Some Results About the Power of a Rank Correlation Test

Alcuni risultati sulla potenza di un test di correlazione tra ranghi

Claudio Giovanni Borroni
Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali
e-mail: claudio.borroni@unimib.it

Keywords: rank correlation tests, asymptotic relative efficiency, FGM distribution

1. Overview and main results

Consider two sequences of ranks $R_{11}, R_{21}, \ldots, R_{n1}$ and $R_{12}, R_{22}, \ldots, R_{n2}$, obtained by sorting a sample of $n$ subjects according to two different criteria. Recently Borroni and Zenga (2007) proposed a new rank association index, which is essentially thought of as a measure of variability of the total ranks $T_i = R_{i1} + R_{i2}$ $(i = 1, \ldots, n)$. In its normalized form, ranging in the interval $[-1, 1]$, the new measure (which can be obviously considered as a competitor of known indexes as Spearman’s rho $\rho_n$, Kendall’s tau $\tau_n$ and Gini’s cograduation index $G_n$) is defined as

$$D_n = \frac{3}{n^2 - n} \sum_{i=1}^{n} \sum_{j=1}^{n} |R_{i1} + R_{i2} - R_{j1} - R_{j2}| .$$

In Borroni and Zenga (2007) the sample properties of $D_n$ are studied; more specifically, the exact expectation and variance of $D_n$ under the null hypothesis of independence are determined. Moreover, the asymptotic normality of $D_n$ is proved. A comparison of the performances of $D_n$ and its classical competitors in small samples is conducted, by Monte carlo simulations, in Borroni and Zenga (2007) and also in Borroni and Cazzaro (2006). As expected, it is not possible to locate the best test-statistic under every different simulated model. However, it can be conjectured that $D_n$ has better performances when it is used to test independence against discordance; moreover, $D_n$ performs very often better than $G_n$ but just occasionally better than $\rho_n$ and $\tau_n$, which are roughly equivalent even for small sample sizes. An important case to verify such conclusions for large sample sizes is the Fairlie-Gumbel-Morgenstern (FGM) model with uniform marginals and pdf $h_\theta(x, y) = 1 + \theta(2x - 1)(2y - 1)$ $(0 \leq x, y \leq 1)$ where the parameter $-1 \leq \theta \leq 1$ regulates association and vanishes under independence. Figure 1 reports the simulated power functions of $D_n$, $\rho_n$, $\tau_n$ and $G_n$ obtained by 50000 samples of size $n = 500$. These functions are plotted against $\theta$ in the range $[-0.1, 0.1]$ to better appreciate their differences. Similar conclusions can hence be drawn for large sample sizes: $D_n$ has a quite different performance with respect to its competitor, being definitely more powerful for discordance alternatives, while performing worse if concordance is considered; under these conditions, $D_n$ performs definitely better than $G_n$ but just slightly better than $\rho_n$ and $\tau_n$.

A good point should be to confirm such conclusions analytically. For instance an important result could be to compute the Pitman asymptotic relative efficiency (ARE) of $D_n$ with respect to its competitors. One has then to know the expression of the asymptotic expected value of each test-statistic under the alternative hypothesis and its asymptotic variance under the null hypothesis. Differently from $\rho_n$, $\tau_n$ and $G_n$, a closed form of the expected value of the test-statistic $D_n$ under a general alternative
hypothesis cannot be unfortunately derived. However the asymptotic equivalence of $D_n$ with a U-statistic proved in Borroni and Zenga (2007) can be exploited. More specifically, it has been proved that the difference $\sqrt{n}(D_n - U_n)$ converges in probability to zero, after defining $U_n = \left(\begin{smallmatrix} n \\ 3 \end{smallmatrix}\right)^{-1} \sum_{(n,3)} \Psi_{0}(X_{i_1}, Y_{i_1}; X_{i_2}, Y_{i_2}; X_{i_3}, Y_{i_3}) - 1$ (the sum $\sum_{(n,3)}$ being taken over the $\left(\begin{smallmatrix} n \\ 3 \end{smallmatrix}\right)$ subsets $1 \leq i_1 < i_2 < i_3 \leq n$ of $\{1, \ldots, n\}$), whose 3-degree symmetric kernel is $\Psi_{0}(X_{1}, Y_{1}; X_{2}, Y_{2}; X_{3}, Y_{3}) = \Psi(X_{1}, Y_{1}; X_{2}, Y_{2}; X_{3}, Y_{3}) + \Psi(X_{2}, Y_{2}; X_{1}, Y_{1}; X_{3}, Y_{3}) + \Psi(X_{3}, Y_{3}; X_{2}, Y_{2}; X_{1}, Y_{1})$, where $\Psi(X_{1}, Y_{1}; X_{2}, Y_{2}; X_{3}, Y_{3}) = [2 S(X_{2} + Y_{2} - X_{3} - Y_{3}) - 1] S(X_{2} - X_{1} + S(Y_{2} - Y_{1}) - S(X_{3} - X_{1}) - S(Y_{3} - Y_{1})] (S(a)$ being 1 if $a \geq 0$ and 0 elsewhere).

The expected value of the kernel $\Psi_{0}$ can then be computed by separating the single elements in its definition and by adding their expected values. When the joint density function has a simple expression, such expected values can be easily computed. For the FGM distribution, some tedious computations give the following expression of the asymptotic expected value of $D_n$: $\lim_{n \to \infty} E(D_n) = -17\theta^2/945 + 26\theta/105 + 2/5$.

For the FGM model, the following results can be also determined: $\lim_{n \to \infty} E(\rho_n) = \theta/3$, $\lim_{n \to \infty} E(G_n) = 4\theta/15$, $\text{Var}(\rho_n|\theta = 0) = 1/n + o(1/n)$, $\text{Var}(G_n|\theta = 0) = 2/3n + o(1/n)$, $\text{Var}(D_n|\theta = 0) = 1751/3150n + o(1/n)$. Hence

$$\text{ARE}(D_n, \rho_n) = \text{ARE}(D_n, \tau_n) = \left(\frac{26/105}{1/3}\right)^2 \frac{2}{1751/3150} = \frac{12168}{12257} \simeq 0.9927$$

$$\text{ARE}(D_n, G_n) = \left(\frac{26/105}{4/15}\right)^2 \frac{2/3}{1751/3150} = \frac{12675}{12257} \simeq 1.0341$$

As above conjectured, test based on $D_n$ is hence slightly less powerful than the ones based on $\rho_n$ and $\tau_n$, but its performance is definitely better than the test based on $G_n$.

Of course, it would be useful to compute the ARE of $D_n$ for other distributions. The above technique can be applied as long as the expected value of $U_n$ can be easily computed. When this task is not feasible, different kind of asymptotic approximations of $D_n$ could be tried. For instance, as $D_n$ is not a linear rank statistic, its projection in the class of such statistics could be used; this issue will be the object of a future research.

References
