A TWO-SAMPLE NONPARAMETRIC TEST

Yi-Liang Chen

Abstract : A nonparametric test for two-sample problem is proposed, and its power compared with Kolmogorov-Smirnov test by Monte Carlo experiments.

Key words and phrases : Empirical distribution function; nonparametric test; random walk.

1. Introduction

Let $X_1, X_2, ..., X_m$ *and* $Y_1, Y_2, ..., Y_n$ *be two independent random samples from continuous distributions with distribution functions* F *and* G*, respectively. We consider the null hypothesis*

$$H_0 : F(x) = G(x), \text{ for every } x,$$

against the general alternative

$$H_1 : F(x) \neq G(x), \text{ for at least one } x.$$

One of the most popular tests for this problem is based on the well-known Kolmogorov-Smirnov two-sample test

$$D_{mn} = sup_{-\infty<x<\infty}|F_m(x) - G_n(x)|$$

$$= sup_{1\leq i\leq m+n}|F_m(Z_{(i)}) - G_n(Z_{(i)})|$$

where F_m *and* G_n *denote the empirical distribution functions for the X-sample and Y-sample respectively and* $Z_{(i)}$ *denotes the ith order statistic of the combined X-sample and Y-sample. It is well-known that the test, under the null*

hypothesis, is distribution-free over the class of all continuous distribution functions and consistent against any differences between F and G;however, it is biased. D_{mn} is simply the greatest deviation between the empirical distribution functions, or in the version of random walk, the maximum height of lattice path from the line of height zero. [*Dwass (1967), p.*1046]

A test statistic which uses all the deviation of two empirical distribution functions and not just the largest one is the Cramér-von Mises test statistic

$$W^2_{mn} = \frac{mn}{m+n} \int_{-\infty}^{\infty} (F_m(x) - G_n(x))^2 d(\frac{mF_m(x)+nG_n(x)}{m+n})$$

[*Cramér (1928), von Mises (1931)*]

In practical work, W^2_{mn} was rewritten by Durbin(1973) as the following:

$$W^2_{mn} = \frac{n}{m+n} \sum_{j=1}^{m} (\frac{r_j - j}{n} - \frac{j-\frac{1}{2}}{m})^2 + \frac{n(2m+n)}{12mn(m+n)}$$

here r_j denotes the rank in the combined sample of the jth observation in the first sample.

Rosenblatt (1952) used the method of Cramér-von Mises scheme to D_{mn} given as following:

$$C_{mn} = \frac{mn}{(m+n)^2} \sum_{i=1}^{m+n} (F_m(Z_{(i)}) - G_n(Z_{(i)}))^2$$

Katzenbeisser and Hackl (1986) suggested using the test statistic

$$T = \sharp\{i | F_n(Z_{(i)}) = G_n(Z_{(i)})\}$$

Which is the number of times when two empirical distribution functions are equal. In the version of random walk, it is the number of times that lattice path visits to the line of height zero. However, all their results were restricted to equal sample sizes.

Chen and Hu (1987) suggested the test statistic D which uses the maximum length of sojourns of lattice path. D , defined as follows,

$$D = max_{0 \le i \le t-1}(r_{i+1} - r_i)$$

where $r_1 < r_2 < ... < r_t$

satisfy $F_m(Z_{(r_i)}) = G_n(Z_{(r_i)})$, *for* $i = 1, 2, ..., t$ *and* $r_0 = 0$

They also generalized the test statistic proposed by Katzenbeisser and Hackl to the case of unequal sample sizes.

In section 2 we propose a new statistic R_{max} *for testing the same problem and derive its null distribution and some properties when two sample sizes are equal. The null distribution of* R_{max} *and its properties when two sample sizes are unequal is obtained in section 3. The asymptotic results of* R_{max} *is derived in section 4. A Monte Carlo experiment is performed to compare the power of* R_{max} *and Kolmogorov-Smirnov test and the result is presented in section 5.*

2. Definition and properties of R_{max}

Let $X_1, X_2, ..., X_m$ *and* $Y_1, Y_2, ..., Y_n$ *be two independent random samples from continuous distributions with distribution functions* F *and* G, *respectively. Also we call X-sample is the first sample, Y-sample is the second sample.*

Definition 2.1 : *Let* R_i *denote the rank in the combined sample, equally sample sizes* n, *of the* i*th observation in the first sample. We define the test statistic* R_{max} *as the maximum rank of the first sample.*

i.e. $$R_{max} = max_{1 \le i \le n} R_i$$

Reject H_0 *for small value of* R_{max}. *Under* $H_0 : F(x) = G(x)$ *for all* x, *it is easy to obtain the probability function of* R_{max}.

Theorem 2.1 : *Under* $H_0 : F(x) = G(x)$ *for all* x, *the test statistic* MR *has probability function as following:*

$$P_0(R_{max} = t) = \frac{\binom{t-1}{n-1}}{\binom{2n}{n}}, \text{ for } t = n, n+1, ..., 2n$$

and 0, for $t = 1, 2, ..., n-1$. *We calculated and listed the result of* $P_0(R_{max} \leq t)$ *for sample sizes* $2n$ *in the table 1.*

To obtain an explicit expression for the first two moments of R_{max}, *the following Lemma will be useful.*

Lemma 2.1 : *For all integer values of* N, *we have*

$$(i)\ \sum_{t=m}^{N} \binom{t}{m} = \binom{N+1}{m+1}$$

$$(ii)\ \sum_{t=m}^{N} t\binom{t}{m} = m\binom{N+1}{m+1} + (m+1)\binom{N+1}{m+2}$$

Proof : *(i) By use of*

$$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}$$

and the first term of the summation $\binom{m}{m} = \binom{m+1}{m+1}$

$$(ii)\ \sum_{t=m}^{N} t\binom{t}{m} = m\binom{m}{m} + (m+1)\binom{m+1}{m} + ... + N\binom{N}{m}$$

$$= m[\binom{m}{m} + \binom{m+1}{m} + ... + \binom{N}{m}]$$

$$+ \binom{m+1}{m} + 2\binom{m+2}{m} + ... + (N-m)\binom{N}{m}$$

Using the quantities (i) again.

Applying the Lemma and collecting the corresponding terms, we obtain the next theorem.

Theorem 2.2 : *Under* $H_0 : F(x) = G(x)$ *for all* x, *the first two moments and variance of the test statistic* R_{max} *has the following forms.*

$$E(R_{max}) = \frac{n}{n+1}(2n+1)$$

$$E(R_{max}^2) = \frac{n^2(2n+1)(2n+3)}{(n+1)(n+2)}$$

$$Var(R_{max}) = \frac{n^2(2n+1)}{(n+2)(n+1)^2}$$

3. Unequally sample sizes

Let the sample size of the first sample is m *and the second sample is* n, *where* $N = m + n$. *On the unequally sample sizes, as the same in the section*

2, we define the test statistic R_{max} is the maximum rank of the first sample.

Definition 3.1 : *Let R_i denote the rank in the combined sample of the ith observation in the first sample with two sample sizes m and n respectively. R_{max}, defined as follows,*

$$R_{max} = max_{1 \le i \le m} R_i$$

Reject H_0 for small value of R_{max}. The following theorems are similar in proofs of the corresponding theorems in section 2, hence the proof is omitted.

Theorem 3.1 : *Under $H_0 : F_m(x) = G_n(x)$ for all x, the test statistic R_{max} has probability function as following:*

$$P_0(R_{max} = t) = \frac{\binom{t-1}{m-1}}{\binom{N}{m}}, \text{ for } t = m, m+1, ..., N$$

and 0, for $t = 1, 2, ..., m-1$.

Theorem 3.2 : *Under $H_0 : F_m(x) = G_n(x)$ for all x, the first two moments and variance of the test statistic R_{max} has the following forms.*

$$E(R_{max}) = \frac{m}{m+1}(N+1)$$

$$E(R_{max}^2) = m(N+1)[\frac{m}{m+1} + \frac{N-m}{m+2}]$$

$$Var(R_{max}) = \frac{m(N+1)(N-m)}{(m+2)(m+1)^2}$$

4. Asymptotic results

In this section, asymptotic equivalents for the distribution of R_{max} will be derived by using the distribution of hypergeometric, because we can take two urns contain m and n labeled-balls as two samples with size m and n, and the test statistic R_{max}, the maximum rank of the first sample, is similar as a ball, with the largest number, is drawn from the first urn and balls, whose number larger than that one from the first urn, are drawn from the second urn. Applying the conception, we obtain the following result.

Theorem 4.1 : Restricted equally sample sizes n, we obtain

$$(2n-t+1)P(R_{max}=t)=\frac{\binom{n}{1}\binom{n}{2n-t}}{\binom{2n}{2n-t+1}}$$

To achieve the asymptotic behavior for the distribution of R_{max}, besides theorem 4.1, we use the conception introduced by Feller.{see Vol.1 of 3rd edition, p.194}

Theorem 4.2 : Let, as $n\to\infty, 1-\frac{t}{2n}\to k, h(1-n+\frac{t}{2})\to x$

where $h=\frac{1}{\sqrt{\frac{nk(1-k)}{2}}}$, *then*

$$(2n-t+1)P(R_{max}=t)\to hn(x)$$

where $n(x)$ is the standard normal distribution.

5. Simulation results

In this section we obtain results from Monte Carlo experiments, and compare empirical power of the test based on the Kolmogorov-Smirnov test and R_{max} under the uniform and the normal distribution with equally sample sizes $n = 20$. The power estimates are based on 1000 replications and with a significance level 0.05. The table of the power contains pure shift and dispersion alternatives as well as for simultaneous shift and dispersion alternatives.

In the case of pure shift alternatives, the Y samples were drawn from distributions such that the expectation was k times the standard deviation of the distribution of X sample under the null hypothesis, where $k = 0.5, 1.0, 1.5, 2.0, 2.5$. For pure dispersion alternatives, the Y samples were drawn from distributions such that the ratio of the standard deviation of the Y- and X- sample is 1, 2 and 3.

Table 2 indicates the test based on Kolmogorov-Smirnov test is less powerful than R_{max} in the uniform distribution for all cases considered. In the normal distribution, for pure shift, the test based on R_{max} is less powerful than

Kolmogorov-Smirnov test, but for pure dispersion and simultaneous shift and dispersion, the test based on R_{max} is more powerful than Kolmogorov-Smirnov test.

Table 1.
The value of P(MR≤t) for sample size 2n

t \ n	5	10	15	20
5	0.0040			
6	0.0238			
7	0.0833			
8	0.2222			
9	0.5000			
10	1.0000			
11		0.0001		
12		0.0004		
13		0.0016		
14		0.0055		
15		0.0163		
16		0.0434		
17		0.1053		
18		0.2369		
19		0.5000		
20		1.0000	0.0001	
21			0.0003	
22			0.0010	
23			0.0031	
24			0.0084	
25			0.0210	
26			0.0497	
27			0.1120	
28			0.2413	
29			0.5000	0.0001
30			1.0000	0.0002
31				0.0006
32				0.0016
33				0.0041
34				0.0100
35				0.0235
36				0.0530
37				0.1154
38				0.2436
39				0.5000
40				1.0000

Table 2

Empirical power of the test based on MR and the Kolmogorov-Smirnov test against shift-, dispersion-, and simultaneous shift- and dispersion-alternatives.

α=0.05, n=20

Distribution	Uniform			Normal		
SIG / DEV	1	2	3	1	2	3
0	0.0500	0.8400	0.9410	0.0530	0.4770	0.7730
	0.0547	0.2747	0.5933	0.0519	0.1515	0.3576
0.5	0.4840	0.9410	0.9720	0.2300	0.7190	0.8840
	0.1903	0.3668	0.6377	0.2563	0.2776	0.4466
1.0	0.9090	0.9800	0.9960	0.5240	0.8650	0.9640
	0.6393	0.6069	0.7112	0.7527	0.5853	0.6134
1.5	0.9930	0.9970	0.9960	0.8070	0.9510	0.9900
	0.9482	0.8001	0.8124	0.9789	0.8191	0.7943
2.0	1.0000	1.0000	1.0000	0.9450	0.9890	0.9970
	0.9993	0.9258	0.9004	1.0000	0.9562	0.9172
2.5	1.0000	1.0000	1.0000	0.9890	0.9980	1.0000
	1.0000	0.9908	0.9598	1.0000	0.9934	0.9825

The empirical power of the test based on MR is show in the first row, the second row contains the empirical power of the Kolmogorov-Smirnov test. SIG = σ_Y/σ_X ; DEV = $K\sigma_X$, K=0.5, 1.0, 1.5, 2.0, 2.5

Reference

Cramer, H. (1928) On the composition of elementary errors. Skand. Aktuarietids. 11, 13-74, 141-180.

Durbin, J. (1973) Distribution theory for tests based on the sample distribution functions. Regional conference series 9, Siam, Philadelphia, U.S.A..

Dwass, M. (1967) Simple random walk and order statistic. Ann. Math. statist. 38, 1042-1053.

Feller, Willian. An Introduction to Probability Theory and Applications. Wiley, New York.

Katzenbeisser, W. and Hackl, P. (1986) An alternative to the Kolmogorov-Smirnov two sample test. Commun. statist. Theor. Meth. 15(4), 1163-1177.

Rosenblatt, M. (1952) Limit theorems associated with variants of the von Mises statistic. Ann. Math. Statist. 23, 617-623.

von Mises, R. (1931) Vorlesungen aus dem Gebiete der angewandten Mathematik. Wahrscheinlichkeitsrechnung und ihre anwendung in der Statistik und Theoretischen Physik. Vol. 1, Franz Deutike Leipzig and Vienna.