Tutorial: Reporting Statistics for Experimental Computational Intelligence

Barend J. Leonard

February 13, 2012

1 Introduction

When reporting experimental results in the field of computational intelligence, it is often necessary to compare the performance of two stochastic algorithms on a given problem. While the average performance over a number of samples might be an indication of the performance of a single algorithm, comparing the averages of different algorithms is by no means scientific proof that one algorithm performs better than the other. What is needed is a way of showing that the results of independent executions of the one algorithm come from a "better" distribution than that of the other. That is, the results produced by one algorithm are *stochastically greater* than those produced by the other algorithm. Only by showing that the distribution of samples differ can one conclude that, on average, one of the algorithms does indeed perform better than the other on the given problem.

In statistics, a number of tests exist to determine various characteristics of distributions of samples. Among them are the Shapiro-Wilk test, the Student's T-Test, the Mann-Whitney U test, and the Friedman test. This tutorial will discuss the use of these tests in the context of experimental computational intelligence. Examples will also be given to show how to apply these tests in situations where algorithms are experimentally compared. The techniques shown in this tutorial should be used to report accurate, scientifically correct results in conference papers, journal articles etc.

The rest of this document is structured as follows: Section 2 explains what sample sets are and what is meant by the distribution of samples. In section 3, an example is given to statistically compare two independent sample sets, while section 4 explains how to compare multiple sample sets.

2 Sample sets and sample distributions

A sample set is a subset of values from a larger population. The sample set is a manageable size and serves as a representation of the larger population for the purpose of calculating statistics. Inferences and extrapolations can then be made from the sample set to the population.

This section describes how a sample set can be created to represent the performance of an algorithm. Furthermore, the significance of the distribution of samples is discussed.

2.1 Creating sample sets by independent algorithm execution

To create a sample set that can be used to evaluate the performance of an algorithm $\mathcal{A}$, a number of independent runs of $\mathcal{A}$ on a particular algorithm $\mathcal{P}$ must be performed. The result of each

Table 1: Sample set of fitness values for gbest PSO on Spherical in 10 dimensions.

2.961498864206842E-12	2.475012468685929E-12	8.0453364689834E-10
3.306738271675018E-10	7.213962688383421E-11	1.4916899428389076E-9
5.591854521483213E-13	3.1340216965064984E-13	6.1555006551530995E-12
2.4677607898066588E-8	2.7727505582359425E-11	2.922069823946636E-9
2.818965889983209E-13	8.881872921966085E-13	4.507794551081403E-11
3.367113702858663E-13	7.698082496373742E-6	5.570927433452115E-10
1.167127936075433E-13	1.1624463876707125E-14	3.7063777878478575E-11
3.702799909172801E-11	4.366140289511752E-12	7.330212307525685E-13
9.552324686373402E-11	3.061669896544799E-10	3.979183789016479E-11
7.653798754680398E-12	4.74389667488567E-10	3.717329062546122E-10
1.2629071221180945E-8	3.336810116813579E-13	6.51881056619394E-11
8.066746832343303E-12	4.825069756148134E-13	1.2954638447496376E-11
7.708724510348941E-14	2.7052527179615396E-10	9.664778030893562E-12
1.5423964793776043E-9	1.465201460493904E-10	1.2463998918916927E-9
9.086170902951647E-11	1.1474141737584282E-12	7.28593554370844E-11
5.1083148251635234E-14	2.538374988582531E-11	3.225702271360826E-10
7.659583992878093E-10	2.50576941307741E-8	3.0071427963141494E-9
4.150779560761687E-12	4.5706859797693943E-10	1.5428698789279826E-9
2.7910088167634803E-12	1.0269469697415547E-10	1.5148862759055265E-10
1.4738953659114527E-12	2.910421842463468E-11	3.3346023833692447E-10
1.527338493123441E-8	9.134573394377564E-14	3.938021154467198E-9
2.3567511444281244E-10	4.91436167317193E-10	8.273786844307705E-11
1.1354804623394119E-10	4.59402230935602E-11	1.1545736064473919E-9
1.7719217184179098E-7	4.6886485096102495E-12	8.152308522362393E-9
3.517287097948712E-8	3.3743679685259788E-12	8.191831141787823E-11
1.6138890349786696E-13	1.0789031482927111E-12	1.0479841340870947E-12
1.2794251747938135E-11	3.239948400323724E-12	2.1113998240473654E-10
6.473541081701952E-13	1.4678029152447757E-10	6.977913551079021E-12
4.1497488377594874E-12	1.54917556729118E-11	6.831488072114509E-9
1.336100467680096E-10	6.359871122777991E-10	8.852641473208391E-10
3.453090664236779E-12	2.4749445723759292E-12	3.379866585130356E-15
1.3719226506496524E-13	9.815591709218173E-11	2.925135325229893E-11
2.3860970053196345E-12	1.2728468242684492E-8	
1.986938185738044E-10	1.1513942008214879E-12	

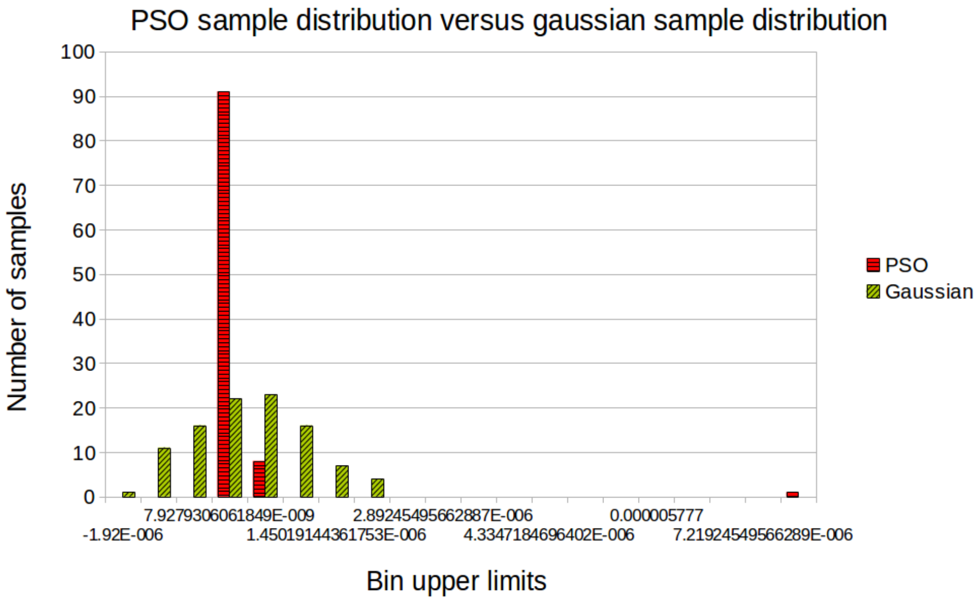

Figure 2.1: Both sample sets in this illustration have the same mean and standard deviation, but are clearly from different distributions.

independent execution of the algorithm is one sample (or observation). The set of all results obtained forms the sample set.

To clarify, consider an experiment wherein a *gbest* particle swarm optimizer is tested on a spherical problem in 10 dimensions. Because of the stochastic nature of the algorithm, independent runs of the algorithm will produce different results. For this example, the algorithm is executed 100 times and the final results (the global best fitnesses after 1000 iterations) are recorded.

The sample set of 100 independent samples is shown in table 1. This sample set represents performance of the algorithm in terms of the fitness (quality) of the final solution obtained on the Spherical problem. To evaluate other aspects of the algorithm's performance (such as diversity), additional sample sets must be created, based on the relevant performance measurements (i.e. a diversity measurement, in addition to a fitness measurement). Furthermore, to evaluate the performance of the algorithm on other problem(s), sample sets must be created using the problem(s) in question.

That is to say that the sample set created here can only be used to make assertions about the performance of a *gbest PSO* on the *Spherical problem* in *10 dimensions*. No claims can be made regarding the performance of PSO on other problems, or regarding variants of PSO on the same problem based on this data.

2.2 Distribution of samples

When comparing two algorithms, it must be shown that the observations in the sample sets generated by the respective algorithms are distributed differently. That is, the probability of samples being close to a certain value, or falling within a certain range differs for the two sample sets. By showing that the observations from one sample set has a higher probability of being close to a desired value, one can assert that the algorithm that produced the better sample set is the better algorithm.

Figure 2.1 shows the binned frequency distribution histograms for the PSO sample set in table 1, as well as a normally (Gaussian) distributed sample set. The Gaussian samples were generated for illustration purposes. The two sample sets have the same mean (8.04×10^{-8}) and standard deviation ($7.6987 \times 10 - 7$) and both contain 100 samples. The bins range from -1.92×10^{-6} to 7.2192×10^{-6} and each bin covers a range of 4.8075×10^{-7}. The reason for the small scale is that most of the PSO samples are distributed extremely close together. Even on this scale, 91% of the PSO samples still fall within the range $(-4.777 \times 10^{-7}, 3.018 \times 10^{-9})$, with 8% in the neighbouring interval, and 1% in the last interval. In contrast, the frequency histogram of the normally distributed samples resembles a Gaussian probability distribution function, centred on the mean.

In the next section an example is given, where the results from two stochastic algorithms are analysed to determine whether or not there is a statistically significant difference between their underlying distributions.

3 Statistical comparison of two samples sets

When comparing sample sets, the goal is to determine whether there is a significant difference between the underlying distributions of the sample sets. A statistical test known as the Student's T-test is a well-known method to show that the distributions of two sample sets differ. However, the Student's T-test is a *parametric* test, meaning that it assumes that the samples in both sample sets are normally distributed. As was seen in section 2.2, this isn't the case for the sample set generated by PSO on a Spherical problem. It is reasonable to assume that samples generated by any optimization algorithm will not be normally distributed. However, tests do exist to verify that samples are normally distributed. To compare two sample sets for which the underlying distributions are unknown, a *non-parametric test* must be used.

This section explains how to test a sample set for normality with the Shapiro-Wilk test, and gives an example of how to use the Mann-Whitney U test to compare two sample sets.

3.1 Testing for normality with the Shapiro-Wilk test (optional)

For the Shapiro-Wilk test, the null hypothesis is that a sample set $x_1, \dots, x_n$ is normally distributed. The test statistic is calculated using the following formula:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{3.1}$$

where $x_{(i)}$ is the i^{th} *order statistic* (the i^{th} smallest number in the sample set), $\bar{x}$ is the mean of the sample set, a_i is given by

$$(a_1, \dots, a_n) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}}, \tag{3.2}$$

where

$$m = (m_1, \dots, m_n)^\top \tag{3.3}$$

and $m_1, \dots, m_n$ are the expected values of the order statistics of independent, but identically distributed random variables sampled from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, and V is the covariance matrix of those order statistics.

For small values of W, the null hypothesis can be rejected. An online implementation of the Shapiro-Wilk test can be found at `http://dittami.gmxhome.de/shapiro/`.

3.2 Comparing two sample sets with the Mann-Whitney U test

The Mann-Whitney U test is a non-parametric test to determine whether observations from one sample set are stochastically greater than observations from another sample set. That is, it determines whether the means of the underlying distributions differ.

The test calculates a U statistic whose distribution under the null hypothesis is known. For sample set sizes above ~ 20, the distribution of U is approximately a normal distribution with mean

$$\mu_U = \frac{n_1 n_2}{2} \tag{3.4}$$

and standard deviation

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \frac{n_1 n_2 \left[\sum_{i=1}^{s} (t_i^3 - t_i)\right]}{12(n_1 + n_2)(n_1 + n_2 - 1)}} \tag{3.5}$$

where s is the number of sets of ties, and t_i is the number of ties in sample set i. In cases where there are not many ties, equation (3.5) can be simplified to

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}. \tag{3.6}$$

For this example, two PSO algorithms were used to optimize a Rastrigin function in 30 dimensions. The first PSO algorithm used the Von Neumann neighbourhood topology, while the second algorithm used the star (gbest) topology. Both algorithms executed for 1000 iterations, and 30 independent samples were performed in each case. The fitness of the best particle after 1000 iterations for each of the samples are shown in table 2.

What follows is a step-by-step example, comparing the two PSO algorithms described above on the Rastrigin problem in 30 dimensions over 1000 iterations.

3.2.1 State the null hypothesis and alternative hypothesis

Under the null hypothesis H_0, the underlying distributions of the two sample sets are equal. The alternative hypothesis H_1 is then that one distribution is stochastically greater than the other. The Mann-Whitney U test assumes that all observations from both sample sets are independent from each other, and that the observations are ordinal. H_0 and H_1 can be stated formally as follows:

$$H_0 : \mu_1 = \mu_2 \tag{3.7}$$

$$H_1 : \mu_1 \neq \mu_2 \tag{3.8}$$

where μ_1 and μ_2 are the means of the underlying distributions of the two sample sets. Note that H_1 does not specify whether μ_1 is greater than or less than μ_2, but simply that they are not equal. For this hypothesis, the Mann-Whitney U test performs what is referred to as a *two-tailed test*. A *one-tailed test* can be performed by changing H_1 accordingly:

$$H_1 : \mu_1 < \mu_2. \tag{3.9}$$

To show that one algorithm is stochastically better than the other, a one-tailed test will be used. Hypothesis (3.7) and (3.9) will therefore be used as H_0 and H_1, respectively.

Table 2: Results From Independent PSO Samples

Von Neumann (S_1)	**Rank**	**GBest (S_2)**	**Rank**
34.8235411672	1	36.8134349602	2
38.8055693611	3	46.763020222	6
41.7882311226	4	54.722677775	11
42.7831929441	5	55.7175507996	12
50.7428466719	7	55.7176012668	13
51.7377905156	8	58.7024885794	17
53.1914061372	9	60.6924066774	18
53.7984643444	10	62.6823047306	19
55.8026744295	14	63.6772586103	20
56.712683432	15	68.65204382	25
58.7024742667	16	71.6369210748	30
64.6724504651	21	72.6318496061	33
64.6735561153	22	72.6318954023	34
66.6620979579	23	73.6268340677	36
66.6621811737	24	75.6166914164	37
69.6982576925	26	75.6167320495	38
70.6419625861	27	76.6116302693	39
70.6424964176	28	77.6065235192	40
70.6431262018	29	77.6065489331	41
71.6369228238	31.5	78.6015663816	43
71.6369228238	31.5	81.5864298897	45
73.6268211998	35	81.5865290086	46
78.6015647735	42	82.5812524367	48
80.5914993612	44	88.5511188029	50
81.5865532549	47	93.5258686585	51
88.551099076	49	95.5156443531	52
99.5544516173	54	96.5106800446	53
105.4652045942	56	101.485517817	55
109.4452833577	58	105.4652249825	57
113.4248978039	59	117.404790114	60
Sum of ranks (R_1):	799	**Sum of ranks (R_2):**	1031

3.2.2 Determine the significance level

A result is considered *statistically significant* if it is unlikely to have occurred by chance. In order to reject H_0, a level of confidence must be established. That is, the amount of evidence required to accept that the observed results happened by chance must be set. This threshold is known as the *significance level* and is denoted by α.

Common choices for alpha include 0.1, 0.05 and 0.01. If a test of statistical significance produces a value less than α, H_0 is rejected at a confidence level of $1 - \alpha$. For this example, a significance level of $\alpha = 0.05$ will be used.

3.2.3 Calculate the U statistic

The U statistic is calculated as follows:

1. Let S_1 be the sample set for the Von Neumann PSO, S_2 be the sample set for the gbest PSO, n_1 and n_2 be the respective sizes of S_1 and S_2, R_1 be the sum of ranks for S_1, and R_2 be the sum of ranks for S_2.

2. Sort samples in S_1.

3. Sort the samples in S_2.

4. Merge S_1 and S_2 to form a single sorted list.
   (Note: In table 2, the lists were unmerged after ranking was done (see step 5) to make the table easier to understand.)

5. Rank each element and resolve ties.
   (Note: In table 2, there is one tie between elements 31 and 32. Both samples are therefore awarded a rank of 31.5.)

6. Compute R_1 and R_2.

7. Compute U_1 and U_2 as

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \\ &= (30)(30) + \frac{30(31)}{2} - 799 \\ &= 1365 - 799 \\ &= 566 \end{aligned} \tag{3.10}$$

   and

$$\begin{aligned} U_2 &= n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \\ &= 1365 - 1031 \\ &= 334. \end{aligned} \tag{3.11}$$

8. Verify that

$$\begin{aligned} n_1 n_2 &= U_1 + U_2 \\ (30)(30) &= 566 + 334 \\ 900 &= 900. \end{aligned} \tag{3.12}$$

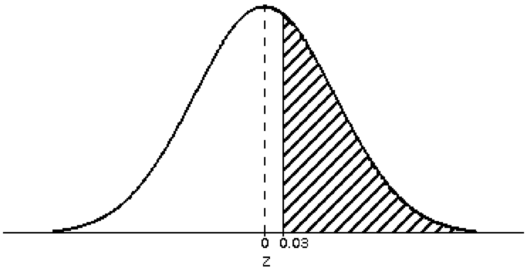

Figure 3.1: Significance of the z-score.

9. Compute the U statistic as

$$U = min\{U_1, U_2\} = 334. \tag{3.13}$$

The values U_1 and U_2 are indications of the number of times samples from one sample set precede or follow samples from the other sample set when all the samples are merged into a sorted list. Recall that the distribution of U is known to be approximately normal under the null hypothesis. Therefore, by standardizing the U value calculated above, the probability of obtaining this specific U value from the normal distribution can be computed.

3.2.4 Compute the standard score (z-score) and find the p-value

The standardized value of U is given by

$$\begin{aligned} z &= \frac{U - \mu_U}{\sigma_U} \\ &= \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \\ &= \frac{-116}{4575} \\ &= -0.025355191 \\ &\approx -0.03. \end{aligned} \tag{3.14}$$

The z-score is approximately the standard normal deviate corresponding to U. That is, if U is a normally distributed random variable with mean μ_U and standard deviation σ_U, then z is the corresponding value from a normal distribution with mean 0 and standard deviation 1. By finding the area under the normal curve above $|z|$, the probability that $|Z|$ will be higher than or equal to $|z|$ is obtained, where Z is a random variable sampled from $U(0, 1)$. This probability is illustrated in figure 3.1 and can be found on the standard normal distribution table (table 3). To read the table, the first two digits of $|z|$ (0.0) must be located in the first column. Then, the second decimal digit of $|z|$ (0.03) must be located in the first row.

The value corresponding to this row and column is the area under the curve and is referred to as the significance p of z. In this case, $p = 0.488$. The p-statistic is linearly related to U and is often reported instead of z-scores. The value of $p \in [0, 1]$ is a non-parametric measure of the overlap between two distributions and gives the probability that the observed difference in the means of the two sample sets occurred by chance. Thus, H_0 is rejected if $p < \alpha$.

Table 3: Standard Normal Distribution Table

$\|z\|$	0	0.01	0.02	0.03	0.04	**0.05**	0.06	0.07	0.08	0.09
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.1	0.001	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.001	0.001
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.002	0.0019
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.003	0.0029	0.0028	0.0027	0.0026
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.004	0.0039	0.0038	0.0037	0.0036
2.5	0.0062	0.006	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.4	0.0082	0.008	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.011
2.1	0.0179	0.0174	0.017	0.0166	0.0162	0.0158	0.0154	0.015	0.0146	0.0143
2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.025	0.0244	0.0239	0.0233
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	**0.0495**	0.0485	0.0475	0.0465	0.0455
1.5	0.0668	0.0655	0.0643	0.063	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.102	0.1003	0.0985
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.123	0.121	0.119	0.117
1	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.166	0.1635	0.1611
0.8	0.2119	0.209	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.7	0.242	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.5	0.3085	0.305	0.3015	.2s981	0.2946	0.2912	0.2877	0.2843	0.281	0.2776
0.4	0.3446	0.3409	0.3372	0.3336	0.33	0.3264	0.3228	0.3192	0.3156	0.3121
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.352	0.3483
0.2	0.4207	0.4168	0.4129	0.409	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0	0.5	0.496	0.492	0.488	0.484	0.4801	0.4761	0.4721	0.4681	0.4641

The value of p can be estimated by dividing U by its maximum value such that

$$\begin{aligned} p &= \frac{U}{n_1 n_2} \\ &= \frac{334}{900} \\ &\approx 0.371. \end{aligned} \tag{3.15}$$

It is worth observing that the values in table 3 get smaller as $|z|$ increases. The first value for $|z|$ that results in a value for $p < \alpha$ is 1.65. This value is called the *critical value* of $|z|$ at a 0.05 level of significance. H_0 can therefore be rejected if

$$|z| \geq 1.65. \tag{3.16}$$

Similarly, for a significance level of $\alpha = 0.01$, H_0 can be rejected if

$$|z| \geq 2.33. \tag{3.17}$$

Note that because $U_1 + U_2 = n_1 n_2$, the mean that is used in the normal approximation is the mean of the two U values. The value of $|z|$ will therefore be the same, regardless of which U value is used.

3.2.5 Rejcet or accept H_0

Once the statistics have been calculated, the null hypothesis must be accepted or rejected. In this example, both the z-score and the p-value indicated that there was not enough evidence to reject H_0 at a 95% confidence level. H_0 is therefore accepted, meaning that the results obtained in this experiment does not show a statistically significant difference in the performance of the two algorithms.

In the case where the H_0 was rejected, that would have indicated a significant difference in the algorithms' performance. The two U values, computed in step 7 of section 3.2.3 could then have been used to determine which algorithm was better by observing which U value was lower. In the case of a minimization problem, the sample set with the lowest U value is the better sample set. The opposite is true for maximization problems.

3.2.6 Report the findings

When reporting statistical findings, at least the following must be done:

- State the null hypothesis and the alternative hypothesis.
- Report either p-values or z-scores for all sample sets being compared.
- Reject or accept H_0 for all sample sets being compared.
- In the case where H_0 is rejected, state which algorithm is better by referring to the U values.
- Statistical findings should be reported in addition to other results, such as fitness profiles, diversity profiles, etc.

Table 4: Average Results From Independent PSO Samples

	Gbest	**Rank**	**Lbest**	**Rank**	**Von Neumann**	**Rank**
Griewank	8.29E-002	3	7.32E-003	1	1.60E-002	2
Norwegian	-2.00E+014	1	-1.27E+003	3	-3.21E+012	2
Schaffer2	1.13E+001	1	1.16E+001	3	1.13E+001	2
R_j		1.67		2.33		2

4 Statistical comparison of three or more sample sets

To compare the performance of multiple algorithms on multiple problems, the *Friedman test* is used. The Friedman test is a non-parametric test, like the Mann-Whitney U test, meaning that it does not assume that the distribution of samples is normal. If the Friedman test shows that there is a significant difference between two or more of the sample sets, a post-hoc test, known as the *Nemenyi test*, can be performed to identify which sample sets differ.

This section demonstrates how to use the Friedman- and Nemenyi tests to compare the performance of three algorithms on three problems.

4.1 Rank the algorithms for the Friedman test

The Friedman test calculates a statistic χ_F^2, whose distribution is known under the null hypothesis, which states that the distributions of samples in the various sample sets are equal.

For this example, three PSO algorithms are tested on three problems. The three algorithms make use of the star- (gbest), lbest-, and Von Neumann topologies, respectively. Each algorithm executed for 1000 iterations on the Griewank, Norwegian, and Schaffer2 functions. To perform a Friedman test, the average results over 30 samples were recorded and are shown in table 4. The algorithms are then ranked for each problem such that the best performing algorithm receives a rank of 1 and the worst performing algorithm receives a rank of k, where k is the number of algorithms. Then, the average rank R_j for each of the algorithms is calculated as follows:

$$R_j = \frac{1}{N}\sum_{i=1}^{k} r_i^j, \tag{4.1}$$

where N is the number of problems, and r_i^j is the rank of algorithm j on problem i.

4.2 Determine the significance level

As for the Mann-Whitney U test, a significance level α must be chosen. See section 3.2.2 for details. For this example, a significance level of $\alpha = 0.05$ will be used.

Table 5: Nemenyi Test Values for $q\alpha$

#Algorithms	2	3	4	5	6	7	8	9	10
$q(0.05)$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$q(0.1)$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

4.3 Calculate the Friedman statistic

The Friedman statistic is given by

$$\begin{aligned}\chi_F^2 &= \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] \qquad (4.2)\\ &= \frac{12(3)}{3(4+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{3(3+1)^2}{4}\right]\\ &= 2.4\,[12.2178 - 12]\\ &= 0.52272.\end{aligned}$$

However, this statistic is known to be undesirably conservative. The following statistic is therefore often used instead:

$$\begin{aligned}F_F &= \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \qquad (4.3)\\ &= \frac{(3-1)0.52272}{3(3-1)-0.52272}\\ &= 0.19086846.\end{aligned}$$

The value of χ_F^2 or F_F is similar in meaning to the U statistic used in the Mann-Whitney U test. That is, it is a measure of the overlap between the different distributions. If all values R_j are equal for $j = 1 \ldots k$, the distributions are the same.

4.4 Check the significance of the Friedman statistic

Which table? Are there different tables for the χ_F^2 and F_F statistics?

If the value is found to be significant, the null hypothesis can be rejected.

4.5 Perform the Nemenyi test

If the null hypothesis was rejected, a Nemenyi test can be performed to find out which algorithms are better that which. The performance of two algorithms are deemed to be significantly different if their average ranks differ by at least the critical difference

$$CD = q\alpha\sqrt{\frac{k(k+1)}{6N}}, \qquad (4.4)$$

where $q\alpha$ can be found in table 5 for a given number of algorithms and a significance level of 0.05 or 0.1.

4.6 Report the findings

When reporting statistical findings, at least the following must be done:

- State the null hypothesis and the alternative hypothesis.
- Report the average ranks as well as the X_F^2 or F_F value.
- Reject or accept H_0 for all sample sets being compared.
- In the case where H_0 is rejected, perform a Nemenyi test to identify which algorithms differ significantly.
- Use the average ranks of algorithms to determine which algorithm is better.
- Statistical findings should be reported in addition to other results, such as fitness profiles, diversity profiles, etc.