Applications of Mathematics

Eugenia Stoimenova

Two-sample rank tests based on exceeding observations

Applications of Mathematics, Vol. 52 (2007), No. 4, 345–352

Persistent URL: http://dml.cz/dmlcz/134680

TWO-SAMPLE RANK TESTS BASED ON EXCEEDING OBSERVATIONS

Eugenia Stoimenova, Sofia

(Received January 1, 2006, in revised version April 24, 2006)

Abstract. Simple rank statistics are used to test that two samples come from the same distribution. Šidák's E-test (Apl. Mat. *22* (1977), 166–175) is based on the number of observations from one sample that exceed all observations from the other sample. A similar test statistic is defined in Ann. Inst. Stat. Math. *52* (1970), 255–266. We study asymptotic behavior of the moments of both statistics.

Keywords: location problem, E-test statistic, M-test statistic

MSC 2000: 62G10, 62G20

1. Introduction

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be random samples from distributions F and G, respectively. We wish to test the hypothesis H_0 that F and G are identical against the alternative that $F(x) \geqslant G(x)$, with strict inequality for some x.

Rank statistics based on the number of observations from one sample which exceed all observations from the other sample give rise to quick and easy tests which are suitable for testing that two samples come from the same distribution. The most popular of tests of this type are Haga's test [1] and Šidák's E-test [4]. Some other tests based on exceeding observations are discussed in [2], [3], [5]. Recently a new test statistic, M-statistic, is proposed in [6] for the two sample problem. In this paper we study the asymptotic behavior of the moments of M-statistic and use the results for deriving analogous properties of E-statistic.

The tests based on exceeding observations can be described with reference to the same basic situation. The notation is adapted from Hájek and Šidák (1967). We define A and B' to be the number of observations among $X_1, \ldots, X_m$ larger than $\max\limits_{1 \leqslant j \leqslant n} Y_j$, or smaller than $\min\limits_{1 \leqslant j \leqslant n} Y_j$, respectively, and A' and B to be the number

of observations among $Y_1, \ldots, Y_n$ larger than $\max\limits_{1 \leqslant i \leqslant m} X_i$, or smaller than $\min\limits_{1 \leqslant i \leqslant m} X_i$, respectively. Clearly, only one of the numbers A and A' (or B and B') is positive, while the other must be zero.

With this notation, the E-statistic for testing H_0 is defined by

$$E = \min(A, B) - \min(A', B'),$$

and the M-statistic is defined by

$$M = \max\{m - A, n - B\}.$$

These statistics are non-linear rank statistics and are not asymptotically normally distributed. In Section 2 we prove the asymptotic distribution of M statistic for largest values. Some intermediate results are given there. In Section 3 we derive the asymptotic distribution of the mean and the variance of M under the null hypothesis.

2. Definitions and presentations

Suppose that the notation is chosen so that $m < n$. The exact distributions of E and M statistics under H_0 are presented in the corresponding papers as follows:

$$P(E \geqslant k) = P(E \leqslant -k) = \binom{m+n}{m}^{-1} \binom{m+n-2k}{m-k}, \quad k = 1, \ldots, m,$$

$$P(M = k) = \begin{cases} \binom{m+n}{m}^{-1} \binom{2k-2}{k-1} \dfrac{3k-2}{k}, & \text{for } 1 \leqslant k \leqslant m; \\ \binom{m+n}{m}^{-1} \binom{m+k-1}{m-1}, & \text{for } m < k \leqslant n. \end{cases} \tag{1}$$

For small sample sizes the above distributions are easily enumerated. For large sample sizes the vast majority of the mass is above $\min(m, n)$. The limit distribution of the upper tail of E-statistic is obtained in [4]. The next theorem gives the limit distribution of the largest values of M-statistic.

2.1. Limit distribution of the M-statistic

Theorem 1. *Let $m, n \to \infty$ and $m/n \to \lambda$ $(0 < \lambda < 1)$. Then for $0 \leqslant k \leqslant n - m - 1$*

$$P(M = n - k) \longrightarrow \frac{\lambda}{1+\lambda} \Big(\frac{1}{1+\lambda} \Big)^k.$$

P r o o f. The probability $P(M = n - k)$ in (1) can be expressed

$$P(M = n-k) = \frac{m!n!(m-k-1)!}{(m+n)!(m-1)!k!} = \frac{mn(n-1)\dots(n-k+1)}{(m+n)(m+n-1)\dots(m+n-k)}$$
$$= \frac{mn^k\big(1-\frac{k-1}{n}\big)\dots\big(1-\frac{1}{n}\big)}{(m+n)^{k+1}\big(1-\frac{k}{m+n}\big)\dots\big(1-\frac{1}{m+n}\big)}$$

which obviously tends to $\lambda(1+\lambda)^{-1}(1+\lambda)^{-k}$ as $m, n \to \infty$ and $m/n \to \lambda$. □

Corollary 1. *The probability mass concentrated in the maximum value of M is asymptotically equivalent to $\lambda(1+\lambda)^{-1}$.*

The approximate number of points in the lower tail $\mathbf{P}\{M \leqslant k\} \leqslant \alpha$ can be calculated using Theorem 1. The approximation depends on the ratios of m and n and on the size of the second sample n as well.

The first columns in Tabs. 1 and 2 contain different ratios of m and n ($0 < \lambda < 1$), and the first rows contain some large values for n. Similar tables are calculated in [6] using the exact distribution for $n = 1, \dots, 25$; $m = 1, \dots, n$; $\alpha = 0.01, 0.05$. A comparison between the exact distribution and the approximate values can be seen for $n = 25$.

$\lambda \setminus n$	25	40	45	50	55	60	65	70	75	80	90	100	110	120	150	200
0.10	$*$	10	15	20	25	30	35	40	45	50	60	70	80	90	120	170
0.14	3	18	23	28	33	38	43	48	53	58	68	78	88	98	128	178
0.18	8	23	28	33	38	43	48	53	58	63	73	83	93	103	133	183
0.22	11	26	31	36	41	46	51	56	61	66	76	86	96	106	136	186
0.26	13	28	33	38	43	48	53	58	63	68	78	88	98	108	138	188
0.30	15	30	35	40	45	50	55	60	65	70	80	90	100	110	140	190
0.34	16	31	36	41	46	51	56	61	66	71	81	91	101	111	141	191
0.38	17	32	37	42	47	52	57	62	67	72	82	92	102	112	142	192
0.42	17	32	37	42	47	52	57	62	67	72	82	92	102	112	142	192
0.46	18	33	38	43	48	53	58	63	68	73	83	93	103	113	143	193
0.50	19	34	39	44	49	54	59	64	69	74	84	94	104	114	144	194
0.54	19	34	39	44	49	54	59	64	69	74	84	94	104	114	144	194
0.58	19	34	39	44	49	54	59	64	69	74	84	94	104	114	144	194
0.62	20	35	40	45	50	55	60	65	70	75	85	95	105	115	145	195
0.66	20	35	40	45	50	55	60	65	70	75	85	95	105	115	145	195
0.70	20	35	40	45	50	55	60	65	70	75	85	95	105	115	145	195
0.74	21	36	41	46	51	56	61	66	71	76	86	96	106	116	146	196
0.78	21	36	41	46	51	56	61	66	71	76	86	96	106	116	146	196
0.82	21	36	41	46	51	56	61	66	71	76	86	96	106	116	146	196
0.86	21	36	41	46	51	56	61	66	71	76	86	96	106	116	146	196
0.90	21	36	41	46	51	56	61	66	71	76	86	96	106	116	146	196
0.94	21	36	41	46	51	56	61	66	71	76	86	96	106	116	146	196
0.98	22	37	42	47	52	57	62	67	72	77	87	97	107	117	147	197

Table 1. Approximate number of points in the lower tail $\mathbf{P}\{M \leqslant k\} \leqslant 0.05$ using Theorem 1.

$\lambda \setminus n$	25	40	45	50	55	60	65	70	75	80	90	100	110	120	150	200
0.10		*	*	3	8	13	18	23	28	33	43	53	63	73	103	153
0.14	*	6	11	16	21	26	31	36	41	46	56	66	76	86	116	166
0.18	*	13	18	23	28	33	38	43	48	53	63	73	83	93	123	173
0.22	3	18	23	28	33	38	43	48	53	58	68	78	88	98	128	178
0.26	6	21	26	31	36	41	46	51	56	61	71	81	91	101	131	181
0.30	8	23	28	33	38	43	48	53	58	63	73	83	93	103	133	183
0.34	10	25	30	35	40	45	50	55	60	65	75	85	95	105	135	185
0.38	12	27	32	37	42	47	52	57	62	67	77	87	97	107	137	187
0.42	13	28	33	38	43	48	53	58	63	68	78	88	98	108	138	188
0.46	14	29	34	39	44	49	54	59	64	69	79	89	99	109	139	189
0.50	15	30	35	40	45	50	55	60	65	70	80	90	100	110	140	190
0.54	15	30	35	40	45	50	55	60	65	70	80	90	100	110	140	190
0.58	16	31	36	41	46	51	56	61	66	71	81	91	101	111	141	191
0.62	16	31	36	41	46	51	56	61	66	71	81	91	101	111	141	191
0.66	17	32	37	42	47	52	57	62	67	72	82	92	102	112	142	192
0.70	17	32	37	42	47	52	57	62	67	72	82	92	102	112	142	192
0.74	18	33	38	43	48	53	58	63	68	73	83	93	103	113	143	193
0.78	18	33	38	43	48	53	58	63	68	73	83	93	103	113	143	193
0.82	18	33	38	43	48	53	58	63	68	73	83	93	103	113	143	193
0.86	19	34	39	44	49	54	59	64	69	74	84	94	104	114	144	194
0.90	19	34	39	44	49	54	59	64	69	74	84	94	104	114	144	194
0.94	19	34	39	44	49	54	59	64	69	74	84	94	104	114	144	194
0.98	19	34	39	44	49	54	59	64	69	74	84	94	104	114	144	194

Table 2. Approximate number of points in the lower tail $\mathbf{P}\{M \leqslant k\} \leqslant 0.01$ using Theorem 1.

Since for large sample sizes the vast majority of the mass is above $\min(m, n)$, Theorem 1 can not be used for equal sample sizes. Analogous approximation for $m = n$ gives

$$P(M = m - k) \longrightarrow \frac{3}{4}\Big(\frac{1}{4}\Big)^k$$

as $m \to \infty$ and $k = 0, \ldots, m - 1$.

Here and further we suppose that $m = n$. Then the expectations of M_n and M_n^2, respectively, under H_0 are represented as follows:

$$\mathbb{E}(M_n) = n - \binom{2n}{n}^{-1} \sum_{k=0}^{n-1} \binom{2k}{k}, \tag{2}$$

$$\mathbb{E}(M_n^2) = n(n-1) - \binom{2n}{n}^{-1} \sum_{k=0}^{n-1} \binom{2k}{k}(k+1). \tag{3}$$

The proof of (2) is via direct comparison with $\mathbf{E}M_n$ computed from the distribution (1) and defined by

$$\mathbb{E}(M_n) = \sum_{k=1}^{n} \binom{2n}{n}^{-1} \binom{2k-2}{k-1} \frac{3k-2}{k}\, k.$$

Straightforward induction on n gives the result. The proof of (3) is via direct comparison with $\mathbb{E}(M_n^2)$ and induction on n.

2.2. Some intermediate results

Proposition 1. *Define the sequences*

$$q_n = \frac{\mathbb{E}(M_n)}{n},$$
$$q'_n = \mathbb{E}(M_n) - n,$$

and

$$q''_n = n\Big[\mathbb{E}(M_n) - n + \frac{1}{3}\Big], \quad n \geqslant 1.$$

1) *Then q_n, q'_n and q''_n satisfy the recurrence relations*

$$q_{n+1} = \frac{n}{4n+2}q_n + \frac{3n+1}{4n+2}, \tag{4}$$

$$q'_{n+1} = \frac{n+1}{4n+2}q'_n - \frac{n+1}{4n+2}, \tag{5}$$

and

$$q''_{n+1} = \frac{(n+1)^2}{n(4n+2)}q''_n - \frac{n+1}{3(2n+1)}. \tag{6}$$

2) *We claim that q_n, q'_n and q''_n converge to finite limits as $n \to \infty$, and*

$$\lim_{n\to\infty} q_n = 1, \quad \lim_{n\to\infty} q'_n = -\frac{1}{3}, \quad \textit{and} \quad \lim_{n\to\infty} q''_n = -\frac{2}{9}. \tag{7}$$

P r o o f. 1) The recurrence relations (4)–(6) follow immediately from (2) and simple algebra.

2) The proof of convergence of q_n, q'_n and q''_n, which we omit, is analogous to the proof for the sequence t_n in Theorem 2. The limits then follow from the recurrence relations. □

Proposition 2. *Define the sequences*

$$y_n = \frac{\mathbb{E}(M_n^2)}{n},$$
$$y'_n = \frac{1}{n}[\mathbb{E}(M_n^2) - n^2],$$

and

$$y''_n = \Big[\mathbb{E}(M_n^2) - n^2 + \frac{2}{3}n\Big], \quad n \geqslant 1.$$

1) *Then* y_n, y_n' *and* y_n'' *satisfy the recurrence relations*

$$y_{n+1} = \frac{n}{4n+2} y_n + \frac{(3n+1)(n+1)}{4n+2}, \tag{8}$$

$$y_{n+1}' = \frac{n}{4n+2} y_n' - \frac{n+1}{4n+2}, \tag{9}$$

and

$$y_{n+1}'' = \frac{n+1}{2n+1} y_n'' - \frac{n+1}{3(2n+1)}. \tag{10}$$

2) *We claim that* y_n, y_n' *and* y_n'' *converge to finite limits as* $n \to \infty$, *and*

$$\lim_{n\to\infty} y_n = 1, \quad \lim_{n\to\infty} y_n' = -\frac{2}{3}, \quad \text{and} \quad \lim_{n\to\infty} y_n'' = -\frac{2}{9}. \tag{11}$$

P r o o f. 1) The recurrence relations (8)–(10) follow from (3) and simple algebra.

2) The proof of convergence of y_n, y_n' and y_n'' is analogous to the proof for t_n in Theorem 2. The limits (11) follow from the recurrence relations. □

3. Asymptotic behavior of the moments

3.1. Expectation of the M-statistic

As a consequence of (7) we obtain the asymptotic behavior of the expectation of the M-statistic.

Theorem 2. *As* $n \to \infty$

$$\mathbb{E}(M_n) = n - \frac{1}{3} - \frac{2}{9n} - \frac{2}{9n^2}(1 + o(1)). \tag{12}$$

P r o o f. Define $t_n = \mathbb{E}(M_n) - n + \frac{1}{3} + \frac{2}{9}n^{-1}$. Utilizing (2) it is easy to check that t_n satisfies the recurrence relation

$$t_n = \frac{n}{4n-2} t_{n-1} - \frac{3n-2}{9(n-1)n(2n-1)}. \tag{13}$$

Now, substitute t_{n-1} by its recurrence relation (13), and repeat substitution until t_1 is reached. Straightforward manipulation of the sums gives

$$t_n = \binom{2n}{n}^{-1} 2t_1 - \binom{2n}{n}^{-1} \sum_{k=1}^{n-1} \binom{2k+2}{k+1} \frac{3k+1}{9k(k+1)(2k+1)}. \tag{14}$$

It is easily seen from (2) that $t_1 = \frac{1}{18}$.

The presentation (14) of t_n implies that $\frac{1}{9}$ is an upper bound for the sequence t_n. To prove the convergence of t_n it is sufficient to show that it is increasing with n. That is,

$$\begin{aligned} t_{n+1} \geqslant t_n &\iff \binom{2n+2}{n+1}^{-1}\frac{1}{9} - \binom{2n+2}{n+1}^{-1}\sum_{k=1}^{n}\binom{2k+2}{k+1}\frac{3k+1}{9k(k+1)(2k+1)} \\ &\qquad \geqslant \binom{2n}{n}^{-1}\frac{1}{9} - \binom{2n}{n}^{-1}\sum_{k=1}^{n-1}\binom{2k+2}{k+1}\frac{3k+1}{9k(k+1)(2k+1)} \\ &\iff \sum_{k=1}^{n-1}\binom{2k+2}{k+1}\frac{3k+1}{9k(k+1)(2k+1)} \geqslant \binom{2n+2}{n+1}\frac{1}{9n(2n+1)}. \end{aligned}$$

The last inequality follows by a simple induction on n. Thus t_n is increasing and hence convergent sequence.

Now, from the recurrence relation (13) we have

$$\lim_{n\to\infty} t_n = \lim_{n\to\infty}\frac{n}{4n-2}\lim_{n\to\infty} t_{n-1} - \lim_{n\to\infty}\frac{3n-2}{9(n-1)n(2n-1)}.$$

Therefore, $\lim\limits_{n\to\infty} t_n = -\frac{2}{9}\lim\limits_{n\to\infty} 1/n^2 = 0$. □

3.2. Variance of the M-statistic

As a consequence of the results (11) we obtain the asymptotic behavior of the variance of the M-statistic.

Theorem 3. *As $n \to \infty$*

$$\operatorname{var}(M_n) = \frac{1}{9} + \frac{1}{9n}(1 + o(1)). \tag{15}$$

P r o o f. Using the definition of the sequence y_n'' and the limits (11) we have

$$\mathbb{E}(M_n^2) = n^2 - \frac{2}{3}n - \frac{2}{9}(1 + o(1)).$$

Then from Proposition 1 and Proposition 2 we have

$$\begin{aligned} \operatorname{var}(M_n) &= \mathbb{E}(M_n^2) - (\mathbb{E}(M_n))^2 \\ &= n^2 - \frac{2}{3}n - \frac{2}{9} - \Big(n - \frac{1}{3} - \frac{2}{9n}\Big)^2 + O\Big(\frac{1}{n^2}\Big) \\ &= \frac{1}{9} + O\Big(\frac{1}{n^2}\Big). \end{aligned}$$

□

3.3. Moments of the E-statistic

It is not hard to relate the moments of the distribution of the E-statistic to those of M-statistic. Note that the distribution of the E-statistic is symmetric about 0 so its expectation under H_0 is 0.

For the second moment we have

$$\mathbb{E}(E_n^2) = 2\sum_{k=1}^{n-1}\binom{2n}{n}^{-1}\left[\binom{2n-2k}{n-k} - \binom{2n-2k-2}{n-k-1}\right]k^2 + 2\binom{2n}{n}^{-1}n^2 \tag{16}$$

$$= 2\binom{2n}{n}^{-1}\sum_{k=0}^{n-1}\binom{2k}{k}(2n-2k-1).$$

Further,

$$\mathbb{E}(E_{n+1}^2) = \frac{n+1}{4n+2}\left[2\binom{2n}{n}^{-1}\sum_{k=0}^{n-1}\binom{2k}{k}(2n-2k-1) + 4\binom{2n}{n}^{-1}\sum_{k=0}^{n-1}\binom{2k}{k} + 2\right].$$

Using (2) it follows that the presentation (16) of $\mathbb{E}(E_n^2)$ satisfies the recurrence relation

$$\mathbb{E}(E_{n+1}^2) = \frac{n+1}{4n+2}[\mathbb{E}(E_n^2) - 4\mathbb{E}(M_n) + 4n + 2]. \tag{17}$$

Now, if $\lim_{n\to\infty}\mathbb{E}(E_n^2) = c$ exists it satisfies the following equation

$$c = \lim_{n\to\infty}\frac{n+1}{4n+2}[c + 4\lim_{n\to\infty}[\mathbb{E}(M_n) - n] + 2].$$

Since $\lim_{n\to\infty}[\mathbb{E}(M_n) - n] = -\frac{1}{3}$ from (11) we find $c = \frac{10}{9}$ and therefore

$$\operatorname{var}(E_n) \sim \frac{10}{9}.$$

References

[1] *T. Haga*: A two-sample rank test on location. Ann. Inst. Stat. Math. *11* (1960), 211–219. zbl

[2] *J. Hájek, Z. Šidák*: Theory of rank tests. Academic Press, Orlando, 1967. zbl

[3] *S. Rosenbaum*: Tables for a nonparametric test of location. Ann. Math. Stat. *25* (1954), 146–150. zbl

[4] *Z. Šidák*: Tables for the two-sample location E-test based on exceeding observations. Apl. Mat. *22* (1977), 166–175. zbl

[5] *Z. Šidák, J. Vondráček*: A simple non-parametric test of the difference in location of two populations. Apl. Mat. *2* (1957), 215–221. zbl

[6] *E. Stoimenova*: Rank tests based on exceeding observations. Ann. Inst. Stat. Math. *52* (2000), 255–266. zbl

Author's address: *E. Stoimenova*, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev str., bl. 8, 1113 Sofia, Bulgaria, e-mail: `jeni@math.bas.bg`.