Notes on Wilcoxon

Paolo Toccaceli

February 27, 2017

1 Introduction

The comparison between the accuracy of different algorithms is a type of problem that is extremely common. One could expect that statisticians had solved this problem satisfactorily a long time ago. The more I studied the problem, the more I realized it is not exactly true.

2 Wilcoxon signed-rank test

As discussed in a previous report, the Wilcoxon signed-rank test is a nonparametric (distribution-free) test that applies to paired observations. The null hypothesis is that the differences in each pair form a symmetric distribution around 0.

1. Compute the differences between pairs
2. Rank the differences in terms of the absolute value. Treat ties by averaging their ranks.
3. Sum together the ranks corresponding to the positive differences and call the result R_+.
4. Sum together the ranks corresponding to the negative differences and call the result R_-.
5. The test statistic is:
$$\min(R_+, R_-)$$

The problem with the Wilcoxon test is that it relies on two circumstances:

1. There are no cases in which the values in a pair have the same value.
2. There are no cases in which pairs have the same difference (in absolute value)

These assumptions are automatically verified if the variates come from a continuous distribution (with a PDF without any Dirac deltas).

But in practice this is never the case, strictly speaking.

There seems to be agreement that ties should be dealt by computing a common rank equal to the average of the ranks.

Wilcoxon and subsequently Pratt discussed how to modify the procedure to make it possible to apply it to the case in which there are pairs with no difference. Wilcoxon suggested to remove the pairs with the same value and then compute the ranking and the sums. Pratt showed how this could lead to undesirable behaviour in some cases and proposed instead that the ranking included the no difference cases, but also that the no differences cases would not be included in R_+ or R_-. Note that in either case the mapping of the statistic into a p-value would then disregard the no-difference cases (e.g. if we have 20 pairs and 5 have no difference, I would use the p-values for the N=20-5=15).

Both author did not offer much in the way of empirical studies or theoretical justification. I did not find references to papers that showed whether these adjustments really work, although textbooks do tend to say that Pratt's method is conservative.

However, these changes are based on intuition and heuristics. They were developed in the 50s, when computing facilities were scarce and of limited power. Surely, today we can use our computers to provide more guidance.

2.1 Recomputation of Wilcoxon statistic

Last week Prof. Vovk helpfully pointed out that one should not just look at a p-value reported by a library but should look at the adjacent values. Unfortunately, the libraries I use for Wilcoxon (in python as well as in R) do not let the user enter the statistic, but they start from the actual variates.

Also, I realized that the python Wilcoxon library computes the p-value using a Gaussian approximation that is unsatisfactory for fewer than 20 pairs.

So I set out to recompute the Wilcoxon statistic.

In Figure 1 you can see the graphs for the one-sided Wilcoxon statistic (i.e. I considered only R_-)

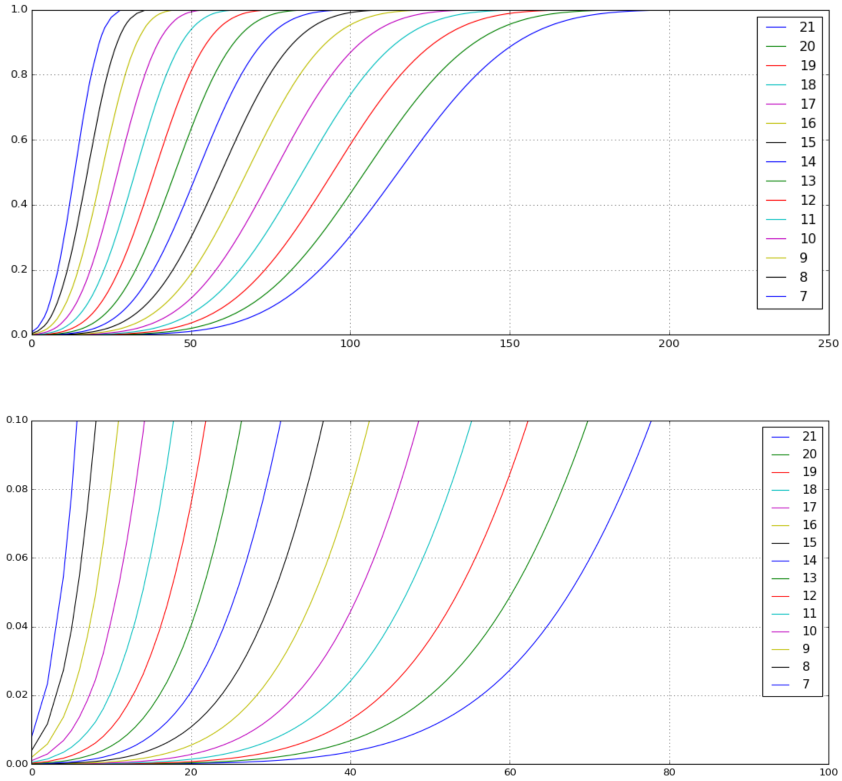

Figure 1: Wilcoxon statistics to p-value. The x axis has the value of the statistic and the y axis the probability to see a value less than or equal to the statistic if the null hypothesis is true. Each line correspond to a different number of paired observations. The bottom plot is for the interesting range of p-values.

The values were obtained by enumerating all the possible assignments of signs to ranks (all assumed to occur with the same probability). The computation took about 8s overall.

I compared the values to those reported for some critical values on statistical tables and they are in agreement.

2.2 Study of the effect of ties and zero differences

The p-values computed above apply if the distribution is continuous. What happens when there are ties and zeros?

Pratt does say that the distribution of the Wilcoxon statistic must change. But how?

I decided that this question was worth some investigation. I set up a simulation in which I generated a large number of sets of N pairs according to some distribution and I count of occurrences of the values of the statistic based on those pairs.

The pairs were generated by drawing from a normal distribution and then quantizing the values into a number of bins of the same width, between -10.5 and +10.5. To study the effect of the discretization, I repeated the simulation with different numbers of bins.

Also, I computed what the p-value would be if one used the Wilcoxon and Pratt methods of dealing with null differences and I counted those below 0.01 and below 0.05. The results are in Table 2.2. Ideally, one would want to find again 0.01 and 0.05 as relative frequencies. In fact, fewer were found, which means than both Wilcoxon and Pratt p-values are too conservative. What they suggest happens 1% of the time, actually occurs less often (so should be considered stronger evidence against the null hypothesis than the p-value would make you think). Surprisingly for me, Wilcoxon method appears to be close to the actual distribution, whereas Pratt is very conservative.

Figure 3 shows the resulting statistic distributions for the case of 23 bins.

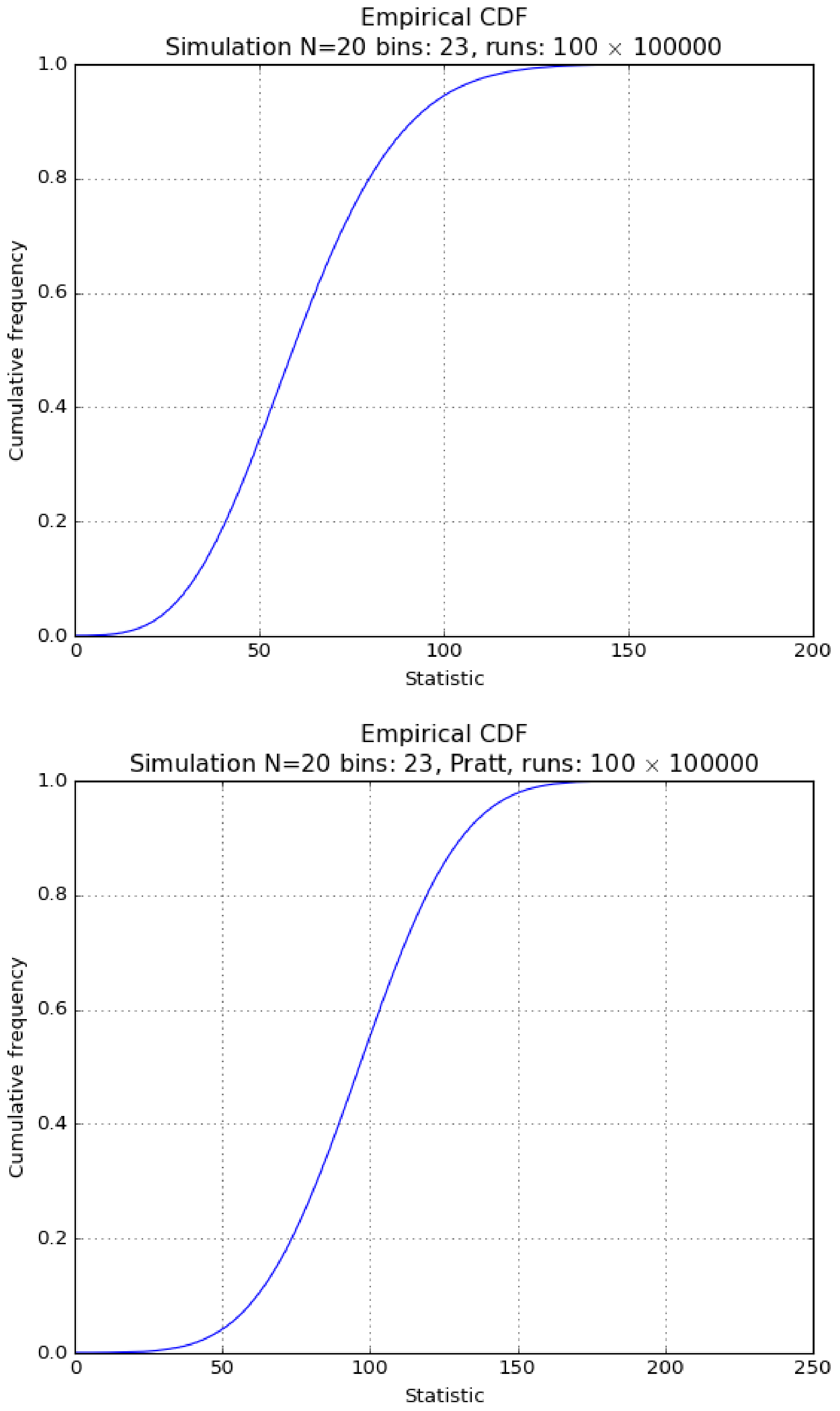

Figure 2: Wilcoxon statistics to p-value for N=20 from simulation that includes ties and zero diffs. The top plot uses Wilcoxon's method, the bottom plot uses Pratt's method.

Figure **??** shows the resulting statistic distributions for the case of 45 bins.

I also tried out an asymmetric distribution for the variates, the gamma function shown in Figure 4. I also

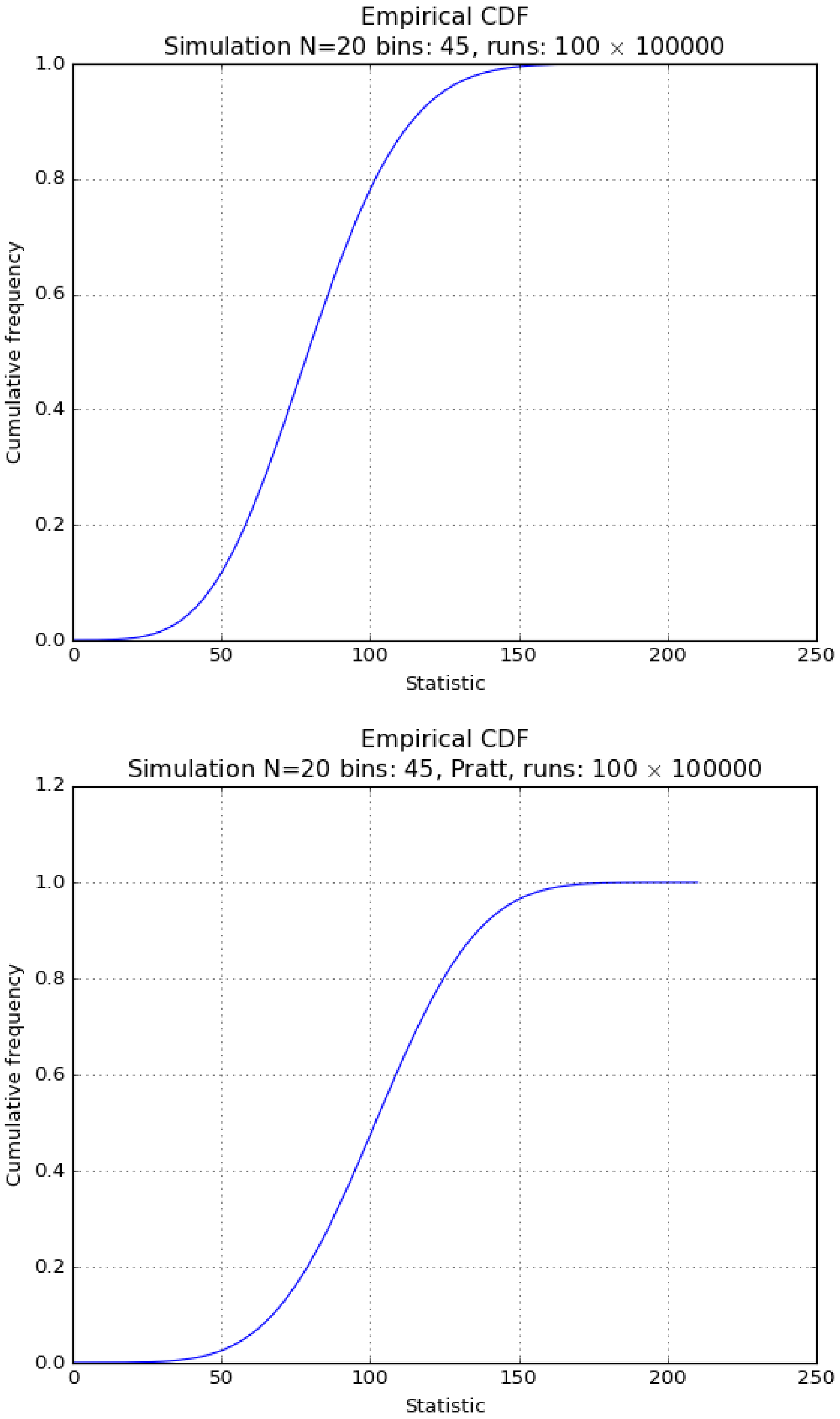

Figure 3: Wilcoxon statistics to p-value for N=20 from simulation that includes ties and zero diffs. The top plot uses Wilcoxon's method, the bottom plot uses Pratt's method.

Bins	Actual	Actual
	rel freq	rel freq
	Wilcoxon	Pratt
23	0.0074016	0.0011322
45	0.0087326	0.0028679
111	0.0093207	0.005883

Bins	Actual	Actual
	rel freq	rel freq
	Wilcoxon	Pratt
23	0.0434252	0.0060545
45	0.0479142	0.0161911
111	0.0490891	0.0313897

Table 1: Observed relative frequencies of variates for which computed p value would be ≤ 0.01 (left table) and ≤ 0.05 (right table)

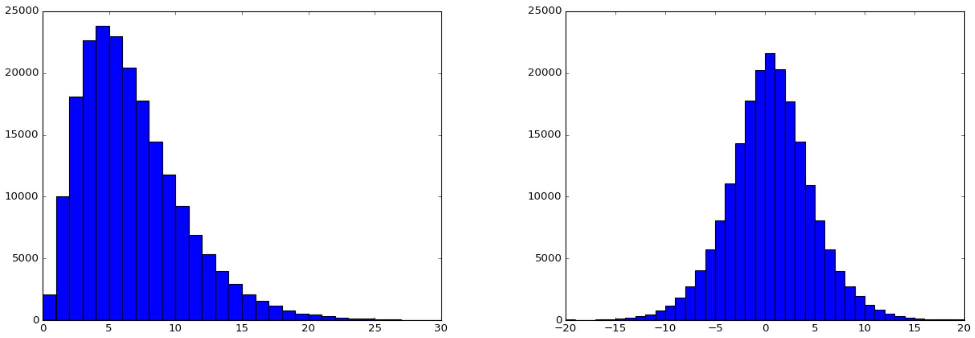

Figure 4: Gamma, shape 3, scale 2.25 (mode at 4.5). The right pane shows the distribution of the differences.

Bins	Actual	Actual
	rel freq	rel freq
	Wilcoxon	Pratt
10	0.0077425	0.0012352
20	0.0088929	0.0030831
30	0.0091832	0.0044841

Bins	Actual	Actual
	rel freq	rel freq
	Wilcoxon	Pratt
10	0.0444789	0.0067979
20	0.0482938	0.017308
30	0.0489291	0.0244664

Table 2: Observed relative frequencies of variates for which computed p value would be ≤ 0.01 (left table) and ≤ 0.05 (right table). In this case, the variates were generated with a gamma distribution, then they were quantized and their difference was taken.

3 Bibliography

- Wilcoxon, F, *Individual comparisons by ranking methods*, Biometrics Bulletin, Vol. 1, No. 6. (Dec., 1945), pp. 80-83

- Pratt, J.W., *Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures* - Journal of the American Statistical Association, Vol. 54, No. 287 (Sep., 1959), pp. 655-667