Tatra Mt. Math. Publ. **39** (2008), 193–203

TATRA MOUNTAINS
Mathematical Publications

ON ACCURACY OF P-VALUES IN BINARY RESPONSE MODELS

Lynn Roy LaMotte

ABSTRACT. A framework is proposed for discussing p-values in models for binary response. Within that framework, the target of a p-value is defined, and accuracy is described relative to that target. Likelihood-ratio, F, and Pearson chi-squared approximate p-values, the exact conditional p-value based on the score statistic, and its mid-p version are examined by examples with models for $2 \times c$ contingency tables.

1. Introduction

Several p-values are available to test for effects in models for categorical response variables. Some are called "approximate" and others, "exact". For the same data, exact p-values can differ among themselves to an extent that leads to different interpretations. Approximate p-values differ among themselves, too. They have been called "inaccurate and misleading" [4].

That exact p-values can differ indicates that there is no unique, correct p-value. It is not always clear what an approximate p-value is approximating, and different p-values may have different targets.

It is not clear exactly what "exact" means, either. Agresti's [1] definition is the most inclusive, saying that an exact p-value is a probability computed from "exactly specified distributions". An exact p-value is not necessarily accurate. The inaccuracy of exact conditional p-values led to the development of mid-p-values. "Accuracy" does not seem to be clearly defined either, because the target of an approximate p-value is not fully agreed upon.

A framework for p-values is suggested in this paper, within which a p-value's target, and hence its accuracy, is defined. A fundamental relation between a p-value and its target is established. Comparisons among p-values are shown

2000 Mathematics Subject Classification: 46N30, 65C60.
Keywords: categorical response, conditional p-values, approximate p-values, exact p-values, unconditional p-values.

in some small-sample examples. Computations for the results shown here were done using SAS/IML [10].

2. The setting and notation

The examples shown here are $2 \times c$ contingency tables. In this setting, there are c populations, and the response variable is dichotomous, coded as 0 or 1. In the jth population, $\Pr(Y = 1) = \pi_j$. From the jth population, n_j subjects are sampled independently, $j = 1, \ldots, c$. Denote the $n = \sum n_j$ responses by $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. The sample space is $\mathcal{S} = \{0, 1\}^n$. Denote the population from which the ith subject came by j_i. Denote the c sample sums by $T_j,\ j = 1, \ldots, c$, where $T_j = \sum_{\{i:j_i=j\}} Y_i$. Let $T_0 = \sum_i Y_i$. Denote realized values by lower case letters, $\boldsymbol{y}$ for the response and $t_j = T_j(\boldsymbol{y})$ for the category sums. The joint probability mass function of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\ \pi_1, \ldots, \pi_c) = \pi_1^{t_1}(1 - \pi_1)^{n_1 - t_1} \cdots \pi_c^{t_c}(1 - \pi_c)^{n_c - t_c}.$$

Clearly, $T_1, \ldots, T_c$ form a sufficient statistic for this family of distributions.

We shall consider p-values for the hypothesis that the distribution of the response is the same under all conditions, that is, $\mathrm{H}_0 : \pi_1 = \cdots = \pi_c$. Let π_0 denote the common $\Pr(Y = 1)$ under H_0. Under H_0, the statistic $T_0(\boldsymbol{Y})$ is sufficient for the joint distribution of $Y_1, \ldots, Y_n$.

3. A framework for p-values

A p-value is taken here to be any statistic $\hat{p}(\boldsymbol{y})$ that takes values between 0 and 1, inclusive. It is supposed to tell us that if H_0 were true, the probability of outcomes as extreme as or more extreme than $\boldsymbol{y}$ would be $\hat{p}(\boldsymbol{y})$. The definition of 'more extreme' is taken to mean lesser values of $\hat{p}$, so

$$C_{\hat{p}}(\boldsymbol{y}) = \big\{\boldsymbol{u} \in \mathcal{S} : \hat{p}(\boldsymbol{u}) \le \hat{p}(\boldsymbol{y})\big\}$$

is the set of outcomes in $\mathcal{S}$ defined by $\hat{p}$ to be as extreme as or more extreme than $\boldsymbol{y}$. Call this the *extreme set* for $\boldsymbol{y}$ defined by $\hat{p}$. It is the critical region for the test that rejects H_0 for outcomes $\boldsymbol{u}$ such that $\hat{p}(\boldsymbol{u}) \le \hat{p}(\boldsymbol{y})$. Regarding $\hat{p}(\boldsymbol{y})$ as a test statistic, H_0 is rejected at the nominal α level of significance when $\hat{p}(\boldsymbol{y}) \le \alpha$. For the data $\boldsymbol{y}$, the least α that would lead to rejection of H_0 is $\hat{p}(\boldsymbol{y})$, and the size of the test with this nominal level of significance is $p(\boldsymbol{y})$, where

$$p(\boldsymbol{y}) = \sup_{\pi_0} \Pr\big[C_{\hat{p}}(\boldsymbol{y});\ \pi_0\big].$$

We shall call $p(\boldsymbol{y})$ the *target* of $\hat{p}(\boldsymbol{y})$ at the outcome $\boldsymbol{y}$. The target $p(\boldsymbol{y})$ of $\hat{p}(\boldsymbol{y})$ is itself a p-value. In the literature it is called an *exact unconditional* p-value.

3.1. Relations between a p-value and its target

In this and the next paragraph it is shown that the extreme sets of $\hat{p}$ and p are the same for all $\boldsymbol{y}$. Let $\hat{p}_1 < \hat{p}_2 < \ldots < \hat{p}_K$ denote the distinct values of $\hat{p}$, and denote the corresponding values of p by $p_1, \ldots, p_K$. Denote corresponding extreme sets by $C_{\hat{p}_j}$ and C_{p_j}, $j = 1, \ldots, K$. Because $\{\hat{p}_i : i = 1, \ldots, K\}$ are distinct, $\hat{p}(\boldsymbol{y}) = \hat{p}_i$ implies that $p(\boldsymbol{y}) = p_i$. It is clear that $p_1 \leq p_2 \leq \ldots \leq p_K$ because $C_{\hat{p}_j}$ is a subset of $C_{\hat{p}_{j+1}}$, $j = 1, \ldots, K-1$. If $\boldsymbol{y} \in C_{\hat{p}_i}$, then $\hat{p}(\boldsymbol{y}) = \hat{p}_j$ for some j, $j \leq i$, so $p(\boldsymbol{y}) = p_j \leq p_i$, which implies that $\boldsymbol{y} \in C_{p_i}$. Therefore $C_{\hat{p}_i} \subset C_{p_i}$, $i = 1, \ldots, K$.

Now we shall show that the inclusion goes the other direction. This is accomplished by showing that $p_1, \ldots, p_K$ are distinct, for then $p(\boldsymbol{y}) = p_i$ implies that $\hat{p}(\boldsymbol{y}) = \hat{p}_i$, and the argument in the last paragraph works when the roles of $\hat{p}$ and p are switched. Note that $\Pr(C_{\hat{p}_j}; \pi_0)$ is a sum of terms corresponding to different values of t_0. Each term with $t_0 \neq 0$ and $t_0 \neq n$ is a positive multiple of $\pi_0^{t_0}(1-\pi_0)^{n-t_0}$. If $t_0 = 0$, the term is $(1-\pi_0)^n$, and if $t_0 = n$, it is π_0^n. Note that $t_0 = 0$ only for the outcome $\boldsymbol{y} = 0$, and $t_0 = n$ only for the outcome $\boldsymbol{y} = (1, \ldots, 1)$. Assume that these two outcomes are only in $C_{\hat{p}_K}$. Then for $j < K$, every term in $\Pr(C_{\hat{p}_j}; \pi_0)$ is continuous for $\pi_0 \in [0,1]$, positive for $\pi_0 \in (0,1)$, and 0 for $\pi_0 = 0$ or $\pi_0 = 1$. Therefore $\Pr(C_{\hat{p}_j}; \pi_0)$ takes its maximum at a value $\pi_{0*} \in (0,1)$. Thus

$$\begin{aligned} p_{j+1} &= \sup_{\pi_0} \Pr(C_{\hat{p}_{j+1}}; \pi_0) \\ &\geq \Pr(C_{\hat{p}_{j+1}}; \pi_{0*}) \\ &= p_j + \Pr\big[\hat{p}(\boldsymbol{Y}) = \hat{p}_{j+1}; \pi_{0*}\big]; \end{aligned}$$

the last term is positive, so we may conclude that $p_{j+1} > p_j$, $j = 1, \ldots, K-1$. Therefore the extreme sets of $\hat{p}$ and p are the same. These two statistics induce the same partial ordering on the sample space.

Although $\hat{p}(\boldsymbol{y})$ and $p(\boldsymbol{y})$ take different values, their differences are not related to their performances as test statistics; as such, they are equivalent. One difference is that $p(\boldsymbol{y})$ is always equal to its target: for each $\boldsymbol{y} \in \mathcal{S}$,

$$p(\boldsymbol{y}) = \sup_{\pi_0} \Pr\big[C_{\hat{p}}(\boldsymbol{y}); \pi_0\big] = \sup_{\pi_0} \Pr\big[C_p(\boldsymbol{y}); \pi_0\big]$$

since $C_{\hat{p}}(\boldsymbol{y}) = C_p(\boldsymbol{y})$.

3.2. Accuracy

The accuracy of $\hat{p}(\boldsymbol{y})$ refers to the difference between $\hat{p}(\boldsymbol{y})$ and its target $p(\boldsymbol{y})$. We may say that $\hat{p}$ is *accurate* at $\boldsymbol{y}$ if $\hat{p}(\boldsymbol{y}) = p(\boldsymbol{y})$. It is *conservative* if

$\hat{p}(\boldsymbol{y}) > p(\boldsymbol{y})$ and *anti-conservative* if $\hat{p}(\boldsymbol{y}) < p(\boldsymbol{y})$. From the result just shown, $p(\boldsymbol{y})$ is accurate for all $\boldsymbol{y}$. If $\hat{p}(\boldsymbol{y}) < p(\boldsymbol{y})$, then for some values of $\pi_0 \in \Omega_0$, it overstates the rarity of outcomes as extreme as $\boldsymbol{y}$. Said differently, it gives the impression that H_0 should be rejected at the $\hat{p}(\boldsymbol{y})$ level of significance when in fact it should be rejected only at the $p(\boldsymbol{y})$ level of significance. On the other hand, if $\hat{p}(\boldsymbol{y}) > p(\boldsymbol{y})$, then H_0 is not rejected at the $p(\boldsymbol{y})$ level of significance when it could be rejected. For these reasons, conservative p-values are preferred to anti-conservative p-values, but it is desirable to have a p-value that is minimally conservative.

The values $p_1 < p_2 < \ldots < p_K$ of p can be used to specify a minimally conservative critical value for a size-α test of H_0 based on $\hat{p}$; it is $\hat{p}_j$, where j is the subscript of the greatest p_i that does not exceed α. By the result just shown, the test that rejects H_0 when $\hat{p}(\boldsymbol{y}) \leq \hat{p}_j$ is equivalent to the test that rejects H_0 when $p(\boldsymbol{y}) \leq p_j$.

3.3. P-values based on different statistics

Most p-values are based on a test statistic, like the likelihood-ratio statistic or the conditional score statistic. Let $W(\boldsymbol{Y})$ be a real-valued statistic. Let $\hat{p}_W(\boldsymbol{y})$ denote a p-value based on W. Commonly, approximate p-values are defined as $1 - F\big[W(\boldsymbol{y})\big]$ where F is a fully-specified cumulative distribution function, like a chi-squared distribution. Let $C_W(\boldsymbol{y})$ denote the extreme set at $\boldsymbol{y}$ defined by $\hat{p}_W$ (instead of $C_{\hat{p}_W}(\boldsymbol{y})$).

A *conditional* p-value for the outcome $\boldsymbol{y}$ based on W is

$$\hat{p}_{cW}(\boldsymbol{y}) = \Pr\big[C_W(\boldsymbol{y}) | T_0(\boldsymbol{Y}) = t_0\big]$$

under H_0, where $t_0 = T_0(\boldsymbol{y})$ and the c in the subscript is for "conditional". These are usually called *exact conditional* p-values. Mehta and Hilton [9] show that such p-values are conservative for all outcomes.

H. O. Lancaster [8] proposed *mid*-p-values to lessen the conservativeness of conditional p-values. A mid-p-value based on W is defined as

$$\begin{aligned} \hat{p}_{cW5}(\boldsymbol{y}) &= \Pr(W > w|t_0) + .5\Pr(W = w|t_0) \\ &= \hat{p}_{cW}(\boldsymbol{y}) - (1 - .5)\Pr(W = w|t_0), \end{aligned}$$

where $w = W(\boldsymbol{y})$, and the 5 in the subscript stands for .5 or 50% (other fractions could be used, too). Clearly, $\hat{p}_{cW5}(\boldsymbol{y}) \leq \hat{p}_{cW}(\boldsymbol{y})$ for all outcomes $\boldsymbol{y}$, and so it is less conservative than $\hat{p}_{cW}$. It is possible that there are outcomes for which it is anti-conservative.

Each specification of a criterion statistic W can lead to several p-values based on it, including an approximate p-value $\hat{p}_W$, a conditional p-value $\hat{p}_{cW}$, and a mid-p-value $\hat{p}_{cW5}$. Each of these has its own target (or unconditional) p-value——p_W, p_{cW}, and p_{cW5}, respectively, and all of these may be different. There are

many other possibilities. For example, a conditional p-value can be constructed based on the statistic p_W, and it is not necessarily identical to $\hat{p}_{cW}$.

At this writing, the p-values available in statistical computing packages for contingency tables and logistic regression include: approximate p-values based on Wald, score, and likelihood-ratio statistics; conditional p-values with outcomes ordered by likelihood-ratio and conditional score statistics, and by conditional probability $\Pr(\boldsymbol{y}|t_0)$ under H_0; and mid-p versions of these same conditional p-values. StatXact 5 [5] includes the unconditional exact test for comparing two binomial proportions, which is the target of the Pearson chi-squared approximate p-value. P-values (approximate, exact conditional, and mid-p) based on the likelihood-ratio statistic, conditional score statistic, and the F-statistic obtained from the least-squares regression of the 0-1 response on appropriately defined predictor variables (an intercept and $c-1$ dummy variables, for example) were examined for this paper.

It can be shown that, in the setting considered here, F and the conditional score statistic order outcomes with the same value of t_0 identically, but differently across different values of t_0. This means that conditional p-values based on these two statistics are identical, that is, $\hat{p}_{cF} \equiv \hat{p}_{cS}$.

4. Examples

4.1. Targets and accuracy in a 2×2 table

Five p-values and their targets are shown in Table 1 for the 2×2 table with frequencies 5 and 1 in the first row. KP is Pearson's chi-squared statistic. It and the F statistic are one-to-one for $2 \times c$ tables (see D'Agostino [6]), so $\hat{p}_F$ and $\hat{p}_{KP}$ are one-to-one, and hence their extreme sets C_F and C_{KP} are identical for all outcomes, and they have the same targets. For this table, the extreme set for the LR approximate p-value, C_{LR}, comprises outcomes corresponding to 86 of the 182 distinct pairs of values of the sufficient statistic. The other extreme sets are subsets of this one, with $C_{cS5} \subset C_{cS} \subset C_F$. Computed under H_0, suprema of $\Pr(C_W; \pi_0)$ occur at different values of π_0 for the different criterion statistics. These approximate p-values and probabilities of their extreme sets are shown in Figure 1.

For this table we have five p-values ($\hat{p}$). The four that are less than .05 are approximate, while the one that is greater is said to be exact. Relative to their targets, $\hat{p}_{LR}$ and $\hat{p}_{cS}$ are the most inaccurate; $\hat{p}_{LR}$ is anti-conservative and $\hat{p}_{cS}$ is conservative. The conditional score mid-p-value is quite accurate for this table; $\hat{p}_F$ is less so, but considerably more accurate than $\hat{p}_{LR}$ and $\hat{p}_{cS}$. The targets, all perfectly accurate, differ considerably, from .0440 to .0851. This illustrates

TABLE 1. P-values and their targets for the 2×2 contingency table with column totals $n_1 = 12$ and $n_2 = 13$ and frequencies 5 and 1 in the first row.

Statistic	$\hat{p}$	p
LR	0.0404	0.0851
F	0.0491	0.0532
KP	0.0469	0.0532
cS	0.0730	0.0484
$cS5$	0.0440	0.0440

that accurate p-values based on different statistics can differ considerably for the same data. Different, equally correct answers are possible. Different approximate p-values have different targets, and so their values relative to one another do not say anything about their accuracy. That they cluster together, as the four lesser ones do here, can't be taken as an indication of a correct value by consensus.

Figure 2 shows sixteen plots of pairs of these p-values and their targets for all the 2×2 contingency tables with $n_1 = 12$ and $n_2 = 13$. The approximate p-values considered are $\hat{p}_{LR}$, $\hat{p}_F$, $\hat{p}_{cS}$, and $\hat{p}_{cS5}$. The 2^{25} outcomes in the sample space produce 182 distinct pairs of values of the sufficient statistics T_0 and T_1 with varying multiplicities. Each point shown in the plots represents one of these outcomes. Each p-value and its target were computed exactly for every possible combination of values of the sufficient statistics.

The diagonal plots in Figure 2 show the p-values versus their respective targets. Points representing outcomes for which $\hat{p}$ is close to its target lie close to the diagonal, equi-angular line. Points above this line show that $\hat{p} > p$, and so $\hat{p}$ is conservative; points below the line, anti-conservative. The LR approximate p-value is inaccurate and consistently anti-conservative in the range depicted. The conditional score p-value $\hat{p}_{cS}$ is conservative, but its inaccuracy varies: around $p_{cS} = .04$ it is only slightly inaccurate, while around $p_{cS} = .05$ it is quite conservative. The F approximate p-value $\hat{p}_F$ is slightly anti-conservative, and it is not more inaccurate than $\hat{p}_{cS}$. The mid-p-value $\hat{p}_{cS5}$ compensates nicely for the conservatism of $\hat{p}_{cS}$. It and $\hat{p}_F$ are comparably accurate, but $\hat{p}_{cS5}$ is slightly more accurate overall. As a rough gauge of accuracy, consider how these p-values would perform as tests of H_0 at the 5% level of significance. For $\hat{p}_{LR}$, the five points in the lower-right quadrant in the plot of $\hat{p}_{LR}$ vs. p_{LR} indicate that $\hat{p}_{LR}(\boldsymbol{y}) < .05$ but $p_{LR} > .05$: that is, for these outcomes the test based on $\hat{p}_{LR}$ would reject H_0 because the critical value (.05) is too great, leading to a true size that is about .08 instead of .05. The test that rejects H_0 if $\hat{p}_F \leq .05$

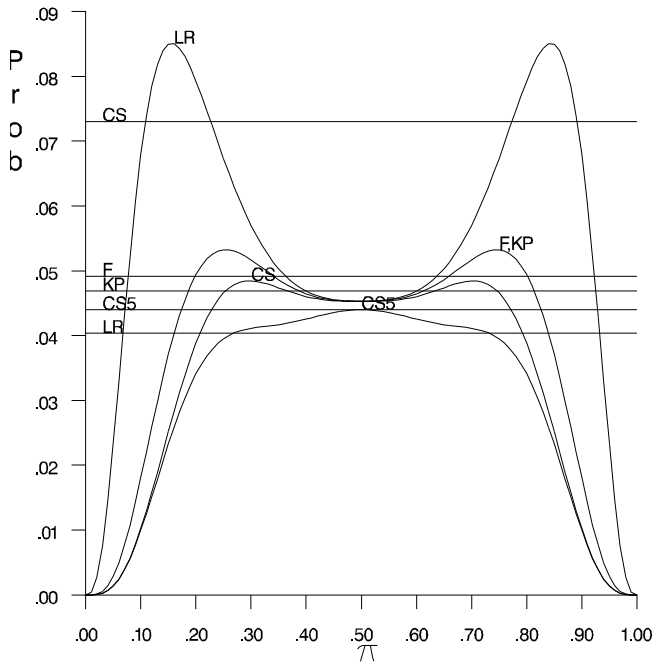

FIGURE 1. Approximate p-values (horizontal lines) and probabilities of extreme sets for a 2×2 contingency table with column totals $n_1 = 12$, $n_2 = 13$ and frequencies 5 and 1 in the first row.

produces only one such discrepant point, as does the test based on $\hat{p}_{cS}$, while the test based on $\hat{p}_{cS5}$ produces none.

The six plots above the diagonal show the extent to which approximate p--values differ for the same outcomes. The six plots below the diagonal depict differences among perfectly accurate p-values. In each graph above and below the diagonal, there is more agreement than disagreement. Below the diagonal, where accurate p-values are depicted, only a few points indicate different conclusions (at the 5% or 10% level of significance). For those few outcomes, though, one accurate p-value would reject H_0 while the other would not—different conclusions, both equally valid.

4.2. Accuracy in Cochran's example

Early in the literature on exact and approximate p-values, W . G . C o c h r a n ([3], pp. 329–330) compared chi-squared approximate p-values and exact conditional p-values for a 2×4 contingency table. His table showed the first, third,

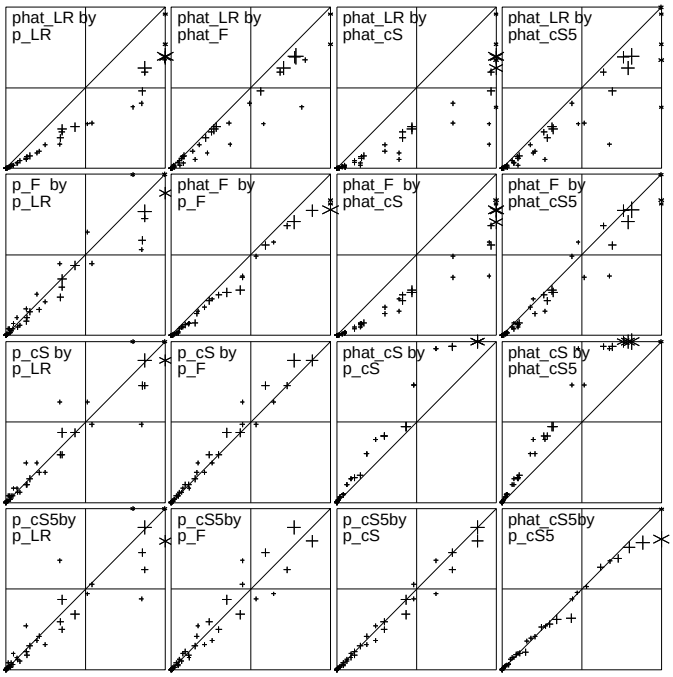

FIGURE 2. 2×2 table, $n_1 = 12$, $n_2 = 13$. Plots of $\hat{p}_{LR}$, $\hat{p}_F$, $\hat{p}_{cS}$, and $\hat{p}_{cS5}$ and their targets against each other. Axes extend from 0 to 0.10. Outcomes with one of the two p-values out of range are shown as *. Plotting symbols are scaled according to multiplicities.

and fifth columns in Table 2, along with a correction for continuity, which is not shown here. Comparing χ^2_3 Table and Conditional p-values, he concluded, "The agreement is not good, the tabular P's being fairly consistently too low". If this setting is a test of homogeneity in independent samples of four subjects each from four populations, and the row totals are not in fact fixed, then the conclusion is not so clear because the two p-values have different targets, shown in Table 2. For outcomes likely to be of interest, say those with $\chi^2_0 \geq 6$, the approximate Table p-value is reasonably close to its target—except for $\chi^2_0 = 16$, closer than the Conditional p-value is to its target. In this sense, the approximate p-value is more accurate than the "exact" Conditional p-value. Regarded as tests at the 10%, 5%, and 1% levels of significance, both approximate p-values (χ^2_3 Table and Conditional) lead to the same conclusions as their targets, except when $\chi^2_0 = 8$, where Conditional fails to reject at the 5% level.

TABLE 2. Chi-squared p-values and their targets for a 2×4 contingency table, for homogeneity of row probabilities. Rows correspond to the 0–1 response, columns to populations. Row sums are 8, column sums are all 4. 'Count' is frequency among all possible tables with $n = 16$ and the fixed column totals. 'Conditional' is the conditional probability computed from this frequency distribution of χ_0^2; for example, .064 = (432+384+6)/12870. Targets are suprema (under H_0) of unconditional probabilities of $\leq$ p--values.

P-values, $\Pr[\chi^2 \geq \chi_0^2]$					
χ_0^2	Count	χ_3^2 Table	Target	Conditional	Target
0	1296	1.000	1.000	1.000	1.000
2	6912	.572	.774	.899	.703
4	1536	.261	.312	.362	.303
6	2304	.112	.126	.243	.133
8	432	.046	.050	.064	.045
10	384	.019	.012	.030	.012
16	6	.0011	.0005	.0005	.0001

5. Discussion

Computing exact conditional p-values, determining the accuracy of a p-value, and finding its target in order to get an accurate p-value are daunting computational tasks. It is reasonable to ask whether there is substantial gain to be had by doing so. There seems to have been more effort devoted to developing algorithms for exact conditional p-values than to assessing the extent to which they are more useful than existing approximate p-values.

Mehta and Hilton [9] examined the accuracy of p-values in 2×3 contingency tables with total sample sizes ranging from 30 to 900. Surprisingly, they did not examine the customary chi-squared approximate p-value in this context. For comparing three binomial probabilities with samples of size 10 each, the size of the nominal 5% level of significance test, which rejects H_0 if the Pearson chi-squared statistic is $\geq$ the upper 5th percentile of a chi-squared distribution with 2 degrees of freedom, is 0.0501. For sample sizes of 20 each it is 0.0524; 5 each, 0.0588. Berger and Boos [2] did not consider the chi-squared approximate p-value in their 2×2 example either; to three decimal places, it gives the same p-value (0.037) that they get by maximizing the probability of lesser p-values over a confidence set for the nuisance parameter. For comparing two binomial probabilities (2×2 tables), D'Agostino *et al.* [7] investigated the accuracies of the Fisher exact test (an exact conditional test) and the chi-squared

and t tests (these two test statistics are one-to-one for two-sided alternatives, but they get approximate p-values from their respective eponymous distributions). Noting that the Fisher exact test is "extremely conservative" and the actual significance levels of the approximate tests are reasonably close to their nominal levels, they recommended that the chi-squared test "should replace the Fisher exact test".

The examples shown in this paper are limited in scope. They share one characteristic, that they are all based on small sample sizes, smaller than in most of the studies mentioned above. Two things stand out. The conditional score mid-p-value performs well in terms of accuracy. And the approximate p-value based on the F statistic performs acceptably, even surprisingly, well, even with the small sample sizes considered here. On the negative side, the LR approximate p-value is anti-conservative, and the conditional score p-value is sometimes quite conservative. The targets of these p-values, considered as (exact unconditional) p-values in their own right, are, of course, accurate, and they can give different indications from the same data.

Until a good reason is found for using LR-based p-values, among those p--values examined for this paper, $\hat{p}_{cS5}$ and $\hat{p}_F$ appear to be acceptable. The amount of computation required to get $\hat{p}_F$ is trivial compared to that required to get $\hat{p}_{cS5}$, except for very small sample sizes. At customary levels of significance, they appear to give similar results, which is not surprising, given the relation noted earlier.

REFERENCES

[1] AGRESTI, A.: *Exact inference for categorical data: recent advances and continuing controversies*, Statistics in Medicine **20** (2001), 2709–2722.

[2] BERGER, R. L.—BOOS, D. D.: *P values maximized over a confidence set for the nuisance parameter*, J. Amer. Statist. Assoc. **89** (1994), 1012–1016.

[3] COCHRAN, W. G.: *The χ^2 test of goodness of fit*, Ann. Math. Stat. **23** (1952), 315–345.

[4] CYTEL SOFTWARE CORPORATION.: *LogXact 5: The Fastest, Most Powerful and Most Reliable Exact Logistic Regression Software Available*, Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA, USA, 2002.

[5] CYTEL SOFTWARE CORPORATION: *StatXact 5*, Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA, USA, 2001.

[6] D'AGOSTINO, R. B.: *Relation between the Chi-squared and ANOVA tests for testing the equality of k independent dichotomous populations*, Amer. Statist. **26** (1972), 30–32.

[7] D'AGOSTINO, R. B.—CHASE, W.—BELANGER, A.: *The appropriateness of some common procedures for testing the equality of two independent binomial samples*, Amer. Statist. **42** (1988), 198–202.

[8] LANCASTER, H. O.: *Significance tests in discrete distributions*, J. Amer. Statist. Assoc. **56** (1961), 223–234.

[9] MEHTA, C. R.—HILTON, J. F.: *Exact power of conditional and unconditional tests: going beyond the* 2×2 *contingency table*, Amer. Statist. **47** (1993), 91–98.
[10] SAS INSTITUTE INC.: *The SAS System for Windows, Release 9.00 TS Level 00M0*, SAS Institute, Inc., Cary, NC, USA, 2002.

Received November 11, 2006

Biostatistics Program
LSU School of Public Health
1615 Canal Street
New Orleans, LA 70112
USA
E-mail: llamot@lsuhsc.edu