DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 1 | Issue 1 Article 4

5-1-2002

Some Locally Most Powerful Rank Tests For Correlation

W. J. Conover

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation

Conover, W. J. (2002) "Some Locally Most Powerful Rank Tests For Correlation," *Journal of Modern Applied Statistical Methods*: Vol. 1 : Iss. 1 , Article 4.
DOI: 10.22237/jmasm/1020254520
Available at: http://digitalcommons.wayne.edu/jmasm/vol1/iss1/4

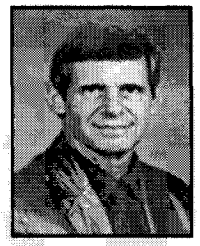
This Invited Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 19-23

Copyright © 2002 JMASM, Inc.
1538 - 9472/02/$30.00

Some Locally Most Powerful Rank Tests For Correlation

W. J. Conover
College of Business Administration
Texas Tech University



Four examples are given to illustrate the ease and practicality of the procedure for finding locally most powerful rank tests for correlation. The first two examples deal with bivariate exponential models. The third example uses the bivariate normal distribution, and the fourth example analyzes the Morgenstern's general correlation model.

Keywords: Bivariate exponential, Bivariate normal, Morgenstern's model, Top-down correlation, Spearman's rho, Normal scores, Savage scores

Introduction

There are two primary difficulties in developing tests with good power properties for testing the null hypothesis of independence between two variables X and Y, based on a bivariate random sample (X_i, Y_i), i = 1, 2, ..., n, against the alternative hypothesis of correlation. One difficulty is in finding a suitable model for the bivariate distribution, and the other is in developing a powerful test for correlation once the model is selected.

The bivariate normal distribution is a convenient model to use for many reasons. The parameter rho is the linear correlation coefficient, so correlation is convenient to address in this model. The most powerful test for correlation is well known, and the locally most powerful rank test (LMPRT) uses Fisher-Yates expected normal scores.

But the bivariate normal distribution does not fit some types of data very well. Therefore other classes of bivariate distributions have been developed in an attempt to find models more appropriate than the bivariate normal distribution, while retaining some of the nice analytical properties found in the bivariate normal distribution. In this paper the general bivariate density function $h(x,y;\Theta)$ is considered, with some fairly general restrictions.

One popular family of bivariate distributions was proposed by Morgenstern (1956). Let F(x) and G(y) be the marginal distribution functions. A bivariate distribution function with those marginals is given by

$$H(x,y;\theta) = F(x)G(y)[1 + \theta\{1 - F(x)\}\{1 - G(y)\}] \qquad (1.1)$$

where θ is the parameter that governs the degree of dependence between the random variables. Farlie (1961) found Spearman's rho to be the optimal correlation coefficient for testing the null hypothesis $\theta=0$ in Morgenstern's model (1.1), and studied the efficiency of less than optimal coefficients. Our derivation of the same result is much simpler and with fewer restrictions on the model.

Morgenstern's model (1.1) was generalized by Farlie (1960) to

$$H(x,y;\theta) = F(x)G(y)[1 + \theta A(x)B(y)] \qquad (1.2)$$

where A(x) an B(y) are bounded functions such that $A(\infty) = B(\infty) = 0$. The model (1.2), and thus (1.1) also, is a special case of the general model studied in this paper.

Konijn (1956) studied correlation tests for the hypothesis $\theta_2 = \theta_3 = 0$ in the model

$$X = \theta_1 W + \theta_2 Z$$
$$Y = \theta_3 W + \theta_4 Z \qquad (1.3)$$

where W and Z are independent random variables. Correlation tests for a similar class of alternatives

$$X = (1 - \theta)U + \theta Z$$
$$Y = (1 - \theta)V + \theta Z \qquad (1.4)$$

where U, V and Z are independent random variables, and the null hypothesis is $\theta = 0$, were investigated by Bhuchongkul (1964). Hájek and Sidák (1967, p.75, or see Hájek et al., 1999, p.77) discuss the nearly identical model

$$X = U + \theta Z$$
$$Y = V + \theta Z \qquad (1.5)$$

These models are more restrictive in their application than the more general model considered in this paper.

In this paper the general bivariate density function $h(x,y;\theta)$ is investigated. A theorem is presented that enables the locally most powerful rank test of $\theta=\theta_0$ to be derived under some fairly general conditions. Four examples are given to illustrate the usefulness of this result.

W. Jay Conover has been publishing papers in nonparametric statistic since1964. He is best known for his books, including *Practical Nonparametric Statistics*, which first appeared in 1971. The third edition was published in 1999. He is a Fellow of the American Statistical Association.

Although the development stops short of finding the efficiencies of the obtained tests, in some cases the tests are well known, and their efficiencies have been studied.

The Locally Most Powerful Rank Test For Correlation

Let (X,Y) have the joint density function $h(x,y;\theta)$ under H_a and the density $h(x,y;\theta_0) = f(x)g(y)$ under H_0, the independence hypothesis, where f(x) and g(y) are the marginal density functions of X and Y respectively. Consider independent copies (X_m,Y_m), m = 1, ..., n of (X,Y), with ranks R_m for X_m, and Q_m for Y_m, where the X's and the Y's are ranked separately. Also define the scores

$$a(i,j;h) = E[\partial \ln \{h(X_n^{(i)},Y_n^{(j)};\theta)\}/\partial\theta \mid_{\theta=\theta 0} \mid H_0] \qquad (2.1)$$

where $X_n^{(i)}$ and $Y_n^{(j)}$ are the i th and j th order statistics in a random sample of size n from f(x) and g(y) respectively.

The following theorem was first proven by Shirahata (1974) in the p-variate case, for general regression alternatives. This is the simplified version for testing correlation in the bivariate case.

Theorem for locally most powerful rank correlation tests. Let J be some open interval around θ_0. If

1. $h(x,y;\theta)/h(x,y;\theta_0)$ exists for $\theta \in J$,
2. $\partial \{h(x,y;\theta)\}/\partial\theta \mid_{\theta=\theta 0}$ exists for $\theta \in J$,
3. $h(x,y;\theta_0) = \lim_{\theta \to \theta 0} h(x,y;\theta)$ exists for $\theta \in J$, and
4. $\lim \theta \to \theta_0 \int\int \mid \partial \{h(x,y;\theta)\}/\partial\theta \mid dx\, dy = \int\int \mid \partial \{h(x,y;\theta)\}/\partial\theta \mid_{\theta=\theta 0} \mid dx\, dy < \infty$,

then the locally ($\delta \to 0$) most powerful rank test of H_0: $\theta = \theta_0$ against H_a: $\theta = \theta_0 + \delta$ is given by the test with the critical region

$$\sum_{m=1}^{n} a(R_m, Q_m; h) > k \qquad (2.2)$$

where k is chosen so the test will have an appropriate size α.

Implementation of the previous theorem to find the scores a(i,j;h) associated with the locally most powerful rank test for correlation involves the following steps.

1. Find the partial derivative of $h(x,y;\theta)$ with respect to θ and set θ equal to θ_0.
2. Divide the result in Step 1 by $h(x,y;\theta_0) = f(x)g(y)$.
3. Substitute $X_n^{(i)}$ for x and $Y_n^{(j)}$ for y in the quotient in Step 2, where $X_n^{(i)}$ and $Y_n^{(j)}$ are the i th and j th order statistics in random samples of size n from f(x) and g(y) respectively.
4. Find the expected value of the random variable in Step 3 under H_0. That is, integrate the product of
   (a) the result of Step 2,
   (b) the density function of the ith order statistic from f(x),
   (c) and the density function of the jth order statistic from g(y), over the entire range of values of X and Y.

Four Examples Of Locally Most Powerful Rank Tests

These four examples show the ease with which the theorem of Section 2 can be applied to obtain locally most powerful rank tests for correlation. In all four examples the resulting test statistic is known, and the literature citations can be consulted to find tables for small sample sizes, and asymptotic approximations for large sample sizes.

The first two examples involve bivariate distributions, where both marginal distributions are exponential. The model in the first example allows only nonnegative correlations, and may be used when the alternative hypothesis is one of positive correlation. The model in the second example allows only nonpositive correlations, and may be used when the alternative hypothesis is one of negative correlation. In both examples, the locally most powerful rank test uses the top-down correlation coefficient of Iman and Conover (1987). The third example involves the bivariate normal distribution, and the fourth example looks at a very general bivariate distribution.

Example 1. Mardia (1970) presented a bivariate exponential distribution

$$h_1(x,y;\theta) = \frac{1}{1-\theta} e^{\frac{-x+y}{1-\theta}} \sum_{r=0}^{\infty} \left[\frac{\theta xy}{(1-\theta)^2}\right]^r /r!r!$$

$$x > 0,\ y > 0,\ 0 \le \theta < 1 \qquad (3.1)$$

which has exponential marginal densities exp(-x) and exp(-y), and which degenerates to the product of those marginal densities exp{-x-y} for $\theta = 0$, representing the case of independence. The correlation coefficient between X and Y is θ. Although this model has standardized exponential distributions, the non-standardized exponential distributions, with a general scale variable, give the same result because the test derived is a function only of ranks within each variable, and the ranks are not changed by a change in the scale variable.

This model first appeared in Mardia (1962) as a special case of a bivariate gamma distribution that appeared in Kibble (1941). It has been attributed to various authors, such as to Downton (1970) by Hawkes (1972) and others, and to Nagao and Kadoya (1971) by Cordova and Rodreguez-Iturbe (1985), Johnson and Kotz (1972), and others. It is widely used as a model for the bivariate

exponential distribution. A parametric test of the null hypothesis of independence is apparently unknown. The locally most powerful rank test is derived in the following.

$$a(i,j;h_1)=E\{(X_n^{(i)}-1)(Y_n^{(j)}-1) \mid H_0\}=(s_n(i)-1)(s_n(j)-1) \quad (3.2)$$

where $s_n(i)$ and $s_n(j)$ are the expected values of order statistics from the exponential distribution. That is, step 1 in the previous section involves finding the derivative of $h_1(x,y;\theta)$ with respect to θ, and setting $\theta = 0$. This gives

$$\frac{\partial}{\partial\theta} h_1(x,y;\theta)\Big|_{\theta=0} = e^{-x-y}(x-1)(y-1) \quad (3.3)$$

The second step is to divide by $f(x)g(y) = e^{-x-y}$, which gives
$(x-1)(y-1)$
In the third step the ith and jth order statistics from the exponential distributions f(x) and g(y) replace x and y respectively. Thus the expected values, in step 4, give the LMPRT scores in (3.2).

The scores in (3.2) are given by the formula

$$s_n(i) = \sum_{j=0}^{i-1} \frac{1}{n-j} \quad (3.4)$$

and are sometimes called Savage scores because they were introduced by Savage (1956). Their use in a rank correlation coefficient

$$r_T = \frac{\sum s_n(R_m)s_n(Q_m) - (\sum s_n(i))^2/n}{\sum [s_n(i)]^2 - (\sum s_n(i))^2/n} = \frac{\sum s_n(R_m)s_n(Q_m) - n}{n - s_n(n)} \quad (3.5)$$

was studied by Iman and Conover (1987), and called the top-down correlation coefficient r_T because of its tendency to emphasize the tail values. That is, items are ranked from least important or lowest, with rank 1, to most important or greatest, with rank n. The top-down correlation coefficient gives more emphasis to the agreement among the most important items, and less emphasis to agreement among the least important items. Spearman's rho and Kendall's tau give equal importance to agreement at all ranks.

Exact quantiles for the null distribution of r_T in a test of independence are given by Iman (1987) for $n \le 14$. Therefore the locally most powerful rank test of H_0: $\theta = 0$ against H_a: $\theta > 0$ in the bivariate exponential distribution given by (3.1) rejects H_0 if and only if $r_T > k$ for a suitably chosen value of k.

Example 2. Gumbel (1960) introduced another bivariate exponential distribution

$$h_2(x,y;\theta) = \{(1+\theta x)(1+\theta y)-\theta\} e^{-x-y-\theta xy} \quad x>0,\ y>0,\ 0\le\theta\le 1 \quad (3.6)$$

with non-positive correlation coefficient:

$$-1 + \int_0^\infty \frac{e^{-y}}{1+\theta y} dy \quad (3.7)$$

Note that the correlation coefficient is zero when $\theta = 0$, and it decreases monotonically as θ increases. Therefore the LMPRT for correlation is also the LMPRT for θ. This distribution degenerates to $\exp\{-x-y\}$ under H_0: $\theta = 0$. This widely known model was studied further by Gumbel (1961) and has been used more recently by Wei (1981) and Barnett (1983). As with the previous model, a parametric test of the null hypothesis of independence is apparently unknown. The locally most powerful rank test is derived in the following.

The optimal scores are again found to be functions of the Savage scores. Specifically the scores are:

$$a(i,j;h_2) = E\{-(X_n^{(i)}-1)(Y_n^{(j)}-1) \mid H_0\} = -(s_n(i)-1)(s_n(j)-1) \quad (3.8)$$

which leads to the locally most powerful rank test that rejects H_0 when $r_T < k$ for some suitably chosen negative number k. Note that the negative value for k is due to the model, which allows only negative correlation in the restricted parameter range for θ.

Example 3. The all-important bivariate normal distribution has density:

$$h_3(x,y;\theta) = (2\pi(1-\theta^2))^{-1/2} \exp\{-(x^2+y^2-2\theta xy)/2\} \quad (3.9)$$

and correlation coefficient θ. The scores for the locally most powerful rank test are given by:

$$a(i,j;h_3) = E(Z_n^{(i)})E(Z_n^{(j)}) \quad (3.10)$$

where $Z_n^{(i)}$ and $Z_n^{(j)}$ are order statistics from the standard normal distribution. These scores are used in the well-known normal scores statistic first given by Fisher and Yates (1957). This derivation of the locally most powerful rank test for the bivariate normal distribution, also given by Shirahata (1974), is much simpler than the previous ones, and uses a more general model than the rather restrictive models (1.3), (1.4) and (1.5).

Example 4. The class of bivariate distributions in introduced by Morgenstern (1956) has the bivariate distribution function:

$$H(x,y;\theta) = F(x)G(y)[1 + \theta\{1 - F(x)\}\{1 - G(y)\}] \quad (3.11)$$

for any marginal distribution functions F(x) and G(y). This model has been extended by Plackett (1965) and often appears in discussions of bivariate distributions (see for

example Mardia, 1970, or Johnson and Kotz, 1972). Due to the unspecified nature of F(x) and G(y) no parametric test is possible. However, rank tests are possible. In fact the locally most powerful rank test is easily derived, as shown in the following.

When H(x,y) is continuous then the density function is

$$h_4(x,y;\theta) = f(x)g(y)[1 + \theta\{1 - 2F(x)\}\{1 - 2G(y)\}] \tag{3.12}$$

which reduces to the independence case f(x)g(y) when $\theta = 0$. This example shows the full power of the method discussed in this paper for finding the locally most powerful rank test for independence. The scores $a(i,j;h_4)$ in this case reduce to

$$\begin{aligned} a(i,j;h_4) &= E\{(2F(X_n^{(i)}) - 1)(2G(Y_n^{(j)}) - 1)\} \\ &= (2E\{U_n^{(i)}\} - 1)(2E\{U_n^{(j)}\} - 1) \end{aligned} \tag{3.13}$$

where $U_n^{(i)}$ and $U_n^{(j)}$ represent order statistics from the uniform distribution on (0,1). These are the scores used in the Spearman rank correlation coefficient, so Spearman's rho is the locally most powerful rank test for correlation for the entire class of Morgenstern distributions, assuming only that the bivariate distributions are continuous. This result was first obtained by Farlie (1961), but this method of proof is much simpler.

Note that $h_4(x,y;\theta)$ is a density function with marginal densities f(x) and g(y) for all density functions f and g. In particular if f and g are exponential density functions, h_4 is another form of a bivariate exponential distribution. In this case the correlation coefficient is $\theta/4$ (Gumbel, 1960) and it varies only within the narrow domain [-.25, .25]. Since the correlation coefficient is a monotonic function of θ, the LMPRT for correlation in this bivariate exponential model uses Spearman's rho, instead of the top-down correlation coefficient of the previous two bivariate exponential models.

Conclusion

Asymptotic normality for the special cases of the test statistic given in the previous section is already known. In general, asymptotic normality under the null hypothesis results from Theorem 3, and under contiguous alternatives from Theorem 4, of Shirahata (1974). Closely related results were also given by Behnan (1971), Ruymgaart et al. (1972), and Ruymgaart (1974).

References

Barnett, V. (1983). Reduced distance measures and transformations in processing multivariate outliers. *Australian Journal of Statistics, 25(1)*, 64-75.

Behnan, K. (1971). Asymptotic optimality and ARE of certain rank-order tests under contiguity. *The Annals of Mathematical Statistics, 42*, 325-329.

Bhuchongkul, S. (1964). A class of nonparametric tests for independence in bivariate populations. *The Annals of Mathematical Statistics, 35*, 138-149.

Cordova, J. R., & Rodriguez-Iturbe, I. (1985). On the probabilistic structure of storm surface runoff. *Water Resources Research, 21*, 755-763.

Downton, F. (1970). Bivariate exponential distributions in reliability theory. *Royal Statistical Society Journal, 32*, 408-417.

Farlie, D. J. G. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika, 47*, 307-323.

Farlie, D. J. G. (1961). The asymptotic efficiency of Daniels' generalized correlation coefficients. *Royal Statistical Society Journal, 23*, 128-142.

Fisher, R. A., & Yates, F. (1957). *Statistical tables for biological, agricultural and medical research.* (3rd ed.). Darien, Conn: Hafner.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association, 55*, 698-707.

Gumbel, E. J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association, 56*, 335-349.

Hájek, J., & Sidák, Z. (1967). *Theory of rank tests.* New York: Academic Press.

Hájek, J., Sidák, Z., & Sen, P.K. (1999). *Theory of rank tests.* (2nd ed.). New York: Academic Press.

Hawkes, A. G. (1972). A bivariate exponential distribution with applications to reliability. *Royal Statistical Society Journal, 34,* 129-133.

Iman, R. L. (1987). Tables of the exact quantiles of the top-down correlation coefficient for n = 3(1)14. *Communications in Statistics, Theory and Methods, 16(5)*, 1513-1540.

Iman, R. L., & Conover, W. J. (1987). A measure of top-down correlation. *Technometrics, 29(3)*, 351-357.

Johnson, N. L., & Kotz, S. (1972). *Distributions in statistics: Continuous multivariate distributions.* New York: John Wiley.

Kibble, W. F. (1941). A two-variate gamma type distribution. *Sankhya, 5*, 137-150.

Konijn, H. S. (1956). On the power of certain tests for independence in bivariate populations. *The Annals of Mathematical Statistics, 27*, 300-323.

Mardia, K. V. (1962). Multivariate pareto distributions. *The Annals of Mathematical Statistics, 33*, 1008-1015.

Mardia, K. V. (1970). *Families of bivariate distributions*. Darien, Conn: Hafner.

Morgenstern, D. (1956). Einfache beispiele zweidimensionaler Verteilungen. *Mitteilungsblatt für Mathematische Statistik, 8*, 234-235.

Nagao, M., & Kadoya, M. (1971). Two-variate exponential distribution and its numerical table for engineering applications. *Bull. Disaster Prev. Res. Inst.,* Kyoto Univ., *20*, 183-215.

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association, 60*, 516-522.

Ruymgaart, F. H. (1974). Asymptotic normality of nonparametric tests for independence. *The Annals of Statistics, 2*, 892-910.

Ruymgaart, F. H., Shorack, G. R., & van Zwet, W. R. (1972). Asymptotic normality of nonparametric tests for independence. *The Annals of Mathematical Statistics, 43*, 1122-1135.

Savage, I. R. (1956). Contributions to the theory of rank order statistics - the two-sample case. *The Annals of Mathematical Statistics, 27*, 590-615.

Shiratata, S. (1974). Locally most powerful rank tests for independence. *Bulletin of Mathematical Statistics, 16*, 11-21.

Wei, L. J. (1981). Estimation of location difference for fragmentary samples. *Biometrika, 68(2)*, 471-476.