1990. Biometrics 46:267-268

CORRESPONDENCE

Alternatives to Fisher's "Exact Test" for Analyzing 2 × 2 Tables with Small Cell Sizes

From: *Richard Engeman*
Denver Wildlife Research Center
Building 16, Denver Federal Center
P.O. Box 25266
Denver, Colorado 80225-0266, U.S.A.
and
George D. Swanson
Department of Anesthesiology
Box B110
University of Colorado School of Medicine
Denver, Colorado 80262, U.S.A.

To the Editor of Biometrics:

Rice (1988a) presents an interesting test for analyzing 2 × 2 contingency tables in the smaller sample size situations. His motivation was to provide an alternative to Fisher's "exact test" and his rationale was based on the application of a prior distribution to the probability of a success, θ, under the null hypothesis. We would like to briefly comment on these aspects of his paper and also discuss another alternative that was proposed a number of years ago.

The statistical literature seems to contain two philosophical camps concerning the use of Fisher's "exact test." One group contends that the test is too conservative and the other believes it to be an appropriate test. This is demonstrated by Rice's paper and the accompanying discussions by Hill (1988) and Barnard (1988). By assuming one theoretical point or another, the two camps can argue the respective philosophical merits of using Fisher's test. For the applied statistician, the ultimate concern is what works "best" in practice. In most data analyses, "best" might refer to the test most likely to correctly indicate a difference.

The properties of various estimation methods seem to be best demonstrated and compared through simulation studies. One could devise many individual examples where Fisher's test produces less conservative results (smaller *P*-value) than a particular competing test. However, Fisher's test generally yields conservative results. Similarly, it has been our experience that Fisher's test produces consistently conservative results, frequently counterintuitive to what would be expected from examination of the data. Thus, we welcome the introduction of potential alternatives.

The test introduced by Rice is conditional, either based on a prior distribution of the values θ could assume under the null hypothesis when no other complete information is available, or based on more specific values when more reliable prior information exists. When describing his probability model, he words his rationale for the prior distribution very carefully, but still finds it a point of contention in his response to the discussants (Rice, 1988b). We find it interesting that statistics texts frequently describe the assumptions under which Fisher's exact test is valid, but go on to recommend its general use for small-cell-size situations. We find Rice's approach reasonable, but the justification feels to us somewhat like "philosophical tightroping."

We have been making use of another test for 2 × 2 tables with small cell sizes that was developed by McDonald, Davis, and Milliken (1977). This unconditional test was also developed in part to provide an alternative to Fisher's test, but seems easier to accept philosophically than Rice's test. More extensive tables for McDonald's test are given in McDonald and Milliken (1975). We have found this test in practice to be far less conservative than Fisher's. We speculate that it leads to results similar to Rice's test, perhaps slightly more conservative. It is far less conservative than Fisher's test. Using Rice's hawk-owl example data yields a one-tailed *P*-value of .022 for McDonald's test versus .016 for Rice's test.

The test by McDonald et al. seems to be frequently referenced in papers on analyzing 2 × 2 tables; however, we have no idea how extensively it is used. It is easy to apply because it has published tables up to cell sizes where Pearson's chi-square can be used. Because it is tabulated, one need not input a program (nor acquire a PASCAL compiler). It was published in a widely read statistics journal in addition to the original university technical report. Therefore, it has potential for extensive use.

We believe that it would be of general interest to receive Rice's perception of the unconditional test by McDonald et al. (or of other tests tabulated for small cell sizes). Of even greater interest (and effort) would be a simulation comparing the Rice and McDonald tests (Fisher's should be included for completeness). Those of us who regularly must analyze 2 × 2 tables with small cell sizes would welcome such information.

References

Barnard, G. A. (1988). Discussion on the paper by William R. Rice. *Biometrics* **44**, 16–18.

Hill, I. D. (1988). Discussion on the paper by William R. Rice. *Biometrics* **44**, 14–16.

McDonald, L. L., Davis, B. M., and Milliken, G. A. (1977). A nonrandomized unconditional test for comparing two proportions in 2 × 2 contingency tables. *Technometrics* **19**, 145–157.

McDonald, L. L. and Milliken, G. A. (1975). A nonrandomized test for comparing two proportions. *College of Commerce and Industry Research Paper 94*. Laramie, Wyoming: University of Wyoming.

Rice, W. R. (1988a). A new probability model for determining exact *P*-values for 2 × 2 contingency tables when comparing binomial proportions. *Biometrics* **44**, 1–14.

Rice, W. R. (1988b). Author's reply. *Biometrics* **44**, 18–22.

The author replied as follows:

Engeman and Swanson have brought up several points concerning the conditional binomial exact test (CBET) procedure (Rice, 1988) for comparing binomial proportions, and have asked that I respond. Their major point is that an alternative testing procedure, which was developed by McDonald, Davis, and Milliken (1977), may have similar power yet require less "philosophical tightroping." My overall response is that I find the test proposed by McDonald et al. to be quite useful, but I also find it to have important limitations.

First I consider the computer simulation analysis requested by Engeman and Swanson. I have examined many of the published tables of McDonald et al. and found the CBET procedure to reject in virtually all cases where the procedure of McDonald et al. rejects, yet I also commonly found, in the case of very small sample sizes, instances where only the CBET would reject. It therefore seems to me that a computer simulation study is unnecessary since it would almost certainly indicate that the CBET has more power than the conservative procedure of McDonald et al., but not much more unless one or both sample sizes are quite small, i.e., when the procedure of McDonald et al. cannot reject while the CBET can, or when their procedure can reject for only the most extreme difference in sample proportions.

The issue of "philosophical tightroping," presuming that I truly know what this means, is a less tangible matter. Once one accepts the fact that an exact Neyman–Pearson-type *P*-value is impossible, an alternative statistical benchmark must be sought. The Fisher/Yates approach is to carry out a permutation test which yields a conceptually different kind of *P*-value that is conservative relative to the criterion of repeated sampling. The approach of the many "corrections" to Pearson's chi-square test (not including Yates' correction) is to determine a conservative upper bound for the Neyman–Pearson *P*-value. The approach of the CBET is to *presume* all values of the unknown probability of success (θ) to be equally likely until data provide evidence to the contrary, and then calculate an exact, nonconservative *P*-value. The rationale for such a *P*-value is motivated by empirical objectivity and is therefore not dependent on properly estimating an unknown prior distribution. It also has the desirable property of representing a weighted average Neyman–Pearson *P*-value, with weights equalling the likelihood of θ given the sample data.

The McDonald et al. approach bypasses the *P*-value calculation altogether and instead determines a rejection region that produces a conservative test relative to the .05 and .01 levels of significance. This seems to me to be a reasonable solution but it also imposes three important limitations. First, rejection at the .05 level is precluded for very small tables where it is possible for the CBET. This is not surprising since the procedure of McDonald et al. is conservative. Second, *P*-values per se are not accessible, except in special cases, since only a rejection region is specified relative to two prescribed levels of significance. While more extensive tabulation would partially eliminate this problem, it does not seem practical due to the large range of cases for which tables are needed (a voluminous book of tables would be required; see below). Any testing procedure that does not provide for *P*-values also eliminates the potential use of combined probability tests (Folks, 1984), which are increasingly being used in many areas of science.

A third limitation with the approach of McDonald et al. is that their test cannot be used (due to the unavailability of tables) in those cases where sample sizes are large but the number of successes is small. For example, suppose mortality data indicate that 6 of 400 randomly marked male animals

die from a natural epidemic, while only 1 of 500 such females die. Is there evidence for sex-specific mortality? Pearson's chi-square test is invalid here due to the small expected numbers of dying animals. Fisher's test is not consistent with the criterion of repeated sampling, tables for the procedure of McDonald et al. are unavailable, yet the two-tailed CBET *P*-value is readily calculated to be .048. This example demonstrates why I believe that Engeman and Swanson's claim that there are "published tables up to cell sizes where Pearson's chi-square test can be used" to be an overstatement.

As a minor point Engeman and Swanson indicate that there are no published tables for the CBET and therefore a PASCAL compiler is required. While this may appear to be the case, it is not since I have offered both compiled and ASCII versions of the program to all of those requesting a copy of the CBET program. Obviously a table for the CBET could easily have been produced, but I find tables to be an antiquated and inefficient means of providing information regarding *P*-values. I chose to provide a computer program instead due to the widespread availability of microcomputers.

Finally, if a practitioner desires a Neyman–Pearson-type of *P*-value then it appears to me that only two alternatives are possible: (1) a conservative upper bound for the true Neyman–Pearson *P*-value, or (2) an alternative nonconservative *P*-value which is philosophically aligned with the criterion of repeated sampling. The conservative approach is "safe" yet "wasteful of sampling effort" since it is virtually guaranteed to yield too large a *P*-value. The *P*-value from the CBET seems to me to be a reasonable compromise between the safe and wasteful attributes of the conservative approach. Unlike the *P*-value from a permutation test, the CBET *P*-value is readily interpretable with respect to the criterion of repeated sampling, yet not conservative. So in summary, I respond to Engeman and Swanson's "philosophical tightroping" statement by saying that I agree and consider those choosing the CBET approach to be philosophically "well balanced."

REFERENCES

Folks, J. L. (1984). Combination of independent tests. In *Handbook of Statistics 4, Nonparametric Methods*, P. R. Krishnaiah and P. K. Sen (eds). New York: North-Holland.

McDonald, L. L., Davis, B. M., and Milliken, G. A. (1977). A nonrandomized unconditional test for comparing two proportions in 2 × 2 contingency tables. *Technometrics* **19**, 145–157.

Rice, W. R. (1988). A new probability model for determining exact *P*-values for 2 × 2 contingency tables when comparing binomial proportions. *Biometrics* **44**, 1–14.

William R. Rice
Biology Board of Studies
University of California
Santa Cruz, California 95064, *U.S.A.*