional approach. An example is the analysis of $2 \times 2 \times K$ tables with large K. When the true odds ratio is identical in each 2×2 table, its unconditional ML estimator is inconsistent when K grows at the same rate as n, such as when each table consists of a case-control matched pair (e.g., Breslow and Day, 1980, page 250). When the true odds ratios are not identical in each table, a Bayesian or empirical Bayesian approach would seem to be a reasonable way of smoothing the stratum-specific estimators of odds ratios, borrowing from the whole to get estimators with improved MSE properties.

Leonardo Epstein and Stephen Fienberg also suggest the worthiness of a Bayesian perspective, pointing out that it may be no more difficult computationally than exact conditional methods. It is interesting to note that, before much development of methodology for multiway contingency tables had taken place, Lindley (1964) argued that Bayesian methods had the advantage (compared to

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science.* www.jstor.org

frequentist methods) of extending to such tables with fewer complications.

Epstein and Fienberg claim that, even for sparse tables, asymptotic methods are often well behaved for certain model-based inferences. I agree with this comment, and my experience also indicates that Haberman's (1977) results are particularly important (Agresti and Yang, 1987). Yet, cases will always exist in which the use of such asymptotics is questionable, such as when a table contains both very large and small counts. Exact methods can help us to highlight situations in which asymptotic answers may be unreliable.

Many of the Epstein and Fienberg criticisms about p-values for exact tests are applicable to any type of statistical test, and I do not disagree with their main thrust. Even though I am interested in exact tests for contingency tables, I realize the limitations of tests in any statistical analysis. I think statisticians generally are in agreement about this, but many practitioners are unaware of such basic considerations as the dependence of results of tests on the sample size and the distinction between statistical significance and practical significance. Because of this, I have taken pains in my writings for nonstatisticians to point out the limitations (e.g., Agresti and Finlay 1986, pages 152–153). Nonetheless, the ubiquitous use of statistical tests is likely to continue, particularly for such common problems as judging whether there is sufficient evidence to distinguish between effects of two treatments.

So, given the obvious limitations, why would one ever want to use a test (exact or otherwise) to check the fit of a loglinear model? I believe the following justification makes sense. We realize that the model being tested is an approximation for reality and does not hold perfectly in the population of interest. In this sense, we know before conducting the test that the null hypothesis is false. Whether the test statistic is "large" may simply depend on how much data we have. Yet, we also know the benefits that accrue from using parsimonious models. For example, unless the sample is extremely large, we obtain better estimators of parameters of interest by using a simple, decently fitting model rather than a more complex model or the saturated model (Bishop, Fienberg and Holland, 1975, Section 9.2; Altham, 1984; Agresti, 1990, Section 6.4.4). Testing goodness of fit for models in a nested set gives us some indication of how much simplification we can reasonably apply in our attempt to obtain such improved estimators. I am curious about how much technical justification can be given for such use of tests.

Epstein and Fienberg claim that exact tests using the efficient score statistic only approximate analogous tests using the likelihood-ratio statistic. Their argument is correct, but there is no reason the efficient score statistic is not itself a valid statistic for measuring departure from the null hypothesis. My paper utilized this statistic because exact tests are then simpler to conduct for models requiring iteration. For each table in the conditional reference set, one can compute the efficient score statistic without needing to fit the model. Of course, if one wants to assume the extra computational burden of fitting the model for each such table, then one could also formulate an exact test using the likelihood-ratio statistic. This is not yet done for indirect models in any commercial software, but in principle it provides no difficulty.

Svend Kreiner has amplified my remarks about tests for conditional independence for higher dimensional arrays often being feasible by exploiting connections with equivalent tests for collapsed tables. As he points out, expressing some models as graphical models is useful for giving insight about when this can be done. I also very much like Kreiner's idea of simulating exact distributions in high dimensions by a simpler sequential approach. I disagree with his comment that there seems to be no practical solution to tests of higher order interactions in the predictable future. Zelen's (1972) influential work was an early start in this direction and is available in StatXact. Also, Morgan and Blumenstein (1991) recently described a simple general algorithm for tests comparing hierarchical loglinear models that can be applied for such a purpose.

I do not share Kreiner's concern about tests involving models fitted by iteration. As previously discussed, for tests using the efficient score statistic, one need only calculate the relevant sufficient statistic for each table in the conditional reference set rather than fit the model. An example already considered is the exact test for the linear-by-linear association model, which is fitted iteratively (Agresti, Mehta and Patel, 1990). Regarding model search strategies, restriction to decomposable models is helpful in many cases, but too severe for applications in which we expect all pairs of variables to be associated. Finally, I imagine that Kreiner agrees that significance tests, exact or asymptotic, should form only part of any model search strategy. As in any statistical setting, we learn less from formal tests than from parameter estimation (in particular, confidence intervals) and model diagnostics.

My paper did not provide much detail on computational algorithms for exact inference, and indeed my knowledge of them is quite limited. The discus-

sion by Cyrus Mehta, one of the foremost experts on this topic, is an informative supplement. He describes the interdisciplinary approach that is necessary to perform the computations in powerful software for exact inference, such as provided by his outstanding package, StatXact. My paper mentions that Mehta has made a long series of important contributions to exact conditional inference, and I fully expect that many of the problems that I posed will be solved by him and his students, and will be computable by the year 2000 version of StatXact! The current version of the StatXact manual is, in fact, a very good source for further discussion, as well as promotion, of the exact conditional point of view.

Samy Suissa's discussion provides arguments for the exact *unconditional* approach for comparing binomial proportions in 2×2 contingency tables. He and Professor Jon Shuster have made impressive progress in constructing algorithms for exact unconditional inference, both for independent samples and matched pairs. Also, Soms (1989a, b) recently gave a program for exact unconditional confidence intervals for differences of proportions. Under the binomial sampling assumption, the unconditional approach is reasonable if one does not mind averaging results for the observed sufficient statistic with results for quite different possible outcomes. The sufficient marginal counts determine the precision with which comparisons can be made. Some world-reknowned statisticians have argued against unconditional averaging (for example, Fisher 1945; Cox and Hinkley, 1974, page 38; Yates, 1984; Cormack and Mantel, 1991), but the issue is clouded by the marginal counts not being ancillary and by practical implications of the extra discreteness occurring with the conditional approach.

L. J. Wei and D. Y. Lin, in analyzing 2×2 tables based on adaptive group sequential designs, also mention conservativeness problems with exact conditional methods. Their research, like that of Suissa and Shuster, indicates superiority for exact unconditional methods. Their point is well taken that many practitioners are more comfortable interpreting a difference or ratio of probabilities rather than an odds ratio. The analyses and the type of application in their work are noteworthy, and I regret that I did not cite them in my paper. In medical studies, the choice of data-analytic method, as well as the choice of design, has ethical implications and can have substantive practical consequences. The work by Lin, Wei and their co-authors is also admirable for giving strong attention to problems of interval estimation.

The conservativeness of the exact conditional approach is compounded for the randomized play-the-winner design discussed by Wei, Smythe, Lin and Parks (1990) by especially severe discreteness. Not only are both margins fixed, but also one of the row totals may be extremely small. In the unconditional approach for this design, neither margin is fixed. Wei, Smythe, Lin and Parks (1990) showed that even the exact conditional test using randomization to achieve the desired size performs poorly for this design.

The exact conditional approach has been strongly attacked by many statisticians for its conservativeness with 2×2 tables. Indeed, it can be quite conservative, and there are settings such as the one Wei and Lin discuss where its performance is inadequate. But one should also keep in mind that conservativeness problems for exact conditional analyses diminish with larger tables. As Epstein and Fienberg point out, in practice, larger tables are the norm; sole 2×2 tables cannot adequately address issues that are more naturally expressed in multivariate terms. One can also diminish conservativeness in some cases by using a test statistic that can assume a greater number of values. For example, in the exact test for the linear-by-linear association model, the conditional distribution is much less discrete when the scores $\{x_i\}$ and $\{y_j\}$ are unequally spaced rather than equally spaced.

For most cases in which conservativeness is an issue, I find recent arguments by Barnard and others supporting the mid-p approach to tests and confidence intervals quite persuasive. For the standard independent binomials design, I will be surprised if the exact unconditional test for 2×2 tables maintains its consistent power and sample size superiority when compared with the modification of the exact conditional test using the mid-p value. In this regard, Routledge (1990) used an Edgeworth expansion to show that at least for large n, the conditional mid-p value and unconditional p-value tend to be closer to each other than each is to the regular conditional p-value. Also, as I mentioned in my paper, I will be surprised if the unconditional test does not suffer from its own conservativeness problems for larger tables with greater numbers of nuisance parameters (e.g., independent *multinomial* samples). Indeed it is a tremendous computational challenge to perform such generalizations of the exact unconditional test, and I look forward to seeing whether Professors Suissa, Shuster, Wei, Lin or others can prove me wrong.

Exact conditional mid-p tests are easily invertible to confidence intervals and generalizable to inference problems for larger tables. Of course, one sacrifices "exactness" with the mid-p approach, but results I have seen so far are very encouraging.

Mid-p inferences tend to behave better than asymptotic methods, particularly in unbalanced cases for small n, yet do not display the conservativeness of the exact conditional and unconditional methods. Consider the choice between (1) a slightly approximate method in which intervals are constructed such that, in the long run, quite close to $p\%$ of them contain the parameter of interest, and (2) "exact" methods in which wider intervals are constructed such that, in the long run, at least $p\%$ contain the parameter of interest. When the intervals obtained by method (2) tend to be much wider and may have actual confidence level considerably higher than the desired $p\%$, my belief is that most practitioners will opt for method (1).

Dupont (1986) gave similar remarks to Suissa's about-potential anomalies in Fisher's exact test with two-sided p-values based on hypergeometric probabilities, caused by the skewness and discreteness. As Duffy notes in her discussion, this is an issue for all tests using highly discrete distributions, not just Fisher's exact test. Yates and discussants (1984) provided an interesting commentary on this issue. In my experience, anomalous behavior is less common for mid-p versions of p-values. It also seems to be less common for the Fisher and Yates recommendation of obtaining two-sided p-values by doubling the single-tail p-value. For instance, the two-sided p-value obtained by doubling the minimum one-tail p-value is 0.114 for table $(0, 35/4, 31)$ and 0.108 for $(0, 36/4, 31)$; it is 0.062 for $(9, 163/2, 169)$ and 0.060 for $(9, 162/2, 170)$. However, this type of two-sided p-value has its own disadvantages, including lack of interpretation, lack of a natural generalization for $I \times J$ tables, greater conservativeness than the ordinary p-value and a null expected value for the corresponding mid-p value unequal to 0.5.

In a personal communication to me, Dr. Duffy has indicated that she expects confidence intervals to be more robust than p-values to the types of perturbations she discusses. This is also my experience and is another reason for preferring interval estimation over tests. For instance, the usual exact conditional 95% confidence interval for the odds ratio is $(0, 1.07)$ for table $(0, 35/4, 31)$ and $(0, 1.04)$ for $(0, 36/4, 31)$; it is $(0.94, 44.85)$ for table $(9, 163/2, 169)$ and $(0.95, 45.39)$ for $(9, 162/2, 170)$.

In summary, I believe the discussants' comments illustrate that each type of "exact" or approximate inferential method for contingency tables has its strengths and weaknesses. Statisticians are unlikely to reach agreement on judging one type of method as superior to others for all purposes. Currently, the exact conditional approach does have the important advantages of wide scope and increasingly feasible computability.

I thank the discussants for their insights, and I look forward to new work on this topic by them and other statisticians.

ADDITIONAL REFERENCES

AGRESTI, A. and FINLAY, B. (1986). *Statistical Methods for the Social Sciences*, 2nd ed. Macmillan, New York.

AGRESTI, A. and YANG, M. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Comput. Statist. Data Anal.* **5** 9-21.

ALBERT, J. H. (1987). Empirical Bayes estimation in contingency tables. *Comm. Statist. Theory Methods* **16** 2459-2485.

ALBERT, J. H. and GUPTA, A. K. (1982). Mixtures of Dirichlet distributions and estimation in contingency tables. *Ann. Statist.* **10** 1261-1268.

ALBERT, J. H. and GUPTA, A. K. (1983a). Bayesian estimation methods for 2×2 contingency tables using mixtures of Dirichlet distributions. *J. Amer. Statist. Assoc.* **78** 708-717.

ALBERT, J. H. and GUPTA, A. K. (1983b). Estimation in contingency tables using prior information. *J. Roy. Statist. Soc. Ser. B* **45** 60-69.

ALTHAM, P. M. E. (1984). Improving the precision of estimation by fitting a model. *J. Roy. Statist. Soc. Ser. B* **46** 118-119.

ASMUSSEN, S. and EDWARDS, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70** 567-578.

BADSBERG, J. H. (1991). *COCO—A Program for Analysis of Complete Contingency Tables*. Inst. Electronic Systems, Aalborg Univ., Denmark.

BARNDORFF-NIELSEN, O. (1976). Nonformation. *Biometrika* **63** 567-571.

BEDRICK, E. J. and HILL, J. R. (1991). Model checking for logistic regression: A conditional approach. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. (E. M. Keramidas, ed.) 208-214. Interface Foundation, Fairfax, VA.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BERGER, J. O. and SELLKE, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.* **82** 112-122.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

CORMACK, R. S. (1986). The meaning of probability in relation to Fisher's exact test. *Metron* **44** 1-30.

CORMACK, R. S. and MANTEL, N. (1991). Fisher's exact test: The marginal totals as seen from two different angles. *The Statistician* **40** 27-34.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London. (Reprinted with corrections in 1982.)

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.

EDWARDS, D. (1984). A computer intensive approach to the analysis of sparse multidimensional contingency tables. In *COMPSTAT 1984. Proceedings in Computational Statistics, 6th Symposium Held at Prague 1984* 355-359. Physica, Heidelberg.

EPSTEIN, L. D. and FIENBERG, S. E. (1991). Bayesian estimation in multidimensional contingency tables. *Bayesian Inference in Statistics and Econometrics: Proceedings of the Indo-US Workshop, 1988. Lecture Notes in Statist.* Springer, New York.

FINNEY, D. J. (1947). *Probit Analysis*, 1st ed. Cambridge Univ. Press.

GOOD, I. J. (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Roy. Statist. Soc. Ser. B* **29** 399-431.

GOOD, I. J. (1975). The Bayes factor against equiprobability of a multinomial population assuming a symmetric Dirichlet prior. *Ann. Statist.* **3** 246-250.

GOOD, I. J. (1990). On the exact distribution of Pearson's χ^2 for the lady tasting beer. *J. Statist. Comput. Simulation* **36** 177-179.

GOOD, I. J. and CROOK, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69** 711-720.

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

HINKLEY, D. V. (1980). Likelihood. *Canad. J. Statist.* **8** 151-163.

HOROWITZ, E. and SAHNI, S. (1983). *Fundamentals of Data Structures.* Computer Science Press, Rockville, Md.

JOE, H. (1988). Extreme probabilities for contingency tables under row and column independence with applications to Fisher's exact test. *Comm. Statist. Theory Methods* **17** 3677-3685.

KREINER, S. (1989). *User Guide to DIGRAM—A Program for Discrete Graphical Modeling.* Statistical Research Unit, Univ. Copenhagen.

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

LIN, D. Y. and WEI, L. J. (1990). Comment on "Investigating therapies of potentially great benefit: ECMO" by J. H. Ware. *Statist. Sci.* **4** 324-325.

LIN, D. Y., WEI, L. J. and DEMETS, D. L. (1991). Exact statistical inference for group sequential trials. *Biometrics* **47** 1399-1408.

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

MEHTA, C. R., PATEL, N. R. and SENCHAUDHURI, P. (1992). Exact stratified linear rank tests for ordered categorical and binary data. *Journal of Computational and Graphical Statistics.* To appear.

MORGAN, W. M. and BLUMENSTEIN, B. A. (1991). Exact conditional tests for hierarchical models in multidimensional contingency tables. *J. Roy. Statist. Soc. Ser. C* **40** 435-442.

NAZARET, W. A. (1987). Bayesian log linear estimates for three-way contingency tables. *Biometrika* **74** 401-410.

PLACKETT, R. L. (1977). The marginal totals of a 2×2 table. *Biometrika* **64** 37-42.

PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9** 705-724.

RAFTERY, A. E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *J. Roy. Statist. Soc. Ser. B* **48** 249-250.

ROYALL, R. M. (1991). Evidential interpretation of statistical tests. Technical Report 737, Dept. Biostatistics, Johns Hopkins Univ.

SHUSTER, J. and SUISSA, S. (1990). Conditional versus unconditional tests in 2×2 trials. Invited paper. *Joint Statistical Meetings of the American Statistical Association*, Anaheim, Calif.

SOMS, A. P. (1989a). Exact confidence intervals, based on the Z statistic, for the difference between two proportions. *Comm. Statist. Simulation Comput.* **18** 1325-1341.

SOMS, A. P. (1989b). Some recent results for exact confidence intervals for the difference between two proportions. *Comm. Statist. Simulation Comput.* **18** 1343-1357.

SPIEGELHALTER, D. J. and SMITH, A. F. M. (1982). Bayes factors for linear and log-linear model with vague prior information. *J. Roy. Statist. Soc. Ser. B* **44** 377-387.

STREITBERG, B. and ROHMEL, R. (1986). Exact distributions for permutation and rank tests. *Statistical Software Newsletter* **12** 10-17.

THOMAS, D. G. and GART, J. J. (1978). Corrigenda. *J. Amer. Statist. Assoc.* **73** 233.

WARE, J. H. (1990). Investigating therapies of potentially great benefit: ECMO (with discussions). *Statist. Sci.* **4** 298-340.

WEI, L. J. and DURHAM, S. (1978). The randomized play-the-winner rule in medical trials. *J. Amer. Statist. Assoc.* **73** 830-843.

WEI, L. J., SMYTHE, R. T., LIN, D. Y. and PARK, T. S. (1990). Statistical inference with data-dependent treatment allocation rules. *J. Amer. Statist. Assoc.* **85** 156-162.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, New York.

WILLIAMS, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *J. Roy. Statist. Soc. Ser. C* **36** 181-191.