Electronic Journal of Statistics
Vol. 3 (2009) 625–632
ISSN: 1935-7524
DOI: 10.1214/09-EJS371

Testing for independence: Saddlepoint approximation to associated permutation distributions

Ehab F. Abd-Elfattah*

Ain Shams University, Cairo, Egypt
e-mail: ehab@asunet.shams.edu.eg

Abstract: One of the most popular class of tests for independence between two random variables is the general class of rank statistics which are invariant under permutations. This class contains Spearman's coefficient of rank correlation statistic, Fisher-Yates statistic, weighted Mann statistic and others. Under the null hypothesis of independence these test statistics have a permutation distribution that is usually approximated by using asymptotic normal theory to determine p-values for these tests. In this note we suggest using a saddlepoint approach that is almost exact and needs no simulations in order to calculate the p-value for tests in this class.

Keywords and phrases: Independence tests, linear rank test, permutation distribution, saddlepoint approximation.

Received February 2009.

Contents

1. Introduction

Suppose we observe N independent pairs of random variables (X_1, Y_1), (X_2, Y_2), $\ldots, (X_N, Y_N)$ and we wish to test the null hypothesis H_0 that the two variables X_i and Y_i are independent for each i. Rearrange all N pairs of observations according to the magnitude of their first coordinate into the sequence (X_{d_1}, Y_{d_1}), $(X_{d_2}, Y_{d_2}), \ldots, (X_{d_N}, Y_{d_N})$ in such a way that $X_{d_1} < X_{d_2} < \cdots < X_{d_N}$. Then put R_i equal to the rank of Y_{d_i} among the observations $Y_{d_1}, Y_{d_2}, \ldots, Y_{d_N}$. Under the assumption of independence and assuming no ties, all $N!$ orderings $(R_1, \ldots, R_N)$ are equally likely with probability $1/N!$. If we are willing to assume that the two variables have a positive association, the $\{R_i\}$

*I would like to express my deepest gratitude to Professor Ronald W. Butler whose guidance and encouragement made this work possible.

should reveal an upward trend, with large values tending to occur on the right of the sequence and low values on the left. An appropriate test statistic that reflects this idea is

$$D = \sum_{i=1}^{N} (R_i - i)^2 \tag{1}$$

with small values of D indicating significance.

The statistic D is related to the well known Spearman's coefficient of rank correlation statistic, S_p, with the relation $S_p = 1 - 6D/N(N^2-1)$, see Gibbons and Chakraborti (2003). It is also related to the weighted Mann statistic, D', by $D' = \frac{1}{6}N(N^2-1) - \frac{1}{2}D$.

Expanding (1), D can be written as

$$D = \frac{1}{3}N(N+1)(2N+1) - 2\sum_{i=1}^{N} iR_i$$

which gives an equivalent simple statistic

$$V' = \sum_{i=1}^{N} iR_i \tag{2}$$

Hajek, Sidak and Sen (1999).

The statistic V' is equivalent to a general class of rank statistics whose null distributions are invariant under permutations. This class can be written as

$$S = \sum_{i=1}^{N} f_N(i) f_N(R_i) \tag{3}$$

which contains the Fisher-Yates normal score test with $f_N(i) = EU_N^{(i)}$, where $U_N^{(1)} < U_N^{(2)} < \cdots < U_N^{(N)}$ are an ordered sample of N observations from the standardized normal distribution; the van der Waerden test statistic with $f_N(i) = \Phi^{-1}(\frac{i}{N+1})$, where Φ is the standard normal distribution function; and the quadrant test statistic with $f_N(i) = sign(i - \frac{N+1}{2})$, Hajek, Sidak and Sen (1999).

This paper is concerned with approximating p-values for tests in the class (3) based on their permutation distributions. The approximation is based on the saddlepoint approximation rather than normal approximation or simulation. Generally, the saddlepoint approximations are accurate up to $O(N^{-3/2})$ when considered over sets of bounded central tendency, in contrast to $O(N^{-1/2})$ for the central limit theorem.

The presented method is a useful tool when accuracy and speed are needed. For example, the need to evaluate a very large number of p-values is increasingly common with modern genetic data. Testing the association between haplotype scores and a trait is a basic problem in genomic studies. In this situation, the asymptotic distribution is not easily available since the haplotype scores are not directly observed but are estimated from genotype data. Permutations are typically used to obtain p-values for large number of tests, but these can be computationally infeasible in large problems. For more details see Seaman and Müller-Myhsok (2005), Lin (2005) and Kustra et al. (2008).

Saddlepoint approximations to randomization distributions were introduced by Daniels (1958) and further developed by Robinson (1982) and Davison and Hinkley (1988). Booth and Butler (1990) showed that various randomization and resampling distributions are the same as certain conditional distributions and that the double saddlepoint approximation attains accuracy comparable to the single saddlepoint approach.

Abd-Elfattah and Butler (2007) and Abd-Elfattah and Butler (2009) used the double saddlepoint approximation to calculate the p-values and confidence intervals for two different problems. The first paper treats the two-sample problem when both treatment and control observations are subject to right censoring. The test statistics are the class of linear rank tests. The second paper deals with extensions of the first paper to three or more treatment levels and considers tests of trend. Both of these papers deal with a survival time data, that is a time to event, which usually is subject to censoring. The current paper concerns nonparametric association with regression type data in which each subject has (x, y) values and we are interested in testing independence of X and Y. Bingyi (1998) also used the double saddlepoint approximation to approximate the p-values of the two sample permutation tests.

Our approximation is also a double saddlepoint approximation which requires no simulations. The following lemma reformulates the class (3) to more appropriate simple form to use the double saddlepoint.

Lemma 1. *The class of statistics (3) can be written in an equivalent form as*

$$V = L^T \sum_{i=1}^{N} f_N(i) Z_i \tag{4}$$

where $L^T = (f_N(1), f_N(2), \dots, f_N(N))$, *and* $Z_1, Z_2, \dots, Z_N$ *are* $N \times 1$ *vectors of the form* $Z_{R_i} = \eta_i$, $i = 1, \dots, N$, *where the* $N \times N$ *identity matrix* $I_N = (\eta_1, \eta_2, \dots, \eta_N)$.

Proof. Simple algebra. □

For example, if $R_1 = 2$ is arithmetical rank so that $Z_2 = \eta_1$ and $\sum_{i=1}^{N} iZ_i$ has a 2 in its first component for R_1.

Section 2 presents the saddlepoint approximation approach. A real data example is illustrated in section 3 along with a simulation study to show the performance of the saddlepoint method. An application of the method to the Cuzick (1982) test statistic in case of interval censoring is discussed in section 4.

2. Saddlepoint approximation for tests of independence

Under the null hypothesis H_0 of independence, the permutation distribution of V places a uniform distribution on the set of $N \times 1$ indicator vectors $\{Z_i\}$. This distribution may be constructed from a corresponding set of i.i.d. $N \times 1$ vectors of $Multinomial(1, \theta_1, \theta_2, \ldots, \theta_N)$ indicator vectors $\zeta_1, \zeta_2, \ldots, \zeta_N$. The permutation distribution over all one way designs for which $\sum_{i=1}^{N} Z_i = (1, \ldots, 1)^T$ is constructed from the i.i.d. Multinomial variables as the conditional distribution

$$Z_1, \ldots, Z_N \overset{D}{=} \zeta_1, \ldots, \zeta_N | \sum_{i=1}^{N} \zeta_i = (1, \ldots, 1)_{N\times 1}^T.$$

The dependence in the statistic can be removed by using the $(N-1) \times 1$ vectors Z_i^- and ζ_i^-, the first $N-1$ components of Z_i and ζ_i respectively, then

$$Z_1^-, \ldots, Z_N^- \overset{D}{=} \zeta_1^-, \ldots, \zeta_N^- \mid \sum_{i=1}^{n} \zeta_i^- = (1, \ldots, 1)_{(N-1)\times 1}^T.$$

Then V can be represented in terms of $\{Z_i^-\}$ as

$$V = L_-^T \sum_{i=1}^{N} f_N(i) Z_i^- + Q$$

where $L_-^T = (f_N(1) - f_N(N), \ldots, f_N(N-1) - f_N(N))$ and $Q = f_N(N) \sum_{i=1}^{N} f_N(i)$.

If v_0 is the observed statistic value of V, and $T(\zeta^-) = L_-^T \sum_{i=1}^{N} f_N(i)\zeta_i^- + Q$, then the null distribution of V is

$$\Pr\{V \geq v_0\} = \Pr\left\{T(\zeta^-) \geq v_0 \mid \sum_{i=1}^{N} \zeta_i^- = (1, \ldots, 1)^T\right\}$$

Assuming any probability vector $\{\theta_1, \theta_2, \ldots, \theta_N\}$ for the Multinomial distribution, the conditional distribution of $T(\zeta_1^-, \zeta_2^-, \ldots, \zeta_N^-)$ given the sufficient statistic $\sum_{i=1}^{N} \zeta_i^-$ is the required permutation distribution.

Let P be a random variable with the required permutation distribution and v_0 the observed value of V. The required p-value is defined as the mid-p-value which is $\text{pr}(P > v_0) + \text{pr}(P = v_0)/2 = \text{mid-}p(v_0)$ and is approximated from the Skovgaard (1987) saddlepoint procedure as the conditional tail probability $\Pr\{T(\zeta^-) \geq v_0 \mid \sum_{i=1}^{N} \zeta_i^- = (1, \ldots, 1)^T\}$

The mid-p-value is approximated from the double saddlepoint procedure using the joint cumulant generating function for $(T(\zeta_1^-, \zeta_2^-, \ldots, \zeta_N^-), \sum_{i=1}^{N} \zeta_i^-)$ given by $K(t, s) = \log M(t, s)$ where

$$M(t, s) = \prod_{i=1}^{N} \left\{ \sum_{j=1}^{N-1} \theta_j \exp(s_j + r_{ij} t) + \theta_N \right\}$$

with $s = (s_1, \ldots, s_{N-1})$ and $r_{ij} = f_N(i)(f_N(j) - f_N(N))$. The approximation is

$$\Pr(V \geq v_0) \simeq 1 - \Phi(\hat{w}) - \phi(\hat{w}) \left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}} \right) \quad (5)$$

where

$$\hat{w} = \operatorname{sgn}(\hat{t})\sqrt{2\left[-\{K\left(\hat{t},\hat{s}\right) - \hat{s}^T 1_- - v_0\hat{t}\}\right]} \tag{6}$$
$$\hat{u} = \hat{t}\sqrt{|K''\left(\hat{t},\hat{s}\right)|/|K''_{ss}(0,0)|}.$$

and 1_- is $(N-1)\times 1$ vector of ones. In these expressions, K'' is the $N\times N$ Hessian matrix and K''_{ss} is the $\partial^2/\partial s\partial s^T$ portion at $(0,0)$, see Skovgaard (1987) and Butler (2005). The saddlepoint $\left(\hat{t},\hat{s}\right)$ solves

$$K'_{sj}\left(\hat{t},\hat{s}\right) = \sum_{i=1}^{N} \frac{\exp(\hat{s}_j + r_{ij}\hat{t})}{\left\{\sum_{l=1}^{N-1}\exp(\hat{s}_l + r_{il}\hat{t}) + 1\right\}} = 1, \qquad j = 1,\ldots,N-1$$
$$K'_t\left(\hat{t},\hat{s}\right) = \sum_{i=1}^{N} \frac{\sum_{j=1}^{N-1} r_{ij}\exp\left(\hat{s}_j + r_{ij}\hat{t}\right)}{\left\{\sum_{l=1}^{N-1}\exp(\hat{s}_l + r_{il}\hat{t}) + 1\right\}} + Q = v_0.$$

By using $\theta_i = 1/N$, the denominator saddlepoint equations have an explicit solution as $\hat{s}_0 = 0$ and this simplifies the calculations.

The saddlepoint expression in (5) uses the saddlepoint approximation as if $T(\zeta^-)$, and consequently P, were continuous random variables. In the permutation setting however, P is discrete and not even a lattice distribution for which a continuity correction would be available. The reason that this continuous formula can and should be used is that it provides the most accurate approximation for the mid-p-value. Pierce and Peters (1992), Davison and Wang (2002), and Butler (2005, §6.1.4) discuss reasons for this accuracy.

These calculations can be summarized as follow; based on the test statistic under consideration, we calculate r_{ij} and Q, then solving the saddlepoint equations yields the saddlepoint $\left(\hat{t},\hat{s}\right)$. Substituting the saddlepoint values at (6) gives $\hat{w}$ and $\hat{u}$ and finally we use (5) to get the saddlepoint p-value.

The saddlepoint method requires solving system of N nonlinear saddlepoint equations. The Newton's method has been used through this paper. The IMSL routine DNEQBF is computationally faster. The DNEQBF routine uses Broyden's (1965) update of Newton's method with the finite-difference method to approximate the initial Jacobian matrix. When the sample size increases, this routine will be very useful to solve a large number of nonlinear equations. A comprehensive treatment of methods for solving nonlinear systems of equations can be found in Dennis and Schnabel (1996), also see Burden and Faires (2003) for methods and related softwares.

3. Example and simulation study

Nayak (1988) gives the failure times of both transmission (X) and transmission pumps (Y) on 15 caterpillar tractors as shown in Table 1.

To test the independence of failure times of X and Y, the test statistic (2) is used with $L = (1,\ldots,N)$, and $Q = L_N\sum_{i=1}^{N} R_i = N^2(N+1)/2$. The true

Table 1

Failure times of transmissions by Nayak (1988)

X	1641	5556	5421	3168	1534	6367	9460	6679
	6142	5995	3953	6922	4210	5161	4732	
Y	850	1607	2225	3223	3379	3832	3871	4142
	4300	4789	6310	6311	6378	6449	6949	

Table 2

Performance under simulation for the independence test statistic (2)

		Sad.	Abs. Err.	Abs. Err.	Rel. Abs.	Rel. Abs.
N	λ	Prop.	Sad.	Nor.	Err. Sad.	Err. Nor.
10	0.0	0.944	0.0010	0.0083	0.0048	0.1057
	0.5	0.945	0.0010	0.0081	0.0050	0.0579
30	0.0	0.943	0.0003	0.0022	0.0012	0.0103
	0.5	0.932	0.0003	0.0022	0.0013	0.0103
50	0.0	0.897	0.0003	0.0013	0.0011	0.0065
	0.5	0.903	0.0003	0.0013	0.0013	0.0069
70	0.0	0.840	0.0003	0.0009	0.0012	0.0052
	0.5	0.867	0.0003	0.0009	0.0013	0.0054

(simulated) mid-p-value was calculated by using 10^6 permutations of the computed test statistic. The simulated mid-p-value is then the proportion of such generations exceeding the observed statistic plus half the proportion of times that attain v_0. The p-value of the saddlepoint approach is compared to the normal p-value calculated using the test statistic $(v' - E(v'))/\sqrt{Var(v')}$. The true mid-$p$-value and the saddlepoint approximated p-value were 0.2768 and 0.2763, respectively, while the normal p-value was 0.2693.

A small simulation study has carried out to assess the performance of the saddlepoint method. Consider the general linear model of dependence

$$X_i = X_i' + \lambda e_i, \qquad Y_i = Y_i' + \lambda e_i, \quad i = 1, \ldots, N$$

where all the variables X_i', Y_i' and e_i are mutually independent and their distributions do not depend on i, and λ is a real non-negative parameter. This general dependence model was considered by Hajek, Sidak and Sen (1999) and Cuzick (1982). In this model the null hypothesis H_0 of independence is equivalent to $\lambda = 0$, whereas for $\lambda > 0$ the variables X_i and Y_i are dependent. Data sets are generated from this model using logistic, extreme value and uniform distributions for X_i', Y_i' and e_i respectively. For each value of $\lambda = 0.0, 0.5$ and sample sizes $(10, 30, 50, 70)$, 1000 data sets were generated and the true, saddlepoint and normal p-values were calculated using the test statistic (2). Table 2 shows the proportion of the 1000 data sets for which the saddlepoint p-value was closer, in absolute error, to the true mid-p-value than the normal p-value "Sad. Prop.", "Abs. Err. Sad." is the average absolute error of the saddlepoint p-value from the true mid-p-value, and "Rel. Abs. Err. Sad." is the average relative absolute error of the saddlepoint p-value from the true mid-p-value, and the remaining listings are the same assessments for the normal approximation.

The saddlepoint p-value was more accurate in 90.8% of the overall cases as compared to the normal approximation. The average absolute saddlepoint error was less than 10^{-3} with average relative error typically less than 0.1%.

4. Extension

The problem of testing the independence between two variables under random censoring has attracted the attention of many authors, see O'Brien (1978), Wei (1980), Oakes (1982), and Gieser and Randles (1997). The saddlepoint approach of section 2 is applicable to test statistics that deal with censored data. When one of the two variables is subject to interval censoring, say the first, and the second random variable is observed, Cuzick (1982) presents a linear log-rank test statistic to test the independence of two vectors in the form $\sum_{i=1}^{N} \xi_i R_{2i}$, where $\{\xi_i\}$ are given scores and $\{R_{2i}\}$ are the ranks of the observed values of the second random variable. In the linear form (4), taking $L = (\xi_1, \xi_2, \ldots, \xi_N)$ and $f_N(R_i) = R_{2i}$, the saddlepoint method is applicable. For example, Cuzick gives a survival times for 20 patients for the analysis of the relation between hemoglobin at presentation and survival in some medical clinic. The normal p-value using his asymptotic approach was 0.0505 while the true mid-p-value and the saddlepoint p-value are 0.0516 and 0.0512, respectively. Generally the saddlepoint approximation can be applied to any linear rank test that takes the form (3).

References

Abd-Elfattah, E.F. and Butler, R. (2007). The Weighted Log-Rank Class of Permutation Tests: P-values and Confidence Intervals Using Saddlepoint Methods, *Biometrika* **94**, 3, 543–551. MR2410007

Abd-Elfattah, E.F. and Butler, R. (2009). Log-Rank permutation tests for trend: Saddlepoint p-values and survival rate confidence intervals. *Canadian Journal of Statistics* **37**, 1, 5–16.

Booth, J.G. and Butler, R.W. (1990). Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* **77**, 787–796. MR1086689

Broyden, C.G. (1965). A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation* **19**, 577–593. MR0198670

Burden, R.L. and Faires, J.D. (2003). *Numerical analysis*. 7th edition, Brooks Cole.

Butler, R.W. (2005). *Saddlepoint Approximations with Applications*. Cambridge University Press. MR2357347

Cuzick, J. (1982). Rank tests for association with right censored data. *Biometrika* **69**, 2, 351–364. MR0671973

Daniels, H.E. (1958). Discussion of paper by D.R. Cox. *Journal of Royal Statistical Society* B **20**, 236–238.

Davison, A.C. and Hinkley, D.H. (1988). Saddlepoint approximations in resampling method. *Biometrika* **75**, 3, 417–431. MR0967581

Davison, A.C. and Wang, S. (2002). Saddlepoint approximations as smoothers. *Biometrika* **89**, 933–938. MR1946521

Dennis, J.E. and Schnabel, R.B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, Philadelphia. MR1376139

Gibbons, J.D. and Chakraborti, S. (2003). *Nonparametric statistical inference*. 4th edition, Marcel Dekker, New York. MR2064386

Gieser, P.W. and Randles, R.H. (1997). A nonparametric test of independence between two vectors. *Journal of the American Statistical Association* **92**, 438, 561–567. MR1467849

Hajek, J., Sidak, Z. and Sen, P.K. (1999). *Theory of Rank Tests*. 2nd Ed. Academic Press. MR1680991

Kustra, R., Shi, X., Murdoch, D., Greenwood, C.M. and Rangrej, J. (2008). Efficient p-value estimation in massively parallel testing problems. *Biostatistics* **9**, 4, 601–612.

Lin, D.Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21**, 6, 781–787.

Nayak, T.K. (1988). Testing equality of conditionally independent exponential distributions. *Communications in Statistics: theory and methods* **17,** 807–820. MR0939644

Oakes, D. (1982). A concordance test for independence in the present of censoring. *Biometrics* **38**, 451–455.

O'Brien, P. (1978). A nonparametric test for association with censored data. *Biometrics* **34**, 243–250.

Pierce, D.A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *Journal of Royal Statistical Society* B **54**, 701–737. MR1185218

Robinson, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *Journal of Royal Statistical Society* B **44**, 1, 91–101. MR0655378

Seaman, S.R. and Müller-Myhsok, B. (2005). Rapid simulation of values for product method and multiple-testing adjustment in association studies. *American Journal of Human Genetics* **76**, 399–408.

Skovgaard, I.M. (1987). Saddlepoint expansions for conditional distributions. *Journal of Applied Probability* **24**, 875–87. MR0913828

Wei, L.J. (1980). A generalized Gehan and Gilbert test for paired observations that are subject to arbitrary right censorship. *Journal of the American Statistical Association* **75**, 371, 634–637. MR0590693