Some Notes on the Cucconi Rank Test for Location-Scale Problems

Alcune Annotazioni sul Test di Cucconi per il Controllo di Ipotesi Congiunte su Parametro di Posizione e Scala

Marco Marozzi
Dipartimento di Economia e Statistica, Università della Calabria
Silvia Osmetti
Istituto di Statistica, Università Cattolica, Milano

Riassunto: Cucconi (1968) ha proposto un test basato sui concetti di rango e contro-rango per il controllo di ipotesi congiunte sui parametri di posizione e scala nel confrontare due popolazioni. Nonostante le sue interessanti caratteristiche, che includono un'evocativa interpretazione in chiave geometrica, il test è del tutto ignorato al di fuori della comunità scientifica italiana. La letteratura straniera, per affrontare problemi su posizione e scala congiuntamente, è solita indicare, tra i test sui ranghi, il test di Lepage (1971). Nel lavoro si mostra come le prestazioni del test di Cucconi siano del tutto paragonabili a quelle del test di Lepage, e che anzi in certe situazioni l'uso del test di Cucconi sia preferibile. Il mancato utilizzo di questo test non è quindi dovuto a ragioni metodologiche.

Keywords: Location-Scale Problems, Nonparametric Inferences, Rank Testing, Cucconi Rank Test, Lepage Rank Test.

1. Introduction

The general null hypothesis of interest in comparing two populations is that they have the same distribution H_0: $F_1(t) = F_2(t)\ \forall\ t \in]-\infty,+\infty[$, where F_1 and F_2 are their respective distribution functions. In the paper we consider the location-scale problem which corresponds to take

$$F_1(t) = G\left(\frac{t-\delta_1}{\sigma_1}\right) \text{ and } F_2(t) = G\left(\frac{t-\delta_2}{\sigma_2}\right),$$

where $G(.)$ is the distribution function for a continuous variable with location 0 and scale 1, δ_1 and δ_2 (σ_1 and σ_2) are the locations (scales) of population 1 and 2, respectively. Therefore, as alternative hypothesis we consider H_1: $\delta_1 \neq \delta_2 \cup \sigma_1 \neq \sigma_2$.

The problem is addressed within the nonparametric framework (Marozzi, 2002) and in particular via rank testing. The best known and most used rank test for location-scale problems is the Lepage (1971) test. We consider the Cucconi (1968) test as well, which is not known outside the Italian statistical school but performs like the Lepage test and it is better in certain situations.

2. The Cucconi and Lepage Rank Test for the Location-Scale Problem

Let $\underline{X}_1 = (X_{11},..., X_{1n_1})$ and $\underline{X}_2 = (X_{21},..., X_{2n_2})$ be independent random samples from populations 1 and 2. Lepage (1971) proposed a test statistic for the location-scale problem which is a combination of the Wilcoxon and Ansari-Bradley statistics. The Lepage statistic L is

$$L = \frac{(W - E(W))^2}{VAR(W)} + \frac{(A - E(A))^2}{VAR(A)},$$

where W is the Wilcoxon statistic, and A the Ansari-Bradley one. It is known that to compute W and A statistics the combined sample $Y = (\underline{X}_1, \underline{X}_2)$ elements should be ordered from least to greatest. Let W_{ji} denote the rank of X_{ji} in the combined sample, then $W = \sum_{i=1}^{n_2} W_{2i}$. To compute A statistic assign the score 1 to both the smallest and largest observations in the combined sample, the score 2 to the second smallest and second largest, and so on. Let A_{ji} denote the score of X_{ji} in the combined sample, then $A = \sum_{i=1}^{n_2} A_{2i}$. E(.) and VAR(.) denote the expected value and variance of W and A under H_0. The corresponding formulae are, for W $E(W) = \frac{n_2(n+1)}{2}$ and $VAR(W) = \frac{n_1 n_2 (n+1)}{12}$, where $n=n_1+n_2$. For A two cases should be distinguished, $E(A) = \frac{n_2(n+2)}{4}$ and $VAR(A) = \frac{n_1 n_2 (n+2)(n-2)}{48(n-1)}$ when n is even, $E(A) = \frac{n_2(n+1)^2}{4n}$ and $VAR(A) = \frac{n_1 n_2 (n+1)(3+n^2)}{48n^2}$ when n is odd. To test H_0 at the $0<\alpha<1$ level of significance reject H_0 if $L \geq l_\alpha$, where the constant l_α is chosen so that the type-one error rate is α. Tables for Lepage test can be found in Lepage (1973). Lepage (1971) showed that since W and A statistics are not correlated under H_0, then L statistic has a limiting chi-squared distribution with 2 degrees of freedom.

Cucconi (1968) proposes a quite different rank solution based on

$$C = \frac{U^2 + V^2 - 2\rho UV}{2(1-\rho^2)},$$

where $U = \frac{6\sum_{i=1}^{n_1} W_{1i}^2 - n_1(n+1)(2n+1)}{\sqrt{n_1 n_2 (n+1)(2n+1)(8n+11)/5}}$, $V = \frac{6\sum_{i=1}^{n_1} (n+1-W_{1i})^2 - n_1(n+1)(2n+1)}{\sqrt{n_1 n_2 (n+1)(2n+1)(8n+11)/5}}$ and $\rho = \frac{2(n^2-4)}{(2n+1)(8n+11)} - 1$. Note that U is based on the squares of the ranks W_{1i}, while V is based on the squares of the counter-ranks (n+1-W_{1i}) of the first sample. Let U' and V' be U and V computed on the second sample, then U'=-U and V'=-V and so it does not matter if one acts on the first or second sample to compute C. Cucconi (1968) proves that under H_0 $E(U)=E(V)=0$ and $VAR(U)=VAR(V)=1$. Of course U and V are (negative) dependent. More precisely, $\rho=CORR(U,V)=COVAR(U,V)$ and if $n>2$ then $-1<\rho<-7/8$. Under H_0, the asymptotic density function of (U,V) is a bivariate normal

$$f(u,v)=\frac{1}{2\pi\sqrt{1-\rho_0^2}}\exp\left(-\frac{u^2+v^2-2\rho_0 uv}{2\left(1-\rho_0^2\right)}\right),$$

where ρ_0=-7/8. When n_1 is not too different to n_2, the normal approximation is already good when $n_1,n_2>6$. H_0 should be rejected when $C>-\ln\alpha$. The Cucconi test is unbiased and consistent. Geometrically, the rejection region of the test is the set of points (u,v) outside the ellipse with equation $u^2+v^2-2\rho_0 uv=-2\left(1-\rho_0^2\right)\ln\alpha$. Note that when $\delta_1>\delta_2$ and $\sigma_1=\sigma_2$ $E(U)>0$ and $E(V)<0$; when $\delta_1=\delta_2$ and $\sigma_1<\sigma_2$ $E(U)<0$ and $E(V)<0$; when $\delta_1>\delta_2$ and $\sigma_1<\sigma_2$ $E(U)$ may be close to 0 but $E(V)<0$. Therefore, under H_0 (U,V) is centered on (0,0) while under H_1 is not.

3. A Comparative Study

A simulation study was performed to evaluate how the Cucconi rank test works and to compare its size and power with those of the Lepage test. We considered seven distributions: a standard normal distribution N(0,1); a double-exponential; a Student's t with 2 df; a standard Cauchy; a 10% outlier distribution obtained as a mixture of a N(0,1) with probability 0.9 and a N(1, 100) with probability 0.1; a 30% outlier distribution obtained as a mixture of a N(0,1) with probability 0.7 and a N(1,100) with probability 0.3; a bimodal obtained as a mixture of a N(-1.5,1) with probability 0.5 and a N(1.5,1) with probability 0.5. We considered (n_1,n_2)=(30,30). The null hypothesis tested was that the populations have the same distribution. The location problem has been considered with a positive value of the location shift δ_1-δ_2 that was specified so that the power of W test for each distribution was close to 50%. The scale problem has been considered with a value of the location shift σ_1/σ_2 greater than one that was specified so that the power of A test was close to 50%. The location-scale problem (with various values of δ_1-δ_2 and σ_1/σ_2) was considered under normal distribution in order to investigate the robustness of t and F parametric tests as well. The power was estimated at α=5% through 10,000 Monte Carlo simulations.

Table 1(a) shows that for the location problem L and C perform very similarly. It should be noted that the power loss with respect to W test ranges from 20% to 25%. Table 1(b) shows that for the scale problem C is better than L test under bimodal, normal, double-exponential, t_2 and 10% outlier distributions; whereas L is better than C under Cauchy and 30% outlier distributions. Table 2 shows that C performs better than L test for location-scale problem under normal distribution. Note that A test is not robust for the scale problem when δ_1-δ_2. Of course, it is not surprising that C and L tests work better than t, F, A and W for the location-scale problem. Moreover, C and L control the type-one error rate. It is important to note that C was always more conservative than L.

Table 1: *Relative efficiency with respect to (a) W test, location problem and (b) A test, scale problem.*

distribution	(*a*) test t	(*a*) test L	(*a*) test C	(*b*) test F	(*b*) test L	(*b*) test C
Normal	104.4	76.8	76.3	144.6	80.6	96.2
Double-exponential	-	80.7	79.8	-	80.8	91.5
t_2	-	81.0	81.0	-	80.9	84.8
Cauchy	-	84.4	83.5	-	80.3	76.7
10% outlier	-	78.5	79.0	-	81.3	84.2
30% outlier	-	83.7	83.7	-	80.4	74.9
Bimodal	-	77.8	75.7	-	79.9	99.1

Table 2: *Power estimates under normal distribution*: (*a*) *position problem when* $\sigma_1/\sigma_2\neq 1$, (*b*) *scale problem when* δ_1-$\delta_2\neq 0$.

(*a*)

σ_1/σ_2	test t	test W	test L	test C
1.5	66.8	63.9	71.6	74.9
2	71.6	68.9	91.9	95.1
3	76.9	72.9	99.7	99.9
5	79.6	74.9	100	100

(*b*)

δ_1-δ_2	test F	test A	test L	test C
0.25	73.0	47.4	50.2	56.7
0.5	72.2	40.8	75.6	79.5
0.75	72.9	32.1	94.3	95.3
1	73.4	21.9	99.5	99.6

4. Conclusion

The best known and most used rank test for the location-scale problem is the Lepage (1971) test. For the same problem, we considered the Cucconi (1968) test as well, which is not known outside the Italian statistical school. It has been shown that it performs like the Lepage test and sometimes it works better. Moreover, it is always more conservative. We suggest that the practitioner should take into account the Cucconi test as an alternative solution when addressing location-scale problems.

References

Cucconi O. (1968) Un Nuovo Test non Parametrico per il Confronto tra Due Gruppi Campionari, *Giornale degli Economisti*, XXVII, 225-248.

Lepage Y. (1971) A Combination of Wilcoxon's and Ansari-Bradley's Statistics, *Biometrika*, 58, 213-217.

Lepage Y. (1973) A Table for a Combined Wilcoxon Ansari-Bradley Statistic, *Biometrika*, 60, 113-116.

Marozzi M. (2002) Some Notes on Nonparametric Inferences and Permutation Tests, *Metron*, LX, 3-4, 141-153.