EMPIRICAL POWER FUNCTIONS FOR NONPARAMETRIC TWO-SAMPLE TESTS FOR SMALL SAMPLES[1]

By D. Teichroew

University of California, Los Angeles

A number of nonparametric or "distribution-free" tests have been proposed [2] for the problem of testing the hypothesis that two samples come from the same population. Some of these tests are functions of the "ranking" sequences which are obtained by arranging the observations from both populations from smallest to largest, and then replacing the observations from the first population by zero and the observations from the second population by one.

If a sample of m is taken from the first population and a sample of n from the second, there are $(m + n)!/m!\,n!$ different sequences. (Since we are dealing with continuous distributions, the probability of a tie is zero. The computer automatically carried about eight decimal places for each normal deviate and therefore there was no need to investigate ties.) If the populations are identical, all sequences have equal probability of occurring. If the populations are not identical, some sequences have a higher probability than others.

The hypothesis can be tested by the use of a critical region. The simplest test consists of selecting, in advance, k particular rankings to be in the critical region. If one of these rankings occurs, the hypothesis will be rejected. The size of the critical region, that is, the probability of rejecting the hypothesis when it is true, is given by

$$\alpha = k \frac{m!\,n!}{(m + n)!}.$$

Non-integer values of k can be obtained by the use of randomized tests.

The problem of which rankings to place in the critical region is settled on the basis of power. Let $P(R_i)$ be the probability of a ranking under an alternative hypothesis. The power, for any alternative, is obtained by summing the $P(R_i)$ for all R_i which belong to the critical region. Therefore, to construct an optimum test, one wants a test which, as α is increased, places the rankings into the critical region in the order of decreasing $P(R_i)$.

Clearly, this test is the optimum test against the alternative, since it maximizes the power. In general it has been possible to construct such optimum tests only against simple alternatives. It is of some interest to determine whether the construction yields optimum tests against classes of alternatives, that is, whether uniformly most powerful tests exist.

Consider the special problem of testing the hypothesis that two samples come

Received November 12, 1954, revised December 15, 1954.

[1] The preparation of this paper was sponsored (in part) by the Office of Naval Research, USN.

TABLE I

$m = 3$		δ								
Ranking	c_1	0.00	.25	.50	.75	1.00	1.25	1.50	2.00	2.50
					$n = 2$					
00011	exact	10.00	6.44	3.94	2.29	1.3	.65	.32	.1	
00011	−1.66	9.0	6.50	4.10	2.25	1.3	.20	.10		
00101	−1.16	11.0	7.30	5.10	3.45	1.1	1.40	.85	.2	
01001	−.66	11.3	8.00	6.30	4.45	2.9	1.95	1.45	.2	
00110	−.50	9.8	8.65	6.65	5.45	4.6	2.30	1.35	.5	
01010	.00	9.3	9.15	8.70	7.40	5.8	4.55	3.20	1.4	
10001	.00	10.0	10.60	9.85	8.10	6.6	5.35	3.70	1.5	
01100	.50	10.3	11.55	11.25	11.15	11.2	8.70	7.20	3.7	
10010	.66	9.0	11.35	12.25	12.00	12.6	11.05	9.20	4.4	
10100	1.16	9.9	12.40	15.00	18.45	19.7	20.90	20.50	18.6	
11000	1.66	10.4	14.50	20.80	27.30	34.2	43.60	52.45	69.5	
11000	exact	10.00	14.77	20.81	28.05	36.24	45.04	54.01	70.72	
					$n = 3$					
000111	exact	5.00	2.859	1.533	.770	.361	.157	.064	.008	.001
000111	−2.11	4.90	2.825	1.675	.875	.400	.175	.075	.025	
001011	−1.71	4.60	2.950	1.875	.750	.400	.275	.100	.025	
010011	−1.27	5.00	3.175	1.825	1.250	.800	.375	.100	.025	.05
001101	−1.27	4.30	3.550	2.075	1.475	.925	.375	.200	.025	
010101	−.83	5.60	3.700	2.375	1.825	1.025	.600	.350	.175	
100011	−.64	5.00	4.275	3.250	2.325	1.750	.800	.475	.175	.05
001110	−.64	5.35	4.275	3.600	2.525	1.475	.950	.550	.225	.025
011001	−.43	5.35	4.225	3.150	2.500	1.675	1.075	.600	.075	.05
010110	−.20	4.40	4.050	3.900	2.675	2.000	1.075	.750	.150	.10
100101	−.20	5.35	4.575	3.650	2.850	1.825	1.525	1.025	.125	.05
101001	.20	4.65	5.050	4.525	3.725	3.375	2.325	1.575	.500	.20
011010	.20	4.40	5.150	4.950	3.675	3.350	2.500	1.775	.675	.15
100110	.43	4.95	5.725	5.400	5.025	4.300	3.225	1.975	.825	.35
011100	.64	5.75	5.500	5.500	6.000	5.375	4.475	3.675	1.725	.70
110001	.64	5.35	5.900	6.350	6.150	4.975	4.150	3.350	1.775	.60
101010	.83	4.65	5.875	6.300	5.975	5.825	5.000	4.100	2.000	.60
101100	1.27	5.15	6.450	7.650	9.050	8.425	8.800	8.400	6.175	3.35
110010	1.27	5.05	6.550	8.300	9.175	9.750	9.850	8.600	5.550	3.15
110100	1.71	5.40	7.825	11.150	13.375	15.850	17.400	18.475	17.350	13.60
111000	2.11	4.80	8.375	12.500	18.800	26.500	35.050	43.850	62.400	76.95
111000	exact	5.00	8.222	12.748	18.697	26.025	34.499	43.721	62.357	78.116

TABLE II

$m = 4$		δ								
Ranking	c_1	0.00	.25	.50	.75	1.00	1.25	1.50	2.00	2.50
					$n = 2$					
000011	exact	6.67	4.02	2.29	1.23	.62		.13	.02	
000011	−1.91	6.28	3.52	2.02	1.26	.70		.12	.04	—
000101	−1.47	6.70	4.86	2.74	1.44	.78		.32	.02	.02
001001	−1.07	6.44	4.68	3.44	2.04	1.14		.26	.06	.06
000110	−.84	5.92	4.74	3.54	2.58	1.68		.48	.14	.02
010001	−.63	6.32	5.46	4.02	2.80	2.14		.60	.26	.04
001010	−.44	6.98	5.52	4.86	3.28	2.26		.98	.44	.04
001100	.00	7.28	6.58	5.58	4.76	3.62		1.48	.50	.20
010010	.00	7.00	6.26	5.44	4.28	3.08		1.64	.54	.24
100001	.00	6.76	6.04	5.28	4.62	3.34		2.00	.70	.18
010100	.44	6.76	7.70	7.46	6.40	5.30		3.22	1.10	.44
100010	.63	6.96	7.82	8.32	7.96	7.04		4.86	2.70	.98
011000	.84	6.64	7.88	9.26	10.04	9.74		7.24	3.90	1.78
100100	1.07	6.52	8.78	9.98	10.70	11.36		9.14	6.64	3.60
101000	1.47	6.94	9.54	13.00	15.96	18.08		19.42	16.92	11.14
110000	1.91	6.50	10.62	15.06	21.88	29.74		48.24	66.04	81.26
110000	exact	6.67	10.45	15.55	21.99	29.66		47.43	65.38	80.11
					$n = 3$					
0000111	exact	2.86	1.49	.72	.32	.13	.05	.02		
0000111	−2.46	2.77	1.50	.82	.36	.15	.06	.02	—	—
0001011	−2.11	3.34	1.68	.84	.50	.14	.08	.05	—	—
0010011	−1.76	2.68	2.06	1.07	.66	.35	.22	.07	—	—
0001101	−1.70	2.80	1.82	.88	.34	.28	.06	—	.01	—
0100011	−1.35	2.85	2.02	1.21	.66	.35	.16	.10	.01	—
0010101	−1.35	3.07	1.62	1.24	.70	.34	.10	.05	—	—
0001110	−1.11	2.78	2.06	1.35	.78	.34	.14	.05	.01	—
0011001	−1.00	2.84	2.14	1.48	.90	.62	.28	.08	.02	—
0100101	−.94	2.52	2.02	1.37	.98	.65	.22	.10	.04	—
0010110	−.76	3.32	2.58	1.68	1.18	.58	.34	.15	—	—
1000011	−.76	2.90	2.26	1.88	1.20	.55	.36	.22	.08	.05
0101001	−.59	2.70	2.04	1.64	1.06	.77	.50	.27	.01	—
0011010	−.41	2.61	2.72	1.75	1.16	.64	.52	.27	.05	—
0100110	−.35	2.80	2.34	1.98	1.38	.94	.52	.31	.05	—
1000101	−.35	2.74	2.26	1.94	1.44	.95	.56	.27	.02	.05
0110001	−.24	3.01	2.68	1.92	1.48	.97	.76	.35	.15	—
1001001	.00	3.08	2.82	2.28	2.04	1.62	.74	.62	.10	—
0101010	.00	2.57	3.06	2.37	2.02	1.50	.76	.48	.12	—
0011100	.00	3.00	2.64	2.51	2.10	1.51	1.00	.45	.12	.10
1000110	.24	2.91	2.90	2.74	2.24	1.60	1.24	.80	.20	—

TABLE II—*continued*

$m = 4$		δ								
Ranking	c_1	0.00	.25	.50	.75	1.00	1.25	1.50	2.00	2.50
					$n = 3$—*continued*					
1010001	.35	2.77	3.16	2.82	2.38	1.74	1.36	.88	.28	.05
0110010	.35	2.85	2.70	3.15	2.52	1.95	1.60	.97	.27	.10
0101100	.41	2.75	3.26	2.82	2.82	2.17	1.52	1.11	.30	—
1001010	.59	2.85	2.54	3.14	2.82	2.48	1.82	1.27	.47	.15
1100001	.76	2.57	3.16	3.58	3.70	3.15	2.66	1.71	.91	.20
0110100	.76	3.00	3.36	3.61	3.16	2.80	2.32	1.88	.77	.25
1010010	.94	3.22	3.68	3.81	4.20	3.52	3.04	2.08	1.10	.25
1001100	1.00	2.75	3.54	3.81	3.80	3.64	3.04	2.44	1.32	.25
0111000	1.11	2.61	3.46	4.15	4.44	4.32	4.14	3.61	2.20	.95
1010100	1.35	2.87	3.84	3.80	5.00	5.31	5.08	4.17	2.42	.90
1100010	1.35	2.92	4.04	4.82	5.28	5.40	5.22	4.50	2.52	1.20
1011000	1.70	2.67	4.48	5.87	6.96	7.98	8.26	7.91	6.27	4.20
1100100	1.76	3.04	4.28	5.57	6.86	8.85	9.76	9.25	6.58	4.25
1101000	2.11	2.74	3.90	7.00	9.42	12.08	14.36	18.00	17.75	13.75
1110000	2.46	2.97	5.38	8.94	13.46	19.60	27.20	35.37	55.71	73.30
1110000	exact	2.86	5.11	8.53	13.39	19.77	27.59	36.57	55.94	73.52

from the same normal population against the alternative that they come from normal populations with the same variance σ^2 but with means which differ by $\sigma\delta$. The optimum test for this problem is, of course, the t test. The optimum rank test for this problem, under the assumption that δ is small, is the c_1 test which was given by Hoeffding [1] and studied in detail by Terry [4].

An interesting practical problem arises as to the behavior of the c_1 test for large values of δ, for example, should one use the c_1 statistic in preference to the rank sum? The question may be phrased in the following way. If the rankings are ordered so that

$$P(R_1) \geqq P(R_2) \geqq P(R_3) \geqq \cdots$$

is true for small values of δ, will this condition hold for all values of δ?

Tables I and II give the empirical frequencies, in percent, of all possible rankings which are obtained when a sample of m from $N(0, 1)$ and a sample of n from $N(-\delta, 1)$ are ranked in order of size and the individual values are replaced by 0 if they come from $N(0, 1)$ and by 1 if they come from $N(-\delta, 1)$. For the two extreme rankings, exact probabilities have been computed [3] and are given for comparison. The tables also give for each ranking the corresponding c_1 value.

The frequencies were obtained as a by-product of a sampling experiment performed on the SWAC. The number of samples computed for the different values of δ varied in the experiment. The values for $n = 2$ in Table I are based on 2000

samples for $\delta = .25, .50, .75, 1.25$, and 1.50, and on 1000 samples for $\delta = 0$, 1.00, and 2.00. The values for $n = 3$ in Table I are based on 4000 samples for each δ except 2.50, in which case 2000 samples were obtained. Table II is based on 5000 samples for $n = 2$ and 7000 samples for $n = 3$, for all values of δ.

The rankings in the tables have been arranged in the order of their c_1 values; in cases where the c_1 value is the same for two or more rankings, the ordering is by increasing probability. The tables show that the probabilities increase essentially monotonically. The deviation from monotonic increase can be accounted for by sampling fluctuation. These tables, therefore, indicate that, at least for some significance levels, it may be possible to construct uniformly most powerful rank order tests for the hypothesis.

The tables may be used to estimate the power function of any rank order test for testing whether two samples from normal populations with the same variance have different means for sample sizes $(m,n) = (3,2), (3,3), (4,2)$, and $(4,3)$. The tests may be randomized or unrandomized and one-sided or two-sided.

I am indebted to I. R. Savage for suggesting the problem and for helpful discussions about the results.

REFERENCES

[1] W. Hoeffding, "Optimum nonparametric tests," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, pp. 83–92 (see p. 88).

[2] I. R. Savage, "Bibliography on nonparametric statistics and related topics," *J. Amer. Stat. Assn.* Vol. 48 (1953), pp. 844–906.

[3] D. Teichroew, "A table giving a probability associated with order statistics in samples from two normal populations which have the same variance but different means," in manuscript.

[4] M. E. Terry, "Some rank order tests which are most powerful against specific parametric alternatives," *Ann. Math. Stat.* Vol. 23 (1952), pp. 346–366.