2. *Exact and Approximate Probabilities for Contingency Table with $n_1 = 20$, $n_2 = 20$, and Random Column Totals*

T	a,c	Exact exceedance probabilities when					Approximate exceedance probabilities based on	
		p = .1	p = .2	p = .3	p = .4	p = .5	T	T_c
10.417	3,13	.00001	.00033	.00090	.00121	.00110	.00125	.00368
10.000	0,8	.00010	.00108	.00157	.00194	.00224	.00157	.00566
10.000	5,15	(same as above because of tie in T)					.00157	.00443
7.025	1,8	.00297	.00917	.00962	.00827	.00727	.00804	.02310
6.667	4,12	.00297	.00921	.01063	.01085	.00949	.00982	.02387
5.227	4,11	.01182	.02259	.02480	.02378	.02045	.02224	.05004
4.444	0,4	.03386	.03372	.03323	.03589	.03857	.03502	.11385
4.286	3,9	.03867	.04933	.04522	.03996	.03927	.03843	.08450
3.956	4,10	.03867	.05022	.05326	.04822	.04253	.04670	.09742
3.243	0,3	.08602	.07007	.06868	.07567	.08075	.07172	.22991
3.137	1,5	.10327	.09020	.07113	.07574	.08075	.07652	.18404
3.135	3,8	.10341	.09930	.08751	.08018	.08127	.07664	.15665
2.849	4,9	.10342	.10252	.10456	.09140	.08423	.09143	.17691
1.111	1,3	.35180	.33677	.30899	.27336	.26973	.29184	.59816
1.026	0,1	.41752	.34584	.34992	.30028	.27683	.31118	1.00000
1.026	5,8	(same as above because of tie in T)					.31118	.49957
.960	6,9	.41752	.34745	.37498	.34071	.30051	.32719	.51363

$p = .5$, and equals .03857, much closer to the uncorrected T estimate .03502 than to the Yates corrected estimate .11385. Table 2 supports Claim 3.

3. WHEN MARGINAL TOTALS ARE RANDOM, T AND T_c PROVIDE DIFFERENT TESTS

The tests indicated by T and T_c are equivalent if and only if for every real number k there is a real number k' such that the sets of contingency tables yielding $T > k$ and $T_c > k'$ are identical. To illustrate we will return to the example cited in Section 1. One sample had no hits out of 20 shots and a second sample had 4 hits out of 20 attempts. If the experiment is conducted at closer range, one might obtain 3 hits from 20 shots with one radar device, and 9 hits out of 20 shots with the other radar device. For purposes of future testing, which range provides a better discriminator between radar tracking units?

The statistic $T = 4.444$ is more extreme for the first set of observations than $T = 4.286$ for the second set, indicating that the longer range tests might provide more information to enable the two tracking devices to be compared. However, results are the opposite if T_c is used. A value of $T_c = 2.976$ for the close range tests is more extreme than $T_c = 2.500$ obtained from the first experiment.

The question is no longer one of choosing between T and T_c to obtain better estimates of the true probability, but rather between T and T_c as a means of ordering discrepancies in observed frequencies. Now T and T_c provide different tests, with different critical regions and different power functions. Claim 3 now is a moot point.

[*Received September 1972. Revised May 1973.*]

REFERENCES

[1] Conover, William J., *Practical Nonparametric Statistics*, New York: John Wiley and Sons, Inc., 1971.

[2] Grizzle, James E., "Continuity Correction in the χ^2-Test for 2 × 2 Tables," *The American Statistician*, 21, No. 4 (October 1967), 28–32.

[3] ———, Appeared in Letters to the Editor, *The American Statistician*, 23, No. 2 (April 1969), 35.

[4] Kendall, Maurice G. and Stuart, Alan, *The Advanced Theory of Statistics, Vol. 2*, 2nd ed., New York: Hafner Publishing Company, 1967.

[5] Kurtz, Thomas E., "A Role of Time-Sharing Computing in Statistical Research," *The American Statistician*, 22, No. 5 (December 1968), 19–21.

[6] Mantel, Nathan and Greenhouse, Samuel W., "What Is the Continuity Correction?" *The American Statistician*, 22, No. 5 (December 1968), 27–30.

[7] Pearson, E.S., "The Choice of a Statistical Test Illustrated on the Interpretation of Data Classed in a 2 × 2 Table," *Biometrika*, 34 (January 1947), 139–67.

[8] Plackett, R.L., "The Continuity Correction in 2 × 2 Tables," *Biometrika*, 51 (December 1964), 327–37.

[9] Yates, Frank, "Contingency Tables Involving Small Numbers and the χ^2 Test," *Journal of the Royal Statistical Society*, Ser. B, Supp. Vol. 1, No. 2 (1934), 217–35.

Comment

C. FRANK STARMER, JAMES E. GRIZZLE and P. K. SEN*

In the discussion of hypothesis testing in 2 × 2 contingency tables, Fisher's exact test is often used as the standard against which competing tests are measured. Statisticians should not be led into a semantic trap by

* C. Frank Starmer is associate professor of computer science and assistant professor of medicine, Department of Medicine, Duke University, Durham, N.C. 27710. James E. Grizzle and P.K. Sen are professors, both at Department of Biostatistics, University of North Carolina, Chapel Hill, N.C. 27514. C. Frank Starmer is recipient of Research Career Development Award 1-K4-HL-70, 102, from the U.S. Public Health Service.

the words "exact test." It is important to know in what sense the "exact test" is exact. We interpret the phrase to mean that it yields the exact probability of observing a result identical to a more extreme probability under the assumption that a particular 2 × 2 table was generated by sampling from a four-variable hypergeometric distribution. It does not give a test with predetermined significance level exactly α. In fact because of the discreteness of the hypergeometric distribution, the observed significance level α^* has the property $\alpha^* \leq \alpha$, where α^* depends on the marginal frequencies (held fixed) and the exact test is therefore always conservative. There seems to be no good reason to use the exact test as the standard of comparison for competing tests.

The result of Tocher [1] shows that the exact test, supplemented by randomization to achieve the desired significance α, is the most powerful test against one-sided alternatives when both, one or no margin totals are fixed in advance. Therefore, the randomized exact test should be the standard to which competing tests are compared.

Even though most statisticians would not use the randomized test in practice, it could be used for judging the value of competing tests. Thus we could search for the best approximation to the most powerful test that does not require the undesirable feature of randomization to achieve the desired significance level.

We have performed a few more simulations to investigate the behavior of T, T_c, and the exact test (E). In addition, we have also made the randomized test (R) which we can use as the basis of comparison. We have tested against a two-sided alternative, and thus our test is not in actuality the most powerful test, but its use conforms to common practice and it should not be far off the mark.

The simulations shown in Tables 1 and 2 show clearly

1. Comparing Two Binomials–Type 1 Errors in 2,000 Simulations

Test	N^a 10	20	30	40	50	60	70	80	90	100
					P = .1					
T	20	77	112	110	94	100	105	96	89	104
T_c	3	11	18	31	38	29	49	52	44	47
E	3	28	18	37	38	39	49	62	55	58
R	90	105	94	96	104	94	102	105	83	95
					P = .2					
T	68	96	97	99	115	102	117	94	83	123
T_c	20	37	40	48	61	52	64	57	48	75
E	20	42	40	62	61	52	69	60	48	77
R	98	94	108	99	108	96	108	96	78	117
					P = .3					
T	91	99	85	105	97	108	100	113	95	112
T_c	28	39	46	56	67	61	59	74	65	86
E	28	46	46	59	67	61	62	74	62	86
R	113	93	84	108	99	98	92	111	92	110
					P = .4					
T	89	111	80	114	113	105	88	80	101	100
T_c	26	38	45	56	66	56	59	57	77	72
E	26	47	45	56	66	56	59	57	77	72
R	105	105	86	115	105	103	99	84	98	94
					P = .5					
T	99	82	98	126	119	101	98	98	88	111
T_c	31	27	58	73	71	61	77	67	62	83
E	31	38	58	73	71	61	77	67	62	83
R	114	98	96	102	103	95	103	96	97	96

[a] Sample size for each binomial distribution.

2. Test of Independence Type 1 Error in 2,000 Simulations in 2 × 2 Table

Test	N^a 20	40	60	80	100	120	140	160	180	200
	a = .56	b = .24	c = .14	d = .06						
T	67	98	117	107	117	80	125	93	101	100
T_c	12	43	36	54	63	62	75	59	58	63
E	8	40	50	54	60	47	79	59	60	69
R	98	113	111	111	116	91	125	98	99	101
	a =.42	b = .28	c = .18	d = .12						
T	94	93	109	106	100	98	90	108	109	99
T_c	21	30	55	59	55	58	57	75	75	70
E	27	32	60	62	60	58	58	77	75	70
R	85	97	100	100	98	90	86	106	109	98
	a = .2	b = .2	c = .3	d = .3						
T	102	105	92	106	120	96	94	102	100	87
T_c	33	42	39	61	65	57	63	74	65	61
E	40	42	39	62	66	57	64	75	66	61
R	99	102	88	99	112	93	94	100	102	83
	a = .54	b = .36	c = .06	d = .04						
T	54	84	86	91	98	100	114	93	90	100
T_c	2	13	21	31	32	71	57	45	47	59
E	2	14	23	37	37	40	57	50	47	59
R	88	110	92	89	97	105	111	86	91	96

[a] Sample size for the single multinomial distribution in the table.

the discrepancy between R and the exact test E. Furthermore, they show that the conservativeness of the exact test and its approximation T_c persist for what would commonly be called large samples. In contrast T achieves a closer approximation to R than T_c or E for moderate or large sample sizes. Even though T is overly conservative for small sample sizes, it is always substantially closer to R in its performance under H_0 than the other competitors evaluated here.

[*Received September 1973.*]

REFERENCE

[1] Tocher, K.D., "Extension of the Neyman-Pearson Theory of Tests to Discontinuous Variates," *Biometrika*, 37 (1950), 130–44

Comment and a Suggestion

NATHAN MANTEL*

In commenting on an earlier version of this article I suggested that Conover's example could be employed to advantage to demonstrate the propriety of using the continuity correction. The reverse demonstration by Conover related not to proper use but rather to misuse of the continuity correction. A note explaining proper use of the continuity correction for situations like the one Conover brought up in his earlier version, I felt, would be a useful contribution—such explanation I will attempt, using one of the examples of his current version.

My thinking here is that in using the continuity correction I should try to parallel the computations I would make if I were estimating tail or class-interval probabilities for a normal distribution with known mean and variance. Use of continuity-corrected chi square for a 2×2 table as displayed by Conover is equivalent to considering the cell frequency a to be normally distributed with expectation n_1c_1/N, variance $n_1n_2c_1c_2/N^3$, but with grouping into class intervals with terminals midway between the integers, e.g., $a = 5$ corresponds to the interval 4.5–5.5. Thus if I wished to get the probability that a is at least as great as 5, I would get the tail area to the right of 4.5 for the distribution $N[E(a), \text{Var}(a)]$—to get the probability that a is exactly 5, I need only subtract from this the tail area to the right of 5.5.

1. Certain Exact and Approximate Probabilities as Given by Conover

T (Uncorrected chi square)	a	Exact probability	Approximate probability based on T_c
9.378	7	0.00270	0.00815
7.677	0	0.00894	0.01857
4.969	6	0.03950	0.06992
3.754	1	0.09480	0.12832
1.948	5	0.22578	0.32752
1.219	2	0.41242	0.49179
0.316	4	0.68893	0.88406
0.073	3	1.00000	1.00000

For Conover's Table 2 example with $n_1 = 19$, $n_2 = 21$, $c_1 = 7$, $c_2 = 33$, $N = 40$ I obtain $E(a) = 3.325$, Var $(a) = 1.44014$, S.D. $(a) = 1.20006$. I can simplify the mechanics of getting tail area differences by taking advantage of the results Conover shows in his Table 1, Part B—this requires only that I interpret the T_c probabilities shown as two-tail probabilities. At the same time the exact cumulative probabilities shown, which Conover treats as ideally correct, can be converted into exact individual term probabilities. Table 1 shows the necessary quantities appearing in Conover's Table 1, Part B.

Exact and approximate individual term probabilities can be derived from this as I next show in Table 2.

2. Exact Individual Term Probabilities and Their Estimated Values Based on Continuity-Corrected Chi Square

a	Exact probability	Continuity-corrected estimate
7	0.00270	0.00815/2 = 0.00408
6	0.03950 − 0.00894 = 0.03056	(0.06992 − 0.00815)/2 = 0.03088
5	0.22578 − 0.09480 = 0.13098	(0.32752 − 0.06992)/2 = 0.12880
4	0.68893 − 0.41242 = 0.27651	(0.88406 − 0.32752)/2 = 0.27827
3	1.00000 − 0.68893 = 0.31107	1 − (0.88406 + 0.49179)/2 = 0.31208
2	0.41242 − 0.22578 = 0.18664	(0.49179 − 0.12832)/2 = 0.18174
1	0.09480 − 0.03950 = 0.05530	(0.12832 − 0.01857)/2 = 0.05488
0	0.00894 − 0.00270 = 0.00624	0.01857 /2 = 0.00928

The excellent agreement between exact individual term probabilities and those based on the use of continuity-corrected chi square, except perhaps at the very extremes, is apparent. Thus for one-sided significance testing the use of continuity-corrected chi square should give much the same results as Fisher's exact test. The same should be true for two-sided testing, but some simple precautions must be taken to conduct such tests properly. The exact one-sided probability for an outcome of 6 or more is Prob $(a = 6)$ + Prob $(a = 7) = 0.03056 + 0.00270 = 0.03326$ while the continuity-corrected chi-square estimate is $0.03088 + 0.00408 = 0.03496$, reason-

* Nathan Mantel, formerly senior mathematical statistician with the National Cancer Institute, is presently research professor, Biostatistics Center, George Washington University, 7979 Old Georgetown Rd., Bethesda, Md. 20014. This research was partially supported by Public Health Service Research Grant No. CA-15686 from the National Cancer Institute.