Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions

BARRY E. STORER and CHOONGRAK KIM*

Using exact unconditional distributions, we evaluated the size and power of four exact test procedures and three versions of the χ^2 statistic for the two-sample binomial problem in small-to-moderate sample sizes. The exact unconditional test (Suissa and Shuster 1985) and Fisher's (1935) exact test are the only tests whose size can be guaranteed never to exceed the nominal size. Though the former is distinctly more powerful, it is also computationally difficult. We propose an alternative that approximates the exact unconditional test by computing the exact distribution of the χ^2 statistic at a single point, the maximum likelihood estimate of the common success probability. This test is a modification of the test of Liddell (1978), which considered the exact distribution of the difference in the sample proportions. Our test is generally more powerful than either the exact unconditional or Liddell's test, and its true size rarely exceeds the nominal size. The uncorrected χ^2 statistic is frequently anticonservative, but the magnitude of the excess in size is usually moderate. Though this point has been somewhat controversial for many years, we endorse the view that one should not use Fisher's exact test or Yates's continuity correction in the usual unconditional sampling setting.

KEY WORDS: Chi-squared statistic; Continuity correction; Exact unconditional test.

1. INTRODUCTION

The comparison of two binomial samples is one of the most commonly encountered problems in statistics. Specifically, we observe r_1 successes out of n_1 observations from population 1 and r_2 successes out of n_2 observations from population 2, and we are interested in testing $H_0 : p_1 = p_2$, where p_1 and p_2 are the unknown true success probabilities in the two populations. Alternatively, we write $p_1 = p$ and $p_2 = p + \Delta$, and are interested in testing $H_0 : \Delta = 0$. The fact that p is an unknown nuisance parameter makes this a less than straightforward problem. This article examines the properties of some exact (and some unexact) procedures used to test H_0 in small-to-moderate sample sizes.

The term *exact* here refers to the use of an exact discrete distribution in computing the significance level (p value) associated with a test procedure; it does not refer to the size of the procedure being exactly as specified. In this discrete setting, only a randomized test can have the latter property. In fact, only the randomized version of Fisher's (1935) exact test is exact in both senses and, indeed, is uniformly most powerful unbiased (Tocher 1950) with either conditional or unconditional sampling. Nevertheless, we think randomized procedures have little or no appeal in applications.

It has been realized for some time that the nonrandomized version of Fisher's exact test is distinctly conservative, relative to the familiar χ^2 statistic, when applied to the case of independent binomial samples (Grizzle 1967; Starmer, Grizzle, and Sen 1974). On the other hand, many authors continue to advocate the use of Fisher's exact test. For example, Fleiss (1981) did not even provide the formula for the uncorrected χ^2 statistic, instead recommending that one *always* use Yates's (1934) continuity correction, since it approximates probabilities associated with Fisher's exact test so well.

Liddell (1978) reviewed these issues and proposed a test based on the exact distribution of the difference in sample proportions, $\hat{p}_1 - \hat{p}_2 = r_1/n_1 - r_2/n_2$, when p is taken to be its maximum likelihood estimate (MLE) under H_0, namely $\hat{p} = (r_1 + r_2)/(n_1 + n_2)$. He also considered the use of a continuity correction to the χ^2 statistic that more nearly corrects for the degree of discreteness inherent in the independent, rather than conditional, sampling scheme (Pirie and Hamdan 1972). This correction is half of the usual Yates correction.

Subsequently, Suissa and Shuster (1985) described the use of the exact unconditional test, which seeks to eliminate the effect of the nuisance parameter by maximizing the size of the test over the domain of the nuisance parameter (Basu 1977). As in Liddell's exact test, the significance level at a particular value of the nuisance parameter is obtained using an exact discrete distribution, except that the rejection region is based on a standardized version of $\hat{p}_1 - \hat{p}_2$. The size of this test cannot exceed the nominal size, yet is distinctly greater than that of the conditional exact test; however, it is computationally inconvenient.

As an alternative, we propose a modification of Liddell's exact test that is based on the exact distribution of the standardized version of $\hat{p}_1 - \hat{p}_2$, though we use a different variance estimate than Suissa and Shuster (1985). We call this the *approximate unconditional test*. Since the significance level is evaluated only at $\hat{p}$, it is far easier to compute than the exact unconditional test; otherwise its properties are very similar.

In Section 2 we describe the test procedures and the methods used to evaluate them. Results are presented in Section 3; they graphically illustrate the inherent discreteness of all of the procedures and the relative merits of each.

* Barry E. Storer is Assistant Professor, Departments of Statistics and Human Oncology, University of Wisconsin, Madison, WI 53706. Choongrak Kim is a graduate student, Department of Statistics, University of Wisconsin, Madison, WI 53706. This work was supported by National Institutes of Health Grant R29-CA45313. The authors thank a referee for helpful comments.

© 1990 American Statistical Association
Journal of the American Statistical Association
March 1990, Vol. 85, No. 409, Theory and Methods

2. METHODS

2.1 Tests Compared

Each of the seven test procedures is described in the following. Throughout we use the notation $R = r_1 + r_2$, $N = n_1 + n_2$, $V = n_1 \cdot n_2 \cdot R \cdot (N - R)/N$, $m_j = n_j - r_j$, $\hat{p}_j = r_j/n_j$, χ^2_α is the upper $100(1 - \alpha)$ percentile of the χ^2 distribution with one degree of freedom, $K(n, r) = n!/(n - r)!/r!$, and $I[\cdot]$ is the indicator function.

1. Exact unconditional (EU) test of Suissa and Shuster (1985):

$$T_{\text{EU}} = \sup_{0 \le p \le 1} \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} b(x, n_1, p)$$

$$\times\ b(y, n_2, p) \cdot I[|Z'(x, y, n_1, n_2)| \ge |Z'(r_1, r_2, n_1, n_2)|],$$

where $b(r, n, p) = K(n, r) \cdot p^r \cdot (1 - p)^{n-r}$ and $Z'(x, y, n_1, n_2) = (x/n_1 - y/n_2)/[(x/n_1) \cdot (1 - x/n_1)/n_1 + (y/n_2) \cdot (1 - y/n_2)/n_2]^{1/2}$, the standardized Z statistic using an unpooled variance estimate; reject if $T_{\text{EU}} \le \alpha$.

2. The approximate unconditional (AU) test:

$$T_{\text{AU}} = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} b(x, n_1, \hat{p})$$

$$\times\ b(y, n_2, \hat{p}) \cdot I[|Z(x, y, n_1, n_2)| \ge |Z(r_1, r_2, n_1, n_2)|],$$

where $\hat{p} = R/N$ is the MLE of p under H_0 and $Z(x, y, n_1, n_2) = (x/n_1 - y/n_2)/[(1/n_1 + 1/n_2) \cdot \hat{p} \cdot (1 - \hat{p})]^{1/2}$, the standardized Z statistic using the pooled variance estimate; reject if $T_{\text{AU}} \le \alpha$.

3. Liddell's (1978) exact (LE) test:

$$T_{\text{LE}} = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} b(x, n_1, \hat{p}) \cdot b(y, n_2, \hat{p})$$

$$\times\ I[|x/n_1 - y/n_2| \ge |r_1 - r_2|];$$

reject if $T_{\text{LE}} \le \alpha$.

4. Fisher's (1935) exact (FE) test:

$$T_{\text{FE}} = \sum_{x=x_1}^{x_2} h(x, R - x, n_1, n_2)$$

$$\times\ I[h(x, R - x, n_1, n_2) \le h(r_1, r_2, n_1, n_2)],$$

where $h(r_1, r_2, n_1, n_2) = K(n_1, r_1) \cdot K(n_2, r_2)/K(N, R)$, $x_1 = \max(0, R - n_2)$, and $x_2 = \min(n_1, R)$; reject if $T_{\text{FE}} \le \alpha$.

5. Uncorrected χ^2 statistic (UC): $T_{\text{UC}} = (r_1 \cdot m_2 - r_2 m_1)^2/V$; reject if $T_{\text{UC}} \ge \chi^2_\alpha$.

6. χ^2 statistic with the Pirie and Hamdan (1972) continuity correction (PC): $T_{\text{PC}} = (\max\{0, |r_1 \cdot m_2 - r_2 \cdot m_1| - N/4\})^2/V$; reject if $T_{\text{PC}} \ge \chi^2_\alpha$.

7. χ^2 statistic with Yates's (1934) continuity correction (YC): $T_{\text{YC}} = (\max\{0, |r_1 \cdot m_2 - r_2 \cdot m_1| - N/2\})^2/V$; reject if $T_{\text{YC}} \ge \chi^2_\alpha$.

2.2 Computations

For given values of sample size, population parameters, and nominal size, the exact probability of rejection for test procedure T_A, based on the true underlying binomial distributions, is computed as

$$S_A(n_1, n_2, p_1, p_2, \alpha)$$

$$= \sum_{r_1=0}^{n_1} \sum_{r_2=0}^{n_2} b(r_1, n_1, p_1) \cdot b(r_2, n_2, p_2) \cdot I[T_A \in R_A(\alpha)],$$

where $R_A(\alpha)$ is the appropriate rejection region for procedure T_A for a nominal size-α test. Hence the true size of a test procedure is obtained by evaluating $S_A(n_1, n_2, p, p, \alpha)$.

The amount of computation involved in evaluating the functions S_A over a wide range of their arguments is considerable, with the exact unconditional test contributing the most to this difficulty. The function

$$\pi(q) = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} b(x, n_1, q)$$

$$\times\ b(y, n_2, q) \cdot I[|Z'(x, y, n_1, n_2)| \ge |Z'(r_1, r_2, n_1, n_2)|],$$

which, for any fixed n_1, n_2, p_1, p_2, and α, must potentially be maximized for every possible pair of outcomes (r_1, r_2), is a linear combination of polynomials and hence a smooth but possibly multimodal function of q. We maximized π using a direct search over $q = .001(.001).5$, using quadratic interpolation whenever $\pi' \ge .1$.

For each of the seven test statistics, we computed $S_A(n_1, n_2, p, p, \alpha)$ over a grid G of quadruplets (n_1, n_2, p, α) intended to represent a range of combinations that might be of interest in the small-to-moderate-sample-size setting. For the equal-sample-size case we considered $n_1 = n_2 = n = 1(1)80$, and for the unequal case $n_1 = 1(1)50$ and $n_2 = 1(1)n_1$; for both cases we considered $p = .01(.01).50$ and $\alpha = .001, .005, .01(.01).20$. The number of points in G is thus 88,000 for the equal-sample-size case and 1,347,500 for the unequal case.

Graphical results for the four exact statistics and T_{UC} are displayed in Figures 1–3. In Figure 1, S_A is plotted as a function of n for fixed values of p and α; in Figure 2, n and α are fixed and p is varied; and in Figure 3, n and p are fixed and α is varied. For the latter figure only, S_A was computed for $\alpha = .005(.005).20$. In each figure the left panels represent the equal-sample-size case and the right the unequal-sample-size case. To improve clarity the two corrected χ^2 statistics are omitted from the plots, but they are included in other comparisons.

Summary comparisons of all of the statistics, relative to the nominal size and to each other, are presented in Table 1. For these comparisons we define $C(A, \alpha, \varepsilon) = k \cdot \Sigma_G I[S_A \le \alpha_\varepsilon]$, where $\alpha_\varepsilon = \alpha \cdot (1 + \varepsilon)$ and $k = 100/\text{card}(G)$. Hence $C(A, \alpha, \varepsilon)$ is simply the percentage of the points in G where S_A exceeded the nominal size α by the fraction ε. For comparing the sizes of tests T_A and T_B, we define $C(A, B, \varepsilon) = k \cdot \Sigma_G I[S_A \le \alpha_\varepsilon] \cdot \{I[S_B > \alpha_\varepsilon] + I[S_B \le \alpha_\varepsilon] \cdot I[|S_A - S_B| > \alpha \cdot \varepsilon] \cdot I[|S_A - \alpha| < |S_B - \alpha|]\}$. That is, for each point in G, test A is counted as better than test B if (a) S_A is less than the tolerated size and S_B exceeds it or (b) both S_A and S_B are within the tolerated size, the difference between the two exceeds an amount propor-

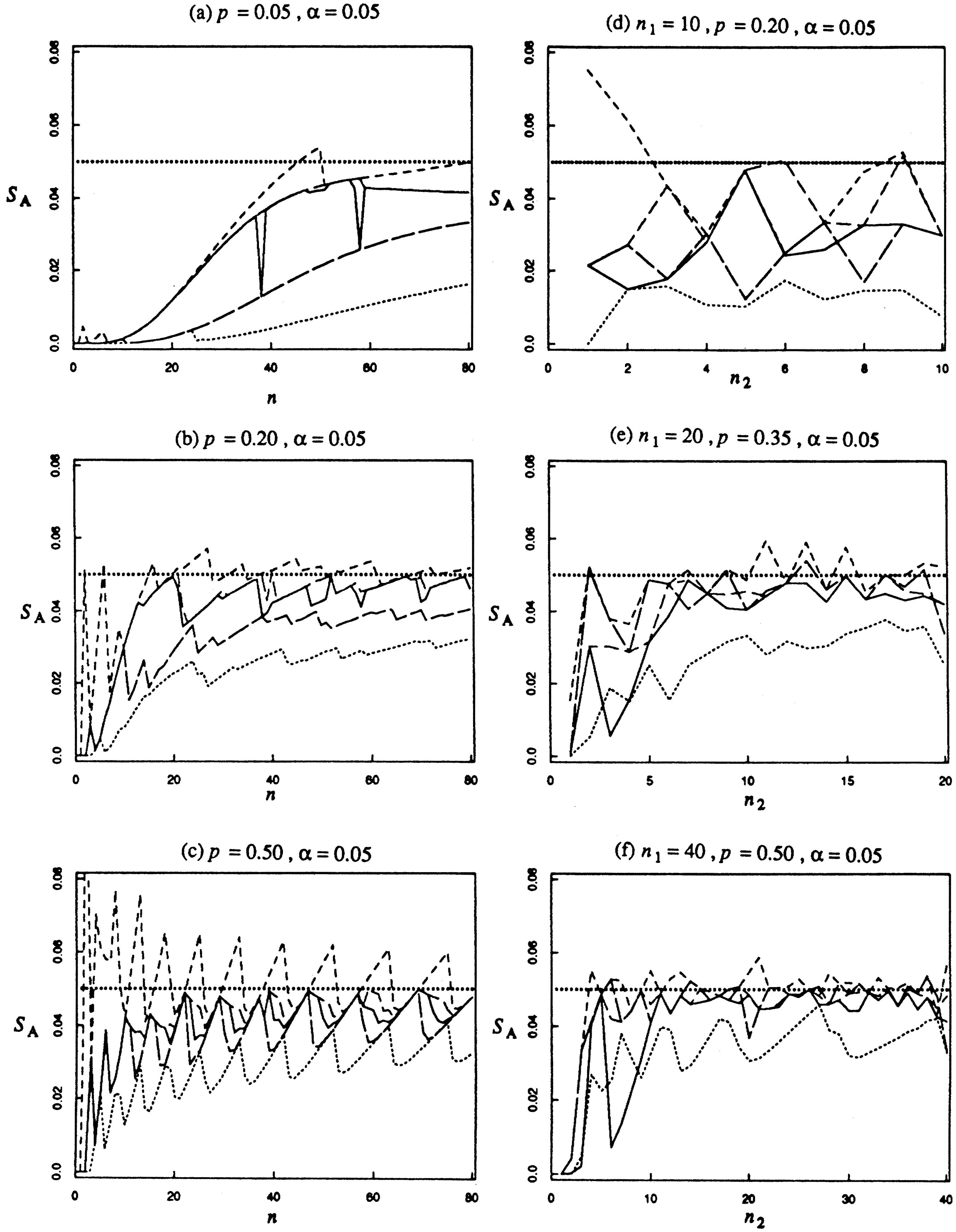
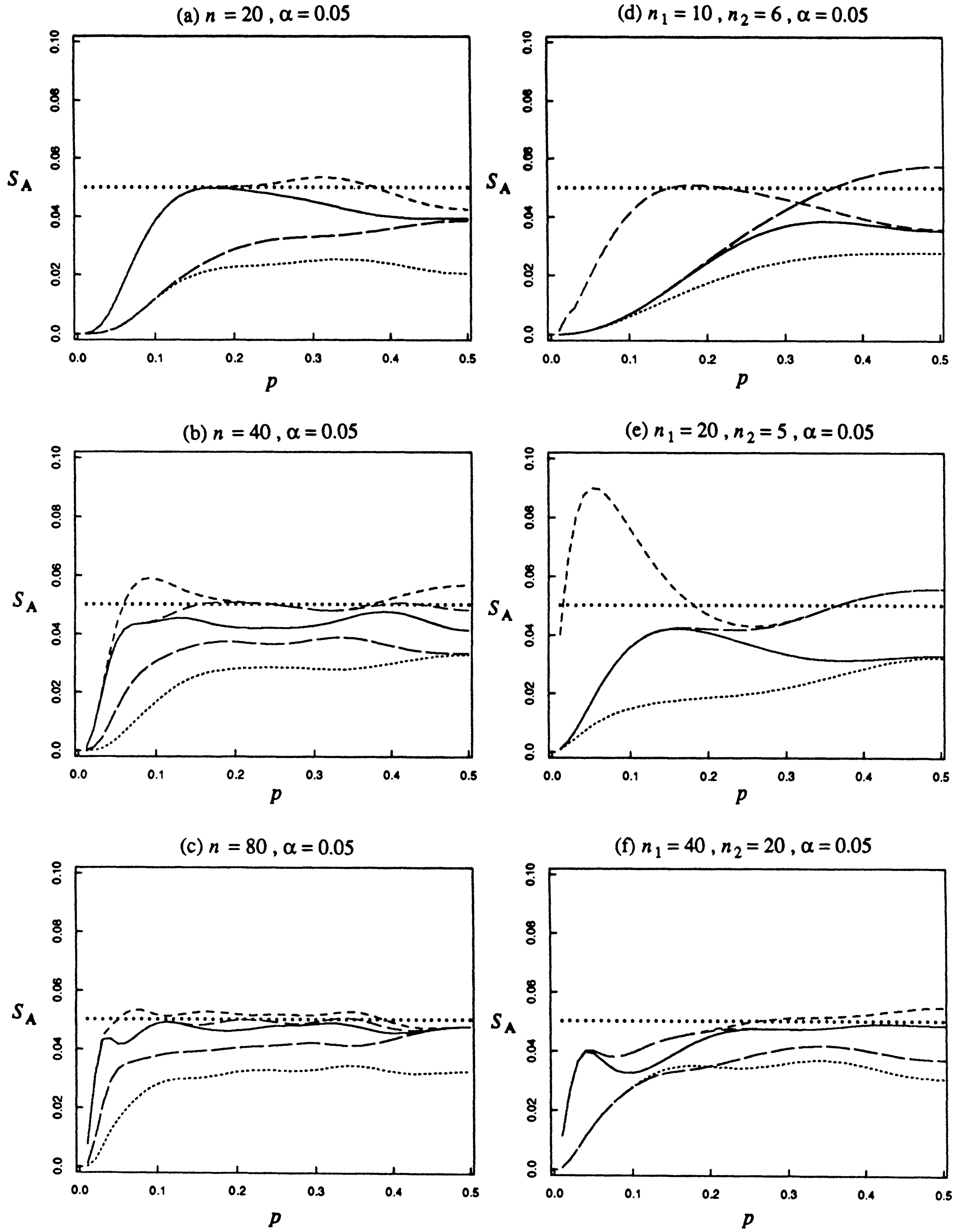


Figure 1. Size of Nominal Size-α = .05 Test Procedures as a Function of Sample Size: ——, T_{EU}; – –, T_{AU}; — —, T_{LE}; ···, T_{FE}; ---, T_{UC}.



(a) $n = 20$, $\alpha = 0.05$

(b) $n = 40$, $\alpha = 0.05$

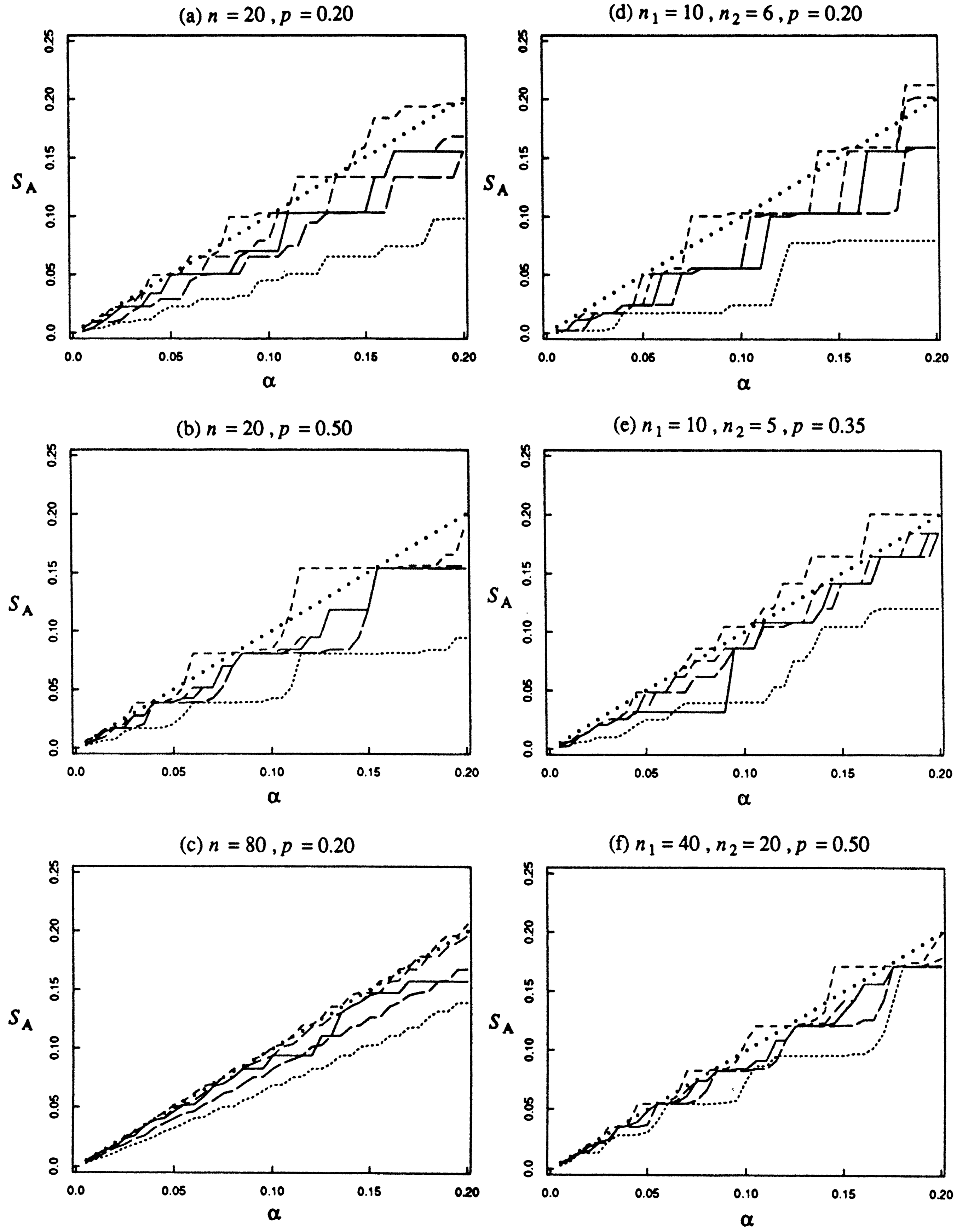
(c) $n = 80$, $\alpha = 0.05$

(d) $n_1 = 10$, $n_2 = 6$, $\alpha = 0.05$

(e) $n_1 = 20$, $n_2 = 5$, $\alpha = 0.05$

(f) $n_1 = 40$, $n_2 = 20$, $\alpha = 0.05$

Figure 2. Size of Nominal Size-α = .05 Test Procedures as a Function of a Nuisance Parameter: ——, T_{EU}; – –, T_{AU}; — —, T_{LE}; ···, T_{FE}; ---, T_{UC}.



(a) $n = 20, p = 0.20$

(d) $n_1 = 10, n_2 = 6, p = 0.20$

(b) $n = 20, p = 0.50$

(e) $n_1 = 10, n_2 = 5, p = 0.35$

(c) $n = 80, p = 0.20$

(f) $n_1 = 40, n_2 = 20, p = 0.50$

Figure 3. Size of Test Procedures as a Function of Nominal Size: ——, T_{EU}; – –, T_{AU}; — —, T_{LE}; · · · , T_{FE}; ---, T_{UC}.

Table 1. Performance of Test Procedures Relative to Each Other (C(A, B, ε)) and to the Nominal Size (C(A, α, ε)) for Various Degrees of Tolerance (ε)

Test procedure		ε (equal sample size)					ε (unequal sample size)				
A	B	0	.01	.05	.10	.20	0	.01	.05	.10	.20
EU	α	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	AU	6.95	4.45	.61	.05	.00	8.51	5.87	2.25	1.03	.32
	LE	77.91	66.40	56.08	44.29	19.94	34.87	27.15	15.09	8.94	4.01
	FE	95.20	92.33	88.84	85.06	74.10	93.52	90.01	85.37	79.04	56.22
	UC	50.89	45.77	28.02	17.03	8.63	50.71	46.21	30.50	19.09	9.17
	PC	81.52	70.53	60.40	48.73	24.20	90.30	86.23	81.63	74.42	49.88
	YC	95.64	92.86	89.48	85.89	75.95	96.18	94.56	92.72	90.97	86.64
AU	α	6.95	4.45	.41	.05	.00	8.51	5.84	2.12	.87	.19
	EU	54.69	41.82	34.12	22.29	8.05	60.01	45.77	37.01	25.74	14.44
	LE	81.99	76.55	73.27	65.52	33.54	46.35	36.60	24.98	17.12	9.37
	FE	88.90	89.99	92.28	90.89	86.75	89.86	90.69	92.14	90.44	76.38
	UC	49.77	45.89	32.27	20.65	11.22	53.73	50.10	37.33	25.06	11.98
	PC	83.71	79.19	76.47	69.39	38.76	87.83	87.95	89.34	87.74	68.35
	YC	89.20	90.24	92.54	91.21	87.31	89.79	91.32	93.75	93.90	92.68
LE	α	.06	.05	.03	.01	.00	19.74	15.75	5.83	1.69	.23
	EU	5.35	3.52	1.99	.81	.37	43.90	36.14	33.92	23.51	10.81
	AU	6.89	4.40	.38	.04	.00	22.30	15.44	11.49	5.35	1.29
	FE	92.64	86.47	79.95	73.43	41.28	77.47	78.08	84.80	85.30	70.97
	UC	50.48	45.45	26.74	15.15	6.45	33.90	33.25	26.75	17.31	8.15
	PC	25.35	15.01	9.04	4.82	2.03	75.67	75.51	81.99	82.30	65.15
	YC	94.82	89.62	83.67	77.86	49.78	78.45	80.69	88.76	91.37	90.23
FE	α	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	EU	.00	.00	.00	.00	.00	4.04	3.57	2.76	1.98	.97
	AU	6.95	4.45	.41	.05	.00	8.52	5.85	2.13	.87	.19
	LE	.63	.49	.29	.07	.00	20.18	16.06	5.97	1.73	.23
	UC	50.54	45.50	26.77	15.15	6.43	49.31	45.15	28.20	14.94	5.97
	PC	.36	.29	.16	.04	.00	35.24	28.61	17.18	7.92	1.32
	YC	38.73	24.35	14.04	6.65	1.02	93.97	89.02	84.29	77.64	47.32
UC	α	50.54	45.50	26.77	15.15	6.41	49.25	45.11	28.19	14.92	5.91
	EU	36.63	33.33	40.80	37.88	22.98	39.51	33.78	39.13	34.41	19.71
	AU	32.04	23.89	23.58	17.16	9.84	28.84	20.16	19.37	13.16	6.49
	LE	44.86	45.59	59.42	66.48	57.20	34.69	25.47	23.14	18.60	11.64
	FE	47.09	50.78	67.98	78.42	85.18	49.96	52.54	67.72	79.12	79.97
	PC	45.69	46.87	61.05	68.42	61.14	49.86	51.89	66.79	78.08	77.90
	YC	47.26	50.96	68.18	78.65	85.61	50.35	53.67	69.48	81.77	89.12
PC	α	.02	.01	.01	.00	.00	.67	.65	.61	.55	.47
	EU	4.52	2.96	1.55	.58	.21	5.30	4.38	3.76	3.21	2.13
	AU	6.93	4.43	.40	.05	.00	9.64	6.65	2.78	1.47	.72
	LE	.70	.47	.28	.10	.02	20.96	16.67	6.59	2.38	.83
	FE	91.67	84.90	77.91	70.57	35.59	55.42	44.60	29.12	16.08	6.11
	UC	50.53	45.49	26.76	15.16	6.45	48.58	44.45	27.58	14.37	5.47
	YC	94.16	88.51	82.15	75.74	44.09	97.62	93.74	90.00	85.42	57.45
YC	α	.00	.00	.00	.00	.00	.05	.05	.04	.04	.03
	EU	.00	.00	.00	.00	.00	1.79	1.35	.97	.59	.20
	AU	6.95	4.45	.41	.05	.00	8.72	5.98	2.23	.96	.26
	LE	.06	.05	.03	.01	.00	19.90	15.86	5.91	1.76	.29
	FE	.00	.00	.00	.00	.00	1.54	1.13	.89	.73	.48
	UC	50.54	45.50	26.77	15.15	6.43	49.20	45.06	28.14	14.88	5.88
	PC	.02	.01	.01	.00	.00	.62	.61	.56	.51	.43

tional to α, and S_A is closer to α than S_B. Thus $C(A, B, \varepsilon)$ is the percentage of the points in G where A is better than B at the indicated level of tolerance. Note that even when $\varepsilon = 0$ it is possible for $C(A, B, \varepsilon) + C(B, A, \varepsilon) < 1$, since the rejection regions for two discrete tests can be identical at some points in G.

Figure 4 presents a further summary of the data in Table 1. For a procedure T_A, one panel plots the minimum level of performance of T_A against all of the other tests as a function of ε, that is, $\min_B C(A, B, \varepsilon)$; similarly, another panel plots the maximum level of performance of all of the other tests against T_A, that is, $\max_B C(B, A, \varepsilon)$. Thus the performance of the tests may be compared informally using a kind of maximin or minimax criterion, respectively.

3. RESULTS

3.1 Equal-Sample-Size Case

Figure 1 dramatically illustrates the inherent discreteness of all of the test procedures, particularly for values of p close to .50 and (of course) for small sample sizes. Although the simple uncorrected χ^2 statistic (T_{UC}) is actually centered closest to the nominal size, this centering means that the test is just as frequently liberal as con-

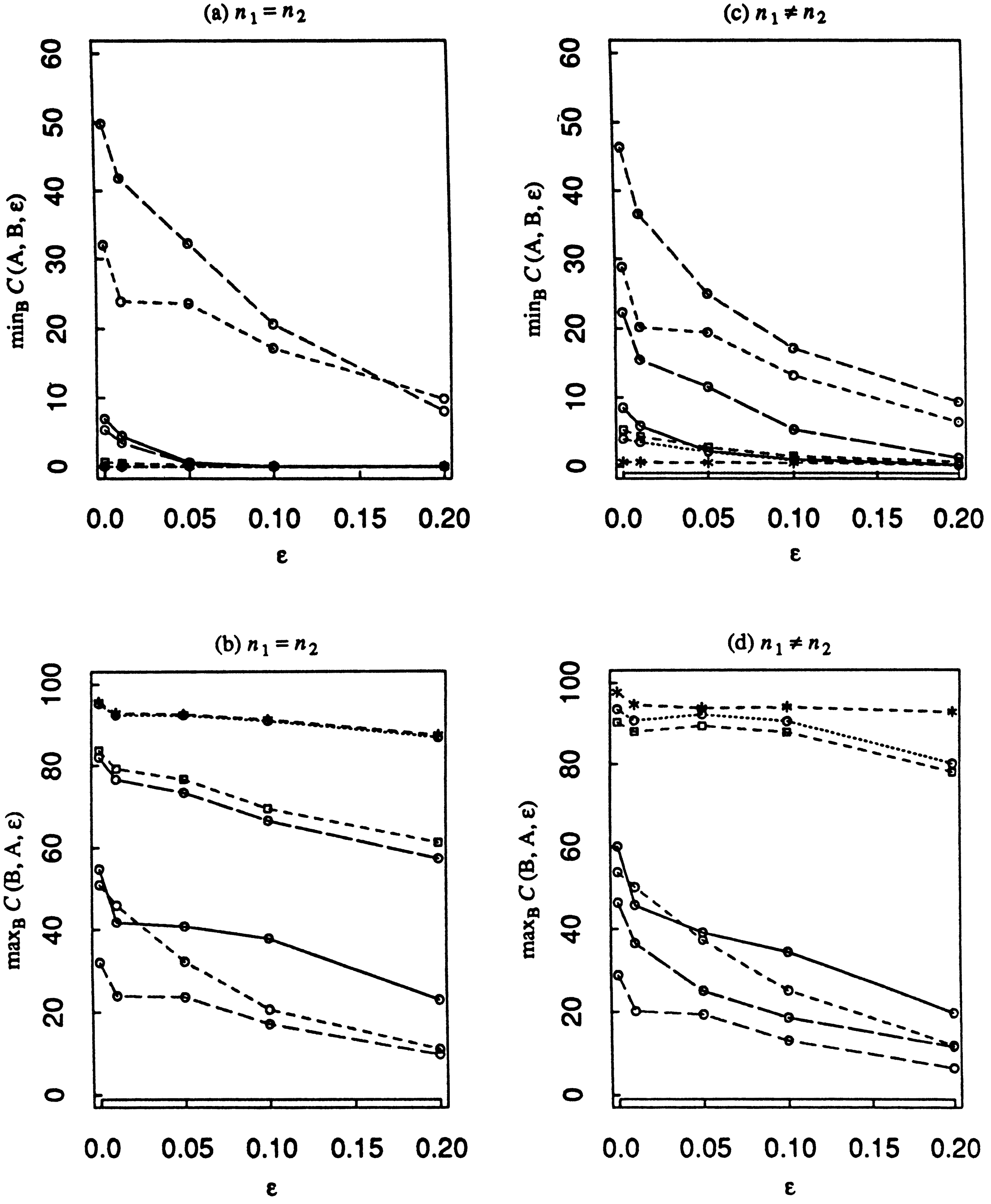


Figure 4. Minimum and Maximum Performance of Test Procedures in Comparison to All Others as a Function of Tolerance: ——O——, T_{EU}; –O–, T_{AU}; —O—, T_{LE}; ···O···, T_{FE}; --O--, T_{UC}; --□--, T_{PC}; --*--, T_{YC}.

servative. The most extreme case occurs when $p = .50$ and $n = 2$, where the outcome (0, 2) or (2, 0), yielding $T_{UC} = 4.0$, occurs with probability .125. Although this particular outcome is not of practical significance, liberal peaks continue to occur in the range .06 up to the largest sample sizes examined.

As it should, the exact unconditional test (T_{EU}) never exceeds the nominal size; however, the simpler test T_{AU} almost shares this property, barely exceeding the nominal size in a very few instances. The noticeable dips in the size of T_{EU} for $p = .05$ are real and occur when a single outcome of substantial probability is in the rejection region of T_{AU} but not T_{EU}. Fisher's exact test is by far the most conservative, and approaches the nominal size very slowly even at the largest sample sizes. The smoother functions in Figure 2 convey the same general impressions. Finally, Figure 3 demonstrates that those impressions are not due to the particular choice of α, although as would be expected the magnitude of differences between the true sizes of the tests and the nominal size is roughly proportional to the latter.

From the figures and table several conclusions may be drawn. First, although both S_{EU} and S_{FE} are guaranteed never to exceed the nominal size, T_{EU} is a much more powerful procedure than T_{FE}. Not only is S_{EU} always at least as large as S_{FE}, but it exceeds S_{FE} by more than 20% of the nominal size nearly 75% of the time. On the other hand, T_{EU} is computationally difficult. By comparison, T_{AU} is relatively easy to compute and exceeds the nominal size infrequently. When it does exceed the nominal size, the error is generally small, exceeding 5% of the nominal size at less than .5% of the examined points. Furthermore, T_{AU} outperforms T_{EU} more than half of the time on an absolute basis, and more than a third of the time at the 5% tolerance level; in addition, it is measurably more powerful than T_{LE}, though the latter exceeds the nominal size even less often.

As mentioned before, the two corrected χ^2 statistics closely approximated one of the exact statistics, though they are conservative. S_{YC} never exceeded S_{FE}, and the two were within 5% of the nominal size at nearly 85% of the examined points. Similarly, S_{PC} almost never exceeded S_{LE}, and the two were within 5% of the nominal size at more than 90% of the examined points. The uncorrected χ^2 is perhaps the most difficult test to place. Although it clearly can be a liberal test, the excess over the nominal size is not necessarily large. If 10% or 20% tolerance is allowed, then T_{AU} and T_{UC} outperform each other, and the other procedures, about equally. Overall, however, by either an informal minimax or maximin criterion T_{AU} outperforms all of the procedures under consideration (see Fig. 4).

3.2 Unequal-Sample-Size Case

Although the aforementioned results are relevant to the planning of experiments and clinical trials (see Sec. 3.3), it is obvious that many, if not most, comparisons made in practice do not involve equal sample sizes. We present these results in a separate section because they are somewhat more ambiguous than those given previously, though the general conclusions are similar.

Although Suissa and Shuster (1985) claimed that the rejection region for an exact binomial test based on the Z statistic with a pooled variance estimator is identical to the one with the unpooled variance estimate that they used in defining T_{EU}, this result is true only for the equal-sample-size case. In general, the choice of Z (using the pooled variance estimate) versus Z' (using the unpooled variance estimate) leads to quite different rejection regions. Consequently, we first compared T_{EU} and a modified procedure using the Z statistic with the pooled variance estimator, denoted by T_{EUP}. These results indicated that T_{EUP} is the preferred procedure. For example, $C(\mathrm{EUP}, \mathrm{EU}, 0) = 67.20\%$ versus $C(\mathrm{EU}, \mathrm{EUP}, 0) = 24.77\%$, $C(\mathrm{EUP}, \mathrm{EU}, .05) = 52.29\%$ versus $C(\mathrm{EU}, \mathrm{EUP}, .05) = 14.29\%$, and $C(\mathrm{EUP}, \mathrm{EU}, .20) = 30.01\%$ versus $C(\mathrm{EU}, \mathrm{EUP}, .20) = 4.06\%$. For this reason we have substituted T_{EUP} for T_{EU} in the unequal-sample-size case.

The right panels in Figure 1 illustrate the essence of the unequal-sample-size results, namely that the ordering of the S_A for the different procedures is not as consistent as in the equal-sample-size case. Still, the inconsistencies do not change the overall impressions of Section 3.1. Although T_{FE} is not uniformly conservative with respect to T_{EUP}, it is nearly so, with S_{FE} exceeding S_{EUP} at only 4% of the examined points and by more than 10% of the nominal size at about 2% of the points. The size of T_{AU} exceeds the nominal size slightly more frequently than in the equal-sample-size case, but by more than 5% at only 2% of the examined points. On the other hand, S_{AU} exceeds S_{EUP} at 60% of the examined points and by more than 5% of the nominal size at more than one-third of the points. In contrast to the equal-sample-size case, the true size of T_{LE} exceeds the nominal size more frequently than T_{AU}. It is still clearly outperformed by T_{AU}, but not with the consistency seen in the equal-sample-size case. By either the minimax or maximin criterion, T_{AU} remains the preferred procedure.

In the unequal-sample-size case, T_{YC} is distinctly more conservative than T_{FE}, the latter exceeding the former in size by more than 20% of the nominal size at nearly half of the examined points. Similarly, the size of T_{LE} exceeds the size of T_{PC} by the same amount at nearly two-thirds of the examined points; in the unequal-sample-size case, T_{PC} more closely approximates T_{FE}.

3.3 Implications for Power

The aforementioned results would be of academic interest if they could not be translated into meaningful differences in power. In our view, a meaningful difference in power is demonstrated by a reduction in the sample size required to achieve a specific power level against a given alternative, as in the planning of a clinical trial, for example. Consequently, we computed $N_A(p_1, p_2, \alpha, \beta) = \min_n\{S_A(n, n, p_1, p_2, \alpha) : S_A \geq 1 - \beta\}$ for some situations that might pertain in small-scale clinical trials. These data are presented in Table 2 for procedures T_{UC}, T_{AU}, T_{EU}, and T_{FE}, along with the true size of the procedure at the indicated sample size.

Although the savings that are realized are not large in terms of absolute numbers, they are significant on a proportionate basis, because the sample sizes required for

Table 2. Minimum Sample Size (true size) Needed to Achieve Power $1 - \beta$ Against the Specified Alternatives for Various Test Procedures

p_1	p_2	$1 - \beta$	N_{UC} (S_{UC})	N_{AU} (S_{AU})	N_{EU} (S_{EU})	N_{FE} (S_{FE})
			Nominal size α = .05			
.05	.25	.80	44 (.048)	46 (.042)	46 (.022)	55 (.009)
		.90	60 (.046)	61 (.043)	61 (.043)	69 (.013)
.05	.35	.80	24 (.018)	27 (.022)	27 (.022)	31 (.002)
		.90	32 (.032)	33 (.031)	33 (.031)	38 (.003)
.05	.45	.80	16 (.006)	17 (.007)	17 (.007)	20 (.002)
		.90	21 (.013)	21 (.013)	23 (.016)	24 (.004)
.15	.45	.80	34 (.054)	36 (.047)	36 (.047)	41 (.026)
		.90	47 (.055)	47 (.047)	47 (.047)	53 (.026)
.15	.55	.80	21 (.051)	22 (.038)	23 (.037)	25 (.019)
		.90	28 (.053)	29 (.041)	29 (.041)	33 (.020)
.25	.55	.80	40 (.050)	41 (.043)	41 (.043)	48 (.029)
		.90	53 (.053)	55 (.049)	56 (.049)	61 (.030)
.25	.65	.80	23 (.052)	24 (.040)	25 (.042)	29 (.024)
		.90	31 (.051)	32 (.044)	33 (.045)	37 (.027)
			Nominal size α = .10			
.05	.25	.80	34 (.101)	37 (.091)	38 (.092)	45 (.019)
		.90	49 (.119)	50 (.093)	50 (.093)	57 (.028)
.05	.35	.80	19 (.052)	22 (.062)	22 (.062)	25 (.005)
		.90	26 (.079)	28 (.079)	28 (.079)	33 (.009)
.05	.45	.80	12 (.021)	13 (.025)	13 (.025)	17 (.001)
		.90	17 (.043)	19 (.050)	19 (.050)	21 (.002)
.15	.35	.80	57 (.098)	57 (.098)	57 (.089)	65 (.059)
		.90	78 (.098)	78 (.097)	79 (.093)	87 (.064)
.15	.45	.80	27 (.110)	28 (.080)	29 (.081)	34 (.044)
		.90	38 (.098)	38 (.093)	40 (.090)	45 (.053)
.15	.55	.80	17 (.103)	17 (.078)	18 (.078)	22 (.040)
		.90	22 (.108)	23 (.081)	24 (.076)	27 (.038)
.25	.55	.80	32 (.108)	33 (.093)	33 (.080)	39 (.061)
		.90	44 (.106)	45 (.096)	45 (.087)	51 (.060)
.25	.65	.80	17 (.095)	20 (.085)	20 (.069)	23 (.051)
		.90	24 (.099)	27 (.089)	27 (.087)	31 (.055)

NOTE: True sizes are in parentheses.

Fisher's exact test are 10%–20% higher than those required for the more powerful procedures. For a fixed resource, this translates into a 10%–20% increase in the number of trials that can be mounted.

4. DISCUSSION

Although our results are formally constructed from the hypothesis-testing perspective of fixing α and computing the true probability of rejecting H_0, we do not mean to suggest that such an artificial framework is important for any particular value of α. From a practical point of view, one is really concerned with the accuracy of the computed significance level or p value obtained when a particular procedure is applied to a given set of data. Nonetheless, by examining the discrepancies in the true size of a procedure over a range of fixed α values, one can infer that a computed p value falling within that range will differ from the true significance level to an approximately similar degree.

In this inherently discrete setting, no nonrandomized procedure can have a size exactly as specified, except coincidentally. Only Fisher's exact test and the exact unconditional test are guaranteed to be strictly conservative. Although it is clear that the latter is distinctly more powerful, computational considerations might still lead one to prefer Fisher's exact test, which in many cases can be implemented with a hand calculator. A slightly relaxed view, however, allows one to entertain procedures that are not strictly conservative to realize the benefits of computational simplicity and increased power. In particular, the previously given results indicate that the size of the approximate unconditional test is almost always within 5% of the indicated significance level, and usually much closer. Though it probably would not be convenient to program on a hand calculator, it can be implemented with a relatively simple set of Minitab or SAS instructions.

The closeness of the approximate unconditional test to the exact version is not surprising, given the relative flatness of π over the range .05 to .95 noted by Suissa and Shuster (1985). Hence typically $\pi(\hat{p}) \approx \pi(p_{\max})$, even though $\hat{p}$ may be quite far from $p_{\max}$, the point where π is maximized. It is possible that the two tests would be more disparate at more extreme values of the nuisance parameter, though if anything the results in Figure 2 suggest the opposite.

Finally, our results suggest that for any reasonable sample size one will not be led far astray by the simple uncorrected χ^2 statistic. Note that 20% excess in a computed significance level of, say, .04 results in a true significance level of .048. We think most practitioners would not consider this a meaningful difference. Conservatism can virtually be guaranteed by employing a continuity correction, even one only half of the usual Yates correction, but with a commensurate loss of power.

[*Received December 1988. Revised July 1989.*]

REFERENCES

Basu, D. (1977), "On the Elimination of Nuisance Parameters," *Journal of the American Statistical Association,* 72, 355–366.

Fisher, R. A. (1935), "The Logic of Inductive Inference," *Journal of the Royal Statistical Society*, Ser. A, 98, 39–54.

Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions* (2nd ed.), New York: John Wiley.

Grizzle, J. E. (1967), "Continuity Correction in the χ^2-Test for 2×2 Tables," *The American Statistician*, 21, 28–32.

Liddell, D. (1978), "Practical Tests of 2×2 Contingency Tables," *The Statistician*, 25, 295–304.

Pirie, W. R., and Hamdan, M. A. (1972), "Some Revised Continuity Corrections for Discrete Distributions," *Biometrics*, 28, 693–701.

Starmer, C. F., Grizzle, J. E., and Sen, P. K. (1974), "Some Reasons for Not Using the Yates Continuity Correction on 2×2 Contingency Tables," *Journal of the American Statistical Association*, 69, 376–378.

Suissa, S., and Shuster, J. J. (1985), "Exact Unconditional Sample Sizes for the 2×2 Binomial Trial," *Journal of the Royal Statistical Society,* Ser. A, 148, 317–327.

Tocher, K. D. (1950), "Extension of the Neyman–Pearson Theory of Tests to Discontinuous Variates," *Biometrika*, 37, 130–144.

Yates, F. (1934), "Contingency Tables Involving Small Numbers and the χ^2 Test," *Journal of the Royal Statistical Society* (supp.), 1, 217–235.