A non-parametric test on ranks in multiple series

Pietro Battiston

This version: July 20, 2015

Consider a situation in which data is available about K samples (possibly, but not necessarily, time series), each containing N observations. The hypothesis of a researcher is that in a given set T, which includes exactly one observation from each sample, there was a positive shift.

For an example, assume that the researcher was told that on April 1st of some year, the temperature around the world was significantly higher than in the other 29 days of the month, and she wants to understand whether this is a joke. She has access to daily temperature measurements from 10 different cities, but the temperatures of such cities are well known to be very etherogeneous (and uncorrelated - as unrealistic as this assumption may seem in the example proposed).

How can the researcher test such hypothesis without making any distributional assumption (i.e. normality of temperatures)? In this short note I will compare three possible methods:

1. a Mann-Whitney U test (Mann and Whitney, 1947) which compares the 10 values observed for April 1st against the $10 \times 29 = 290$ values observed on the other days,
2. a set of 10 different Mann-Whitney U tests (one for each city), the p-values resulting from which are then aggregated in a single p-value with the Fisher method for meta-analysis,
3. a new (to the best of my knowledge!) test which I present in the following section.

1 Description of the "ballrank" test

Consider city i, and sort the observations from the warmest day to the coldest one. Let r_i be the (0-based) rank of the observation for April 1st: it can take any value from 0 to 29, and under the null hypothesis that the temperature on April 1st is not exceptionally high, any of these values is equally probable. Hence, the sum across cities of the ranks,

$$R = \sum_{i=1}^{K} r_i,$$

is, still under the null hypothesis, distributed as the sum of K discrete uniform variables from 0 to $N-1$. What does this distribution look like?

As already observed, any value from 0 to $N-1$ is equally probable for each r_i, and hence each of the N^K combinations in $\{0,\ldots,N-1\}^K$ is also equally probable. How many combinations result in a given value of R? The answer is the same as to the question of how many ways there are to distribute R balls in K bins each with place for maximum $N-1$ balls, which is well known to be

$$\sum_{i=0}^{K}(-1)^i\binom{K}{i}\binom{R+k-1-iN}{K-1},$$

hence giving a pdf for a given R of

$$p(R,K,N)=\frac{\sum_{i=0}^{K}(-1)^i\binom{K}{i}\binom{R+k-1-iN}{K-1}}{N^K}.$$

By comparing the observed R with the (CDF of the) distribution just described, a p-value can be easily computed. Although the procedure is relatively computationally cheap, for large values of the parameters it can also simply be approximated with a normal distribution: more precisely, since R is the sum of K discrete uniform distributions, each with mean

$$\mu=\frac{N-1}{2}$$

and variance

$$\sigma^2=((N)^2-1)/12,$$

the Central Limit Theorem guarantees that

$$\frac{R-K\mu}{\sqrt{\sigma^2\cdot K}}\overset{K\to\infty}{\to}\mathcal{N}(0,1).$$

2 Comparison of the tests

Intuitively, the main conceptual difference between the pooled Mann-Whitney U test (method 1) on one side and the two other methods on the other side is that the former sacrifices some information (the composition of the samples - e.g. the fact that the observations come from different cities, with different climates) in order to obtain a single large sample on which to run the test (the combinatorial nature of the test implying that its power tends to increase with the sample size). So the relative power of such method compared to the others two should decrease the larger the differences between cities.

A more subtle difference between methods 2 and 3 is related to the fact that the p-value associated to any given statistics S, calculated against any given distribution $\mathcal{D}$, is obtained as the probability of a result being, "just by chance", *at least as extreme* as S. When the domain of $\mathcal{D}$ is composed by an infinite or very large set of points (and $\mathcal{D}$ is not concentrated in few of them), the fact that the inequality is weak is close to irrelevant; but if instead such domain contains few points, the probability of a result being *at least as extreme* as S is significantly larger than the probability of a result being *strictly more extreme* than S. Now, according to method 2 a Mann-Whitney test is first ran on each subsample. When, as in this case, one of the two sets is composed by only one element, the Mann-Whitney test is equivalent to dividing the rank of such element by the total number of observations. If such total number is small, the fact that the inequality is weak is relevant, and this implies a loss of power of the test. On the other hand, if the ranks are first summed and *then* their sum is compared to its theoretical distribution under the null hypothesis, the fact that the inequality is weak becomes much less relevant, because each point in the domain of such theoretical distribution has a low probability mass. Notice that while there are alternatives to the Fisher test for aggregating p-values from independent tests of a same hypothesis (e.g. the z-transform method - also known as "Stouffer's approach" - or the Edgington method), and they may perform better or worse depending on the kind of data analyzed, the choice of which to use is irrelevant for the problem just described.

2.1 Simulations

The three methods were compared using a Monte Carlo approach. Samples were created according to the following model:

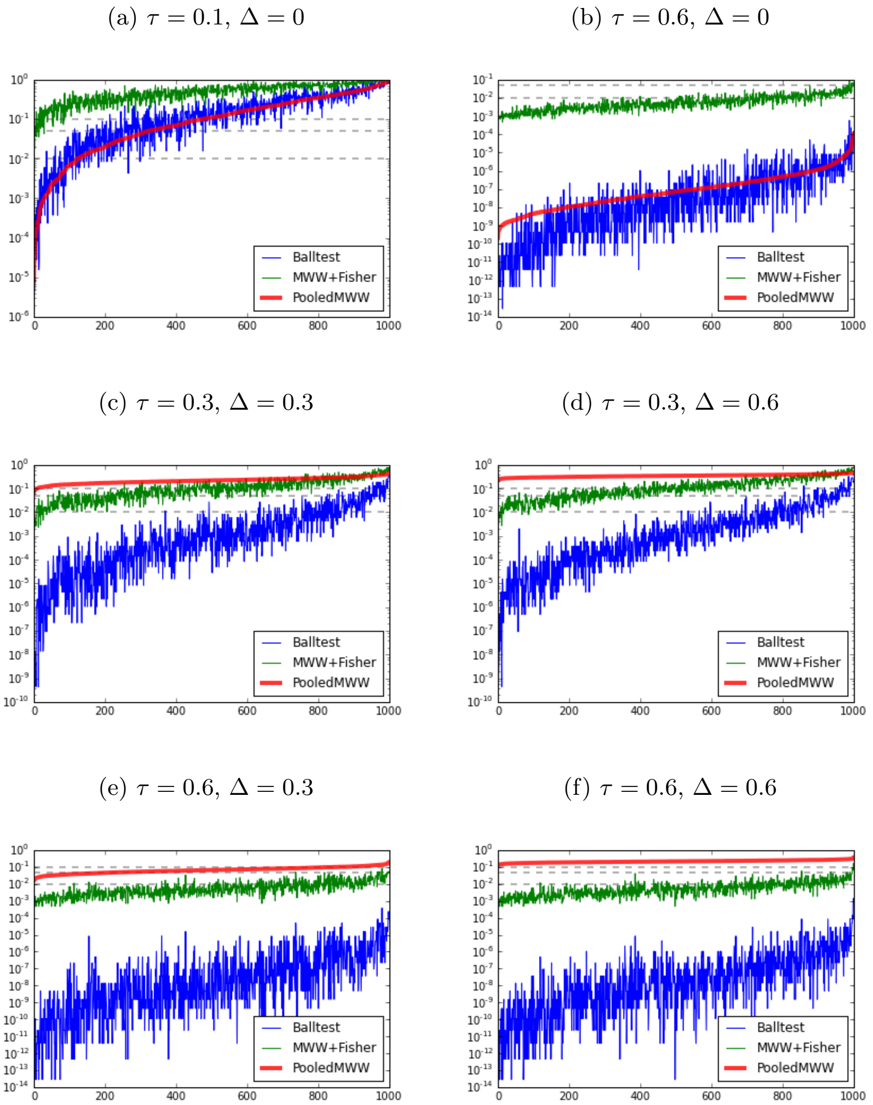
$$x_{i,j} = \tau T_i + \Delta j + \epsilon_{i,j}$$

where the parameters τ and Δ denote respectively the effect that the researcher wants to identify and a group-specific fixed effect, and $\epsilon_{i,j}$ is a random component extracted from a uniform distribution over the interval $[0, 1]$. Notice that the model is set up in such a way that for large values of Δ (i.e. $\Delta > 1+\tau$), any observation from group $j + 1$ is larger than any observation from group j. On the other hand, if $\Delta = 0$, then all groups are identically distributed. The boolean variable T_i, which determines the sample on which the effect is present, is defined as true only for a given value of i, and false otherwise: without loss of generality, let us set $T_i = 1 \iff i = 1$.

Several combinations of the parameters τ and Δ were simluated: for each, 1000 different samples were created, each with i ranging from 1 to $N = 8$, and j ranging from 1 to $K = 15$, and on each sample the three methods were ran in order to compare the distributions of the resulting p-values.

The distinction drawn at the beginning of the present section between the pooled MWW test and the other two methods can be reformulated by stating that such method is relatively advantaged for a value of Δ close to 0, i.e. when the information about the different groups is of negligible importance; the other two methods are expected to perform relatively better for values of Δ which are larger (relatively to the value of τ).

This is indeed what emerges from Figure 1. For larger values of Δ, the new test is much more powerful than the pooled MWW test: when $\Delta = 0.6$, the



(a) $\tau = 0.1$, $\Delta = 0$

(b) $\tau = 0.6$, $\Delta = 0$

(c) $\tau = 0.3$, $\Delta = 0.3$

(d) $\tau = 0.3$, $\Delta = 0.6$

(e) $\tau = 0.6$, $\Delta = 0.3$

(f) $\tau = 0.6$, $\Delta = 0.6$

Figure 1: (Logs of) p-values obtained for different values of the parameters τ and Δ. The results are sorted according to the reference p-values obtained with method 1 (pooled MWW). The dashed horizontal lines signal the 1%, 5% and 10% confidence intervals.

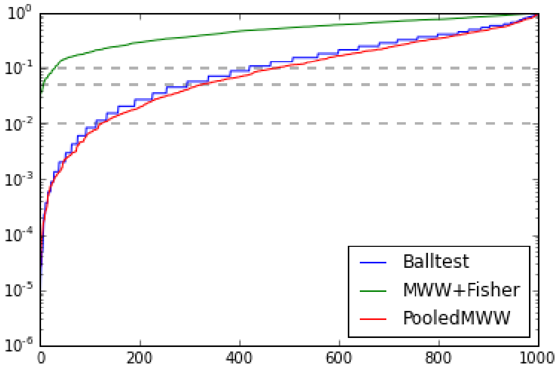


Figure 2: Same p-values as in Figure 1 (a), but with the results for each method sorted independently from the others (in order to reconstruct three cumulated distribution functions).

latter is never significant, even at the 10% level, while the former usually is (and is *always* significant at the 1% level for $\tau = 0.6$). However it is interesting to notice that in the case in which the informational advantage is the lowest ($\Delta = 0$), the pooled MWW test is more powerful than the new test by a very small margin (which can be seen more clearly in Figure 2).

Even more intriguingly, the difference blurs with increasing values of τ, even keeping $\Delta = 0$, as can be seen in Figure 1 (b): the new method performs consistently better than the pooled MWW for small p-values. This may be of limited interest, however, since both methods reach significant levels in all cases.

Summing up, while the pooled MWW emerges as more powerful than the new test only under specific assumptions, and by a slight margin, the new test achieves significance in many cases in which the pooled MWW does not.

Concerning method 2, when comparing with the pooled MWW, it is again advantaged whenever the difference between groups is relevant information (when $\Delta > 0$), but it is dominated by the new test in virtually all cases. As already mentioned, this could reflect two different phenomena: the specific aggregation method used (the Fisher method), and the fact that information is "wasted" when applying the weak inequality in each group. While disaggregating the two effects is out of the scope of the present note, Figure 3 provides clean evidence that method 2 is inefficient: the expected distribution of p-values in the case in which the null hypothesis holds ($\Delta = 0$) should be uniform between 0 and 1, and this is what happens, at least approximately, when using the two "direct" methods (1 and 3). When aggregating the results of the group-level Mann-Whitney tests, instead, the distribution is clearly biased towards higher values.

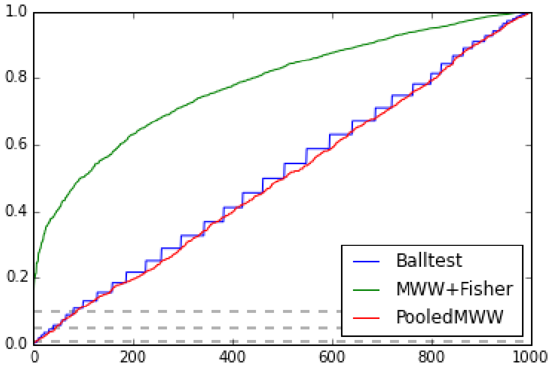


Figure 3: P-values in the case $\Delta = \tau = 0$: results for each method sorted independently.

3 Conclusions

A test is presented which allows to compare one value against some others across a given number of subsamples of equal size. If the subsamples are not identically distributed, the new sample is able to exploit this information in order to achieve higher power than a Mann-Whitney U test ran on all the values pooled together: on the other hand, if the subsamples are identically distributed, the Mann-Whitney U is more powerful only in some cases, and by a slight margin.

The new test could in principle be generalized to the case in which the groups have different sizes. One simple way to do so is to make all groups of the same size by adding observations in the smaller ones: in order to obtain conservative results, however, such artificial observations should be larger than any other items in the groups they are added to, and this would decrease the power of the test (although to a small extent if the number of artificial observations needed is small - i.e. if groups are *almost* the same size). In principle, the sum of ranks from groups of any size could be also compared to its theoretical distribution, but such theoretical distribution is less obvious to compute than the one described in Section 1. Whatever approach is adopted in such extension, it should also be noted that so far each observation (or equivalently, each rank position) was assumed to be equally informative: if subsamples have different sizes, this may not necessarily be true.

Finally, the test could also be extended to the case in which the intersection of the set T with each subsample does not necessarily contain exactly one element. Again, it is in principle straightforward to calculate the theoretical distrubution of the resulting sum of ranks, but it is computationally far from trivial. Since the distribution of the U statistics is asymptotically normal, the same can be said for the sum of several U statistics, but this observation is of limited utility if the tests analyzed in the present note are to be used, as in the example provided, for studying small samples.

References

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.