Publication III

Vesa P. Vaskelainen and Vesa Riihimäki. 2009. Bounds for the sample size in performance testing of combinatorial algorithms. InterStat, July 2009, paper no. 6, 14 pages.

Bounds for the sample size in performance testing of combinatorial algorithms

Vesa P. Vaskelainen* and Vesa Riihimäki†

Department of Communications and Networking
Helsinki University of Technology TKK
P.O. Box 3000, 02015 TKK, Finland

June 7, 2009

Abstract

The performance of combinatorial algorithms is often evaluated by using the computational times of a certain number of inputs. The run times of algorithms with certain type of data appear in computational results as the mean of a small sample. Still the choice of sample size is rarely based on the distribution of run times. In this work we use statistical tests to compare the performance of combinatorial algorithms. Furthermore, we determine confidence intervals for the run times of combinatorial algorithms. As a consequence, we get easy-to-use bounds of sample size for justifying the accuracy of computational results.

Keywords: Confidence interval, combinatorial algorithms, CPU performance, normal approximation, normality, statistical tests, skewness.

*Supported by the Academy of Finland under Grant No. 127493 and by Walter Ahlström Foundation (Walter Ahlströmin säätiö).

†Supported by the Academy of Finland under Grant No. 120196.

1 Introduction

In the past, statistical techniques have not been exploited widely in performance testing of combinatorial algorithms. This work attempts to show the benefit which could be obtained by using them, and in the broader sense the paper deals with the problem of the choice of sample size. We examine the use of statistical tests for comparing the performance of algorithms for the maximum transitive subtournament problem [9] and the minimum spanning tree problem [8]. In addition, the statistical theory of error in normal approximation [4] is applied to justify the required sample size for computational tests in order that the confidence intervals for means can be stated.

The paper is organized as follows. Section 2 deals with the statistical tests that are applied in Section 4. Section 2.1 surveys the relevant parts of the theory of error in normal approximation to this context. In Section 3 we discuss the combinatorial algorithms from which the experimental data of run times is obtained for Section 4.

2 Preliminaries

Statistical tests enable us to determine the statistical significance of an observation. For a general introduction to statistical tests see for example [14, 15]. In our case, when the inputs for different algorithms are the same and thus the samples are dependent, some possible tests for the equality of run time means are the *Student's paired t-test*, the *Wilcoxon signed rank test* and the *Sign test*. The t-test assumes the sample to be from normal distribution and the Wilcoxon signed rank test assumes symmetric distributions but not normality. The Sign test is even more tolerant by accepting nonsymmetric distributions. These ground assumptions should be taken into account when applying these tests. It can be said that the stricter assumptions are, the more efficient test is when the assumptions are fulfilled. An analytical comparison of the Wilcoxon and Sign tests to t-test is presented in [5]. A similar comparison based on examples is done in [10].

For the paired differences $d_i = x_i - y_i$, $1 \leq i \leq N$, only the ratio of the sample mean $\bar{d}$ and the sample standard deviation s_d affect the test value of the Student's paired t-test if we have certain sample size N. Using this, we can determine the minimum sample sizes for verifying the statistically significant difference of means between two normal distributions with signif-

icance level α, see Table 1. Note that these values are indicative also for the Wilcoxon signed rank test and the Sign test.

Table 1: Minimum sample sizes for determining the difference of the mean of two normal distributions with Student's paired t-test.

$\lvert\bar{d}/s_d\rvert$	$\alpha = \mathbf{0.1}$	$\alpha = \mathbf{0.05}$	$\alpha = \mathbf{0.02}$	$\alpha = \mathbf{0.01}$
0.1	273	387	545	669
0.3	32	46	64	78
0.6	10	14	19	23
1.0	5	6	9	11
1.5	4	5	6	7

Many theoretical models use normal distribution as a basis for calculations. To help decide whether to use normal models or not, different tests for normality are used. The most intuitive normality tests are graphical comparisons [13]. See more information for numerical/analytical tests in [11] for the *Kolmogorov–Smirnov test*, [12] for the *Lilliefors test*, or [2, 6] for the *Jarque–Bera normality test*.

2.1 Estimating Confidence Intervals for Skewed Distributions

The *skewness* of a distribution X is [3]

$$\gamma_1 = \frac{\mu_3}{\sigma^3}\,, \tag{1}$$

where μ_3 is the third moment about the mean and σ is the standard deviation. The *skewness of the sample* can be determined as follows [7]

$$g_1 \;=\; \frac{m_3}{m_2^{3/2}}\,, \tag{2}$$

where m_2 and m_3 are the second and the third central moment of the sample.

Let us consider a random sample x_i from a distribution of run times X. We are interested in the confidence interval of the mean $\bar{x}$. The approximation of the distribution of the mean can be done with normal distribution, and error in the approximation is controlled by sample size. That is, the normalized sample mean

$$\frac{\bar{x}-\mu}{s}\sqrt{N} \tag{3}$$

is asymptotically normally distributed [14]. Determining the confidence interval for the mean of normal distribution is explained in detail in many textbooks and it is [14]

$$\bar{x} \pm \frac{s}{\sqrt{N}} z_\alpha \,, \tag{4}$$

where $\bar{x}$ and s are sample mean and estimate for standard deviation, respectively. The fractile z_α is calculated from normal distribution.

Since (3) is only approximately and asymptotically normally distributed, the error in normal approximation must be considered. The suitable research for practical purposes is the work by Höglund [4]. Höglund approximates the distribution of the normalized sample mean (3) by Edgeworth expansion [4]

$$F_n(t) = \Phi(t) + \frac{P(t)\phi(t)}{\sqrt{N}} + \frac{Q(t)\phi(t)}{N} + o(\frac{1}{N}) \,, \tag{5}$$

where Φ and ϕ are cumulative and density functions of the standard normal distribution, respectively. P and Q are polynomials depending on the population distribution and sampling procedure. Using this he estimates the error ϵ in α when determining two-tailed confidence interval. To be exact, Höglund proves that

$$|\alpha^+_{N^+(\alpha,\epsilon)} - \frac{\alpha}{2}| = \epsilon + O(\epsilon^2), \tag{6}$$

where

$$\alpha^+_{N^+(\alpha,\epsilon)} = 1 - F_{N^+(\alpha,\epsilon)}(z_{\alpha/2}) \tag{7}$$

is the probability of the sample mean to be in the upper $\frac{\alpha}{2}$-fractile while the sample size is $N^+(\alpha,\epsilon) = \frac{\beta^2_\alpha}{\epsilon^2}$ and $\beta_\alpha = -P(z_{\alpha/2})\phi(z_{\alpha/2})$. Practically, this gives us a lower bound

$$N > \frac{P^2(z_{\alpha/2})\phi^2(z_{\alpha/2})}{\epsilon^2} \tag{8}$$

for the sample sizes when determining the confidence intervals for skewed distributions. Höglund proves that if the identically and independently distributed events in a sample are from the continuous distribution of the random variable X with unknown variance and $E[X^8] < \infty$, and the error in

both upper and lower bound of the confidence intervals does not exceed $\epsilon/2$, the sample size must fulfill [4]

$$P(t) = \gamma_1 \frac{2t^2 + 1}{6}\,, \tag{9}$$

$$N > \frac{(2z^2_{\alpha/2} + 1)^2 \phi^2(z_{\alpha/2})}{9} \frac{\gamma_1^2}{\epsilon^2}\,. \tag{10}$$

where γ_1 is the skewness (1) of the distribution and ϕ is density function of the standard normal distribution. The minimum sample sizes for determining confidence interval with skewness $\gamma_1 = 1$ are shown in Table 2. Reasonable range of use in (10) is $\alpha \geq \epsilon$ since the error should not exceed the magnitude of the nominal value. Note that the sample size increases relatively to γ_1^2. In addition to making the interval more credible, the increase in sample size narrows the confidence interval, i.e. makes it more accurate.

Table 2: The minimum sample sizes with skewness $\gamma_1 = 1$. The bounds for other values of γ_1 can be calculated by multiplying the numbers by γ_1^2.

ϵ	$\alpha =$ **0.1**	$\alpha =$ **0.05**	$\alpha =$ **0.02**	$\alpha =$ **0.01**
0.005	1944	1145	442	190
0.01	486	287	111	48
0.02	122	72	28	*(12)*
0.03	54	32	*(13)*	*(6)*
0.04	31	18	*(7)*	*(3)*

Example 1 Lognormal distributions with 0 mean and variance 0.314, 0.5 and 1 have skewnesses 1.00, 1.75 and 6.18. Table 2 gives bounds for the first case directly. For the other two that are clearly skewed, Table 2 gives required sample sizes for $\epsilon = 0.02$ and $\alpha = 0.1$ to be $N = 374$ and $N = 4660$.

In Example 1, the skewnesses were exact but when we apply Höglund's theorem for real experimental data skewness is always estimated from a sample. Höglund [4] advices to proceed with the following iterative method. Start with reasonably large n_0 and decide α and ϵ. After this evaluate an estimate for γ_1 by using g_1 in (2) with sample of size n_0. Then calculate $\hat{n}_0(\alpha, \epsilon)$ using (10) and check if $n_0 \geq \hat{n}_0(\alpha, \epsilon)$. If not, continue by evaluating a more accurate estimate for γ_1 based on the sample of size $n_1 \geq \hat{n}_0(\alpha, \epsilon)$. Then calculate $\hat{n}_1(\alpha, \epsilon)$ and check if $n_1 \geq \hat{n}_1(\alpha, \epsilon)$. Iterate this until $n_k \geq \hat{n}_k(\alpha, \epsilon)$ is fulfilled. Then the sample is sufficiently large.

3 Combinatorial Algorithms

The idea for this work arose during the computational experiments of algorithms for the *maximum transitive subtournament problem* [9]. Run times with the similar random input deviated so much that the commonly used average of ten run times did not make computational results reproducible. When examining the distributions of means it turned out that the distributions were strongly skewed. Two basic backtrack algorithms (A1 and A2) and the Russian doll search algorithm (A3) for finding the maximum transitive subtournaments are experimentally evaluated in [9]. Run time data from that work is shown in Table 3.

In addition, to illustrate the use of statistical pair tests we chose to use Prim's and Kruskal's algorithm for the *minimum spanning tree problem* [8]. Prim's algorithm slows down when the number of vertices is increased while Kruskal's algorithm is nearly invariant for this change if the number of edges is kept constant. An increase in the number of edges has a slowing effect on both. By trial one can quickly find the number of edges and vertices so that run times are nearly equal. Then randomly generated instances with these parameters produce an experimental data that is used in section 4.2.

These algorithms are deterministic i.e. in principle, in the same computational environment they always go through the same computational steps and use the same run time with the same input. Even so, we cannot generally say for a random input what the accurate run time will be without testing it experimentally. The main reason for this is that the properties of the input that influence the run time are difficult to know in full. The next section applies statistical methods to experimental data. It appears that inferences about the performance differences of combinatorial algorithms can be quite varying if based only on sample means.

4 Results

Experimental data is now analysed with statistical tests which were referred to in Section 2 and with the technique described in Section 2.1. The outcomes are from the inputs for which no isomorph rejection was carried out and the inputs are identical for the compared algorithms. The algorithms were implemented in C++ and runs made in a 2.0-GHz PC with Linux operating system. Directed random graphs were generated with standard C function

rand and undirected random graphs with *minstd_rand* in Boost library [1]. The population is the set of run times for the inputs with the same number of vertices and edges. The significance level $\alpha = 0.05$ is used unless noted otherwise. For a detailed description of the statistical tests used see e.g. [14, 15].

In Figure 1 the graphical comparison is done for the samples of run times in Table 4 on page 10. The unit of the X-axis is second and the width of bars (in Figures 1(a) and 1(b)) is chosen to be one fifth of the estimator of variance of the population, $s/5$. To avoid the choosing of width, consider a *sample distribution* for sample $\{x_1, x_2, ...x_N\}$ from population X

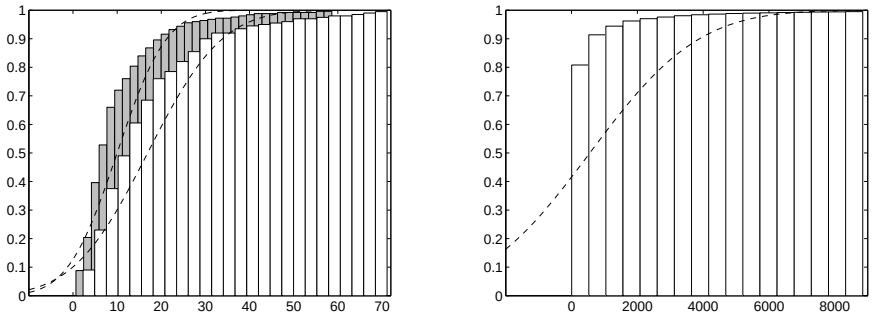
$$X_N(x) = \begin{cases} 0, & x < x_1^* \\ k/N, & x_k^* \leq x < x_{k+1}^* \\ 1, & x \geq x_N^* \end{cases} \tag{11}$$

where x_1^* is the smallest event in sample and x_2^* the second smallest and so on. By plotting $X_N(x)$ one can obtain a more accurate representation of the cumulative distribution function of the sample. This is demonstrated in Figures 1(c) and 1(d).

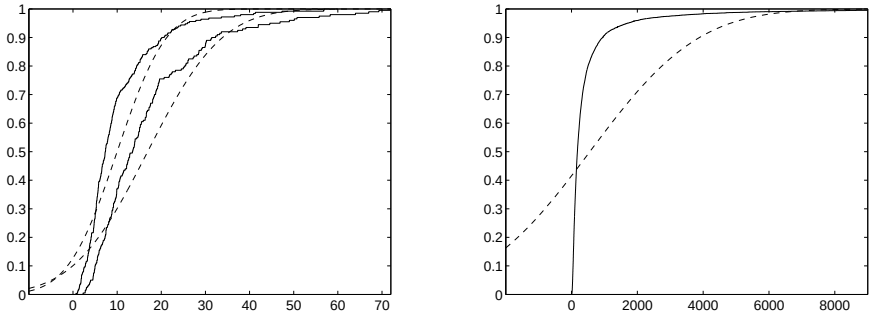
All samples deviate notably from the corresponding normal distribution and with Algorithm 1 the most. In the following the significance of the deviation is estimated with Lilliefors or Jarque–Bera tests to check if t-test can be used.

4.1 Algorithms for finding the maximum transitive subtournaments

To describe the insufficiency of ten runs, let us have 10 consecutive runs sampled by using algorithms for the maximum transitive subtournament problem. The run times are as stated in Table 3. A1 is one order of magnitude slower on average and is the quickest only in one case. The shape of the distribution of its run time seems to be highly skewed. A3 is 4.1 s quicker per run (or 32%) than A2 on average in these 10 runs. However, s_{A2} and s_{A3} are larger than $\bar{x}_{\mathrm{A2\text{-}A3}}$, and thus the significance of the observed difference should be determined. Lilliefors or Jarque–Bera tests cannot reject the normality of either of A2 or A3. The P-values are approximately 0.15 or more. Now, the null hypothesis H_0 is that the population mean of run times of A2 and A3



(a) A3 and A2 (b) A1



(c) A3 and A2 (d) A1

Figure 1: Graphical comparisons of cumulative distribution functions of run times of Algorithms 3, 2, and 1 (see section 4.1).

are equal, and we calculate the test value for the Student's paired t-test:

$$t = \frac{\bar{d}}{s_d}\sqrt{N} \approx \frac{4.1}{6.3}\sqrt{10} \approx 2.06\,. \tag{12}$$

That is lower than the critical test value $t_{0.05,10} = 2.26$ for significance level $\alpha = 0.05$ with sample size $N = 10$.

Thus the observed difference is not significant and we do not reject H_0. However, the test value is higher than $t_{0.1,10} = 1.83$ which means that the P-value for our case is between 0.05 and 0.1. With $\alpha = 0.1$, H_0 would have been rejected.

The ranks of the differences in run times for ten runs are combined

Table 3: Ten consecutive sample runs for A1, A2 and A3 (inputs had 50 vertices and 1960 arcs). Run times are in seconds. Differences and ranks are for A2 and A3.

	x_1	x_2	x_3	x_4	x_5	x_6
A1	295.9	68.5	114.3	5.0	256.1	121.3
A2	7.59	13.60	16.68	12.40	12.14	10.21
A3	1.92	6.53	18.79	9.85	4.65	8.40
A2-A3	5.66	7.07	-2.10	2.55	7.49	1.81
Rank	7	8	3	4	9	2

x_7	x_8	x_9	x_{10}	$\overline{x}$	s	g_1
1401	201.4	143.1	120.3	273	406	2.91
7.54	27.41	16.61	3.88	12.8	6.5	1.10
3.71	8.75	19.29	5.22	8.7	6.0	1.08
3.83	18.67	-2.68	-1.34	4.1	6.3	1.36
6	10	5	1			

in Table 3. The Wilcoxon signed rank test thus has test values

$$w_- = \sum_{i,sign(d_i)<0} rank(d_i) = 3 + 5 + 1 = 9\,, \tag{13}$$

$$w_+ = \frac{n(n+1)}{2} - w_- = \frac{10(10+1)}{2} - 9 = 46\,. \tag{14}$$

Since $w = max(w_+, w_-) = 46$ is lower than critical test value $w_{0.05,10} = 47$, we do not reject H_0. As $w > w_{0.1,10}$ the P-value for the test is between 0.05 and 0.1. With $\alpha = 0.1$, H_0 would have been rejected.

The test value for the Sign test, $S = |\{i : x_i < y_i\}| = 3$, is not in the region of rejection, i.e. $max(S, N - S) = 7 < 9 = S_{0.05,10}$. Thus the Sign test does not reject H_0.

Since for the difference A2-A3, $\bar{x}/s \approx 0.65$, 10 to 20 runs should be sufficient for the Student's paired t-test to differentiate A2 and A3 (see Table 1), depending on the significance level α. Taking ten sample runs more, we get mean, deviation, and skewness for A2 18.7 s, 17.3 s and 2.44 respectively. For A3 the values are 10.0 s, 7.4 s and 1.46, respectively. The Lilliefors and Jarque–Bera tests reject the normalities and thus Student's paired t-test cannot be safely used. Wilcoxon signed rank test gives test values $w_- = 26$,

$w_+ = \frac{N(N+1)}{2} - w_- = 184$ and normalized test value

$$z = \frac{max(w_+, w_-) - 0.5 - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}} \approx 2.93\,. \tag{15}$$

This is greater than $z_{0.05} = 1.96$ and H_0 is rejected. The test value for the Sign test is now $S = 4$ which is in the region of rejection for significance level α more than or equal to 0.02, i.e. $N - S = 16 \geq S_{\alpha,20}$. Therefore, we accept the alternative hypothesis that the population mean of run times of A2 and A3 are not equal. Further, the observed means and $w_+ > w_-$ and $N - S > S$ suggest that $A3$ is quicker than $A2$ for this population.

The distributions of the run times of algorithms are skewed, see Table 3. By using $n_0 = 10$ we have estimates for skewnesses (2) $g_{1,\mathrm{A1}} = 2.91$, $g_{1,\mathrm{A2}} = 1.10$ and $g_{1,A3} = 1.08$. Testing the sufficiency of the sample size for confidence interval with $\epsilon = 0.03$ and $\alpha = 0.05$ as proposed in section 2.1 gives $N > 32g_1^2$ (see Table 2). Thus we need sample sizes of at least 271, 39 and 38 for A1, A2, and A3, respectively. Let us next choose $n_{1,\mathrm{A2}} = n_{1,\mathrm{A3}} = 100$, then we get new estimates for skewnesses $g_{1,\mathrm{A2}} = 1.91$ and $g_{1,\mathrm{A3}} = 2.24$ which lead to sample sizes of at least 117 and 161. We choose $n_{2,\mathrm{A2}} = 200$ and $n_{2,\mathrm{A3}} = 250$, then the estimates are $g_{1,\mathrm{A2}} = 1.95$ and $g_{1,\mathrm{A3}} = 2.45$. These give sample sizes of at least 121 and 192 which means $n_{2,\mathrm{A2}}$ and $n_{2,\mathrm{A3}}$ are large enough. With A1 the iteration did not stop even we had a sample size of 8000 ($g_{1,\mathrm{A1}} = 22.9$). The required sample size becomes impractically large because of the high skewness.

For A2 and A3, 95% confidence intervals of means are calculated by using (4), resulting in $\mathrm{CI}_{\mathrm{A2}} = 16.9 \pm 1.8\,\mathrm{s}$ and $\mathrm{CI}_{\mathrm{A3}} = 10.0 \pm 1.1\,\mathrm{s}$. Table 4 combines the descriptive statistics of the run times of the three algorithms. For the confidence intervals, the error in the upper and lower tail probabilities is at most 0.015, i.e. the propability for the real mean to be outside the region is less than $\alpha/2 + \epsilon/2 = 0.04$.

Table 4: Descriptive statistics of A1, A2, and A3.

	Sample size	**Mean** (s)	**Deviation** (s)	**Skewness**	**CI** (s)
A1	8000	555	2600	22.9	-
A2	200	16.9	13.2	1.95	16.9 ± 1.8
A3	250	10.0	8.8	2.45	10.0 ± 1.1

4.2 Prim versus Kruskal

Consider Prim's and Kruskal's algorithms as described in Section 3 with parameters 200000 and 495000. Let us have 10 consecutive paired runs for both algorithms to compare the mean run times of algorithms. The times are as stated in Table 5. The normality test does not reject the normality of these samples. The resulting P-values for Prim and Kruskal from Jarque–Bera are 0.51, 0.57 and from Lilliefors >0.2, 0.17.

Kruskal is 3.0 ms quicker per run (or 0.7%) on average in these 10 runs. However, s_{Prim} is larger than $\bar{x}_{\text{Difference}}$ and the Student's paired t-test gives test value $t \approx \frac{3.0}{13.1}\sqrt{10} \approx 0.72$. The test value is below critical value $t_{0.05;\,10} = 2.26$ and the null hypothesis H_0 that the population mean of run times of Prim and Kruskal are equal is not rejected.

When running the Wilcoxon test it is important to notice that for one of the pairs the run times are equal, and thus we have only 9 runs to test. The test value for Wilcoxon signed rank test is $w = max(w_+, w_-) = 31.5$ which is below critical value $w_{0.05;\,9} = 40$ and therefore the observed difference is not significant. The Sign test has test value $S = 2$ which does not reject H_0.

Table 5: Ten consecutive sample runs for Prim and Kruskal (inputs had 200000 vertices and 495000 edges). Run times are in milliseconds.

	x_1	x_2	x_3	x_4	x_5	x_6
Prim	424	415	388	421	420	430
Kruskal	410	412	415	412	413	411
Difference	14	3	-27	9	7	19
Rank	7	1.5	9	4,5	3	8

x_7	x_8	x_9	x_{10}	$\overline{x}$	s	g_1
402	415	422	410	414.7	12.2	-1.23
411	412	411	410	411.7	1.5	1.14
-9	3	11	0	3.0	13.1	-1.39
4,5	1,5	6	-			

Since for the Differences $|\bar{x}/s| \approx 0.11$ is small we increase our sample size to $N = 340$ which according to Table 1 should be sufficient for verifying the statistically significant difference of means. For descriptive statistics of run times with the new sample size, see Table 6. When looking at the sample means in Table 5, Kruskal seems to be quicker than Prim. Nevertheless,

the larger data indicates the opposite. There is always such possibility of drawing wrong conclusions when they are based on the sample. However, with statistical tests we are more aware of the probabilities of errors.

Table 6: Descriptive statistics of Prim and Kruskal with $N = 340$.

	Mean (ms)	**Deviation** (ms)	**Skewness**	**CI** (ms)
Prim	409.8	13.9	-0.38	409.8 ± 1.5
Kruskal	414.1	7.1	3.12	414.1 ± 0.8
Difference	-4.3	14.5	-0.67	

The run times of Prim and Kruskal are skewed (see Table 6) and thus potentially not normal. The resulting P-values for Prim and Kruskal from Jarque–Bera are 0.001 and <0.001 and from Lilliefors <0.01. Hence, Student's paired t-test is not used for this sample. With sample size $N = 326$ (and 14 pairs equal) the test values for Wilcoxon signed rank test are $w_+ = 19181.5$ and $w_- = 34094.5$. The normalized test value is

$$z = \frac{max(w_+, w_-) - 0.5 - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}} \approx 4.37\,, \tag{16}$$

which strongly indicates that the population mean of run times of Prim and Kruskal are not equal. The P-value for the test is less than 0.01. The test value for the Sign test is $S = 183$. Using the normalized test value, we have

$$z = \frac{max(S, N-S) - 0.5 - N/2}{\sqrt{N/4}} = 2.16\,. \tag{17}$$

The Sign test also rejects H_0 with P-value between 0.02 and 0.05. The observed means and $w_- > w_+$ and $S > N - S$ support our conclusion that Prim is quicker than Kruskal for this population.

The estimates of the skewnesses (2) of run time distributions with sample size $N = 340$ are $g_{1,\mathrm{Prim}} = -0.38$ and $g_{1,\mathrm{Kruskal}} = 3.12$. The proposed sample size for confidence interval with $\epsilon = 0.03$ and $\alpha = 0.05$ is $N > 32g_1^2$ (see Table 2). The estimates give sample sizes of at least 5 and 310 which means that sample is sufficiently large for estimating the confidence intervals. The 95%-confidence intervals for Kruskal and Prim are combined in Table 6.

5 Conclusions

In this paper we showed how statistical techniques help us make inferences concerning the performance differences of combinatorial algorithms and how they also bring reliability to computational experiments. One should be able to adapt the presented methods easily to one's own needs. A natural future study on this topic would be to examine the possibilities to reduce the bounds for sample size while still maintaining the same reliability.

References

[1] Boost C++ Libraries, `http://www.boost.org/` [referenced 10.4.2008]

[2] G. Brys, M. Hubert, and A. Struyf, A Robustification of the Jarque-Bera Test of Normality, in COMPSTAT 2004 – International Conference on Computational Statistics by IASC, 753–760, Prague, Physica-Verlag, 2004.

[3] H. Cramér, Mathematical Methods of Statistics, No. 9 in Princeton Mathematical Series, Princeton University Press, Princeton, 1946.

[4] T. Höglund, Bounds for the Sample Size to Justify Normal Approximation of the Confidence Level, Ann. Inst. Statist. Math, Vol. 43, 565–578 (1991)

[5] J. L. Hodges, Jr, E. L. Lehmann, The Efficiency of Some Nonparametric Competitors of the t-Test, The Annals of Mathematical Statistics, Vol. 27, 324–335 (1956)

[6] C. M. Jarque and A. K. Bera, Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals, Econ. Lett., Vol. 6, 255–259 (1980).

[7] D. N. Joanes and C. A. Gill, Comparing measures of sample skewness and kurtosis, The Statistician, Vol. 47, 183–189 (1998).

[8] D. Jungnickel, Graphs, Networks and Algorithms, 2nd ed., Springer, Berlin, 2005.

[9] L. Kiviluoto, P. R. J. Östergård, and V. P. Vaskelainen, Algorithms for Finding Maximum Transitive Subtournaments, submitted.

[10] E. L. Lehmann, H. J. M. D'abrera, Nonparametrics: Statistical Methods Based on Ranks, Holden-Day, San Francisco, 1975.

[11] F. J. Massey, The Kolmogorov–Smirnov Test for Goodness of Fit, J. American Statistical Assoc., Vol. 46, 68–78 (1951).

[12] J. S. Milton and J. C. Arnold, Introduction to Probability and Statistics, McGraw-Hill, New York, 1995.

[13] D. Öztuna, A. H. Elhan, and E. Tccar, Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions, Turkish Journal of Medical Sciences, Vol. 36, 171–176 (2006).

[14] V. K. Rohatgi, Statistical Inference, John Wiley & Sons, New York, 1984.

[15] D. Zwillinger and S. Kokoska, CRC Standard Probability and Statistics Tables and Formulae, Chapman & Hall/CRC, Boca Raton, 2000.