

Performance Evaluation

Winter 2024

How good is the performance?

- Most meta-heuristics possess stochastic natures
(performance varies from time to time)
- It is difficult to obtain an analytical prediction of meta-heuristics to assess
 - either **the solution** achievable within a given computation time
 - or the **time taken** to find a solution of a given quality (convergence)

Fair Comparison

- **Verification** is often necessary to compare the performance of a new algorithm with existing ones
- Comparison can be made in terms of four issues
 - **Effectiveness** (Quality of solutions)
 - **Efficiency** (Speed of convergence)
 - **Consistency** (Variance of performance)
 - Robustness (Applicability to different problems)

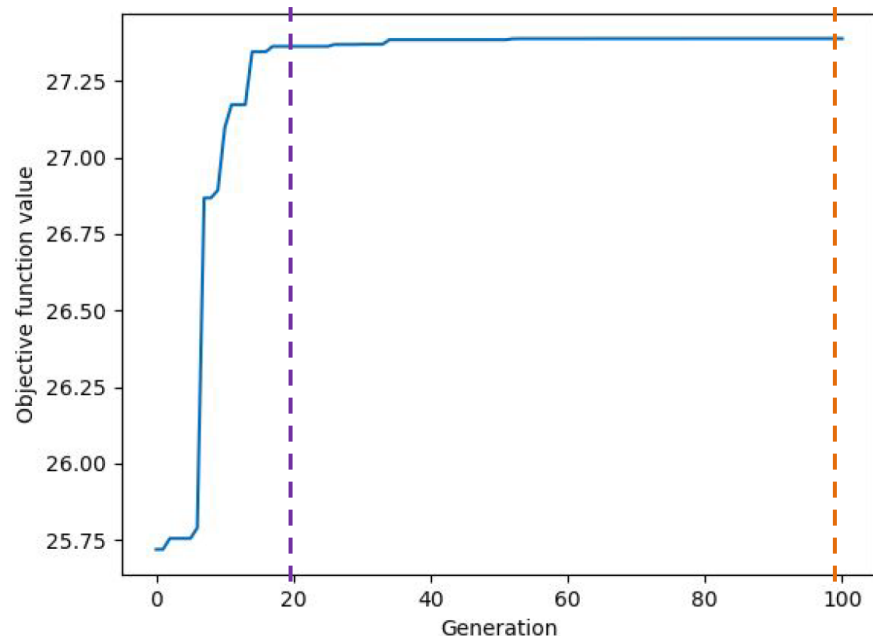


Experimentation

- Performance assessments are best carried out by **experimentation**
- Measures of performance: **solution quality** and **running time** can be seen as random variables
- We are also interested to know whether the performance is **consistent** (with a smaller variance)

Solution quality

- Quality of solution can be measured by the **objective values reached at the end**
- Compare the solution quality within **a fixed period of computation time (generation/iteration)**



Running time



- The computation time may depend on programming skill, hardware and software used
- It would be better to report the **number of objective function calls** instead
(usually $= \text{iterations} * \text{population size}$)
- However, if **communication time** among processors has to be considered, **clock-time** may be used
- Indicate computing resources
e.g., a Core i9 processor at 3.3 GHz with 16 GB of RAM

Use of statistics

- Statistics offers the advantage of providing
 - a systematic framework for the collection and evaluation of data, thus maximizing the objectivity and replicability of experiments
 - a mathematical foundation that supplies a probabilistic measure of events on the basis of inference from the empirical data
- It takes **multiple experiments** to confirm the advantage

Characterization

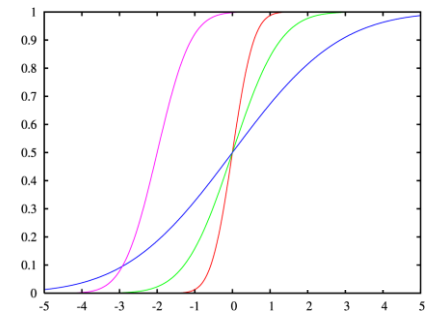
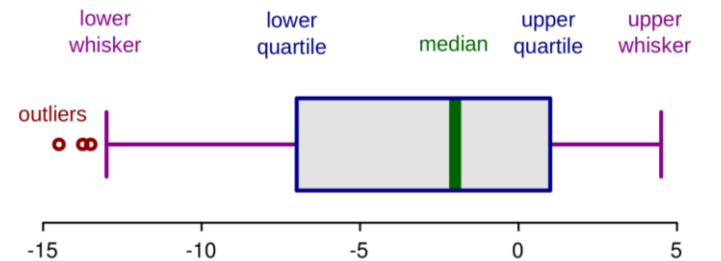
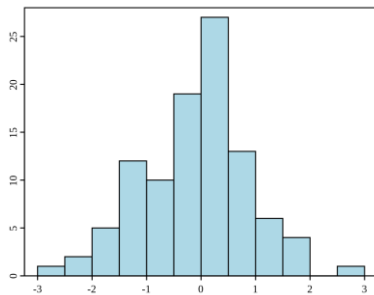
- The performance measure X (solution or running time) of a meta-heuristic on a single instance can be described by its probability distribution or cumulative distribution function

$$p(x) = \Pr [X = x] \qquad F(x) = \Pr [X \leq x] = \sum_{x_i \leq x} p(x_i).$$

- Alternatively, some parameters of the probability distribution are known, e.g., **mean and variance**

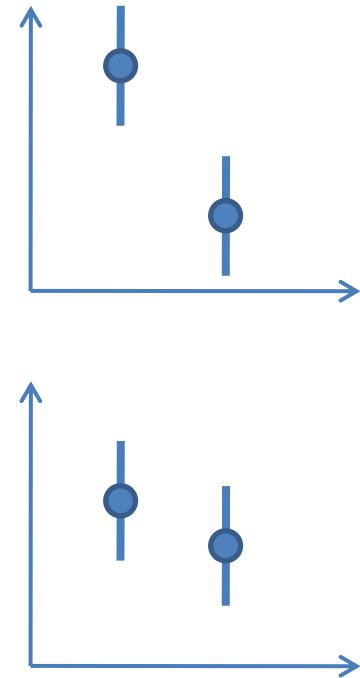
Summary measures

- Summary measures for sample data are divided into measures of locations (sample mean and q-quantiles), and measures of dispersion (sample variance).
- Histograms, boxplots and CDF plots are used to provide a more complete view of the data.
- Summary measures tend to hide part of the information contained in the sample data



Rule of thumb

- Compare variables X and Y
- If the confidence intervals do not overlap, X and Y are **significantly different**.
- If the mean of one variable lies within the other variable's confidence interval, **the difference is NOT significant**.



Comparison: descriptive statistics

- It is necessary to test the **performance difference** of multiple meta-heuristics
- Descriptive statistics (**sample mean** and **standard deviation**) are fundamental but often inadequate.

Comparison: descriptive statistics

- Multiple runs are conducted to minimize an objective function
 - Meta-heuristic A has mean=100, std_dev=10
 - Meta-heuristic B has mean=105, std_dev=4
 - Which one is better?

Hypothesis test

- Other than descriptive statistics, we may perform hypothesis tests
- If the test does not reject the absence of differences, the analyst should
 - either collect more data in order to increase the power of the test and detect also small differences or
 - stop if differences are insignificant

Hypothesis testing

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Typical data	Scale	Ordinal or Nominal Scale as well
Usual central measure	Mean	Median (associated with a particular run)

Parametric hypothesis

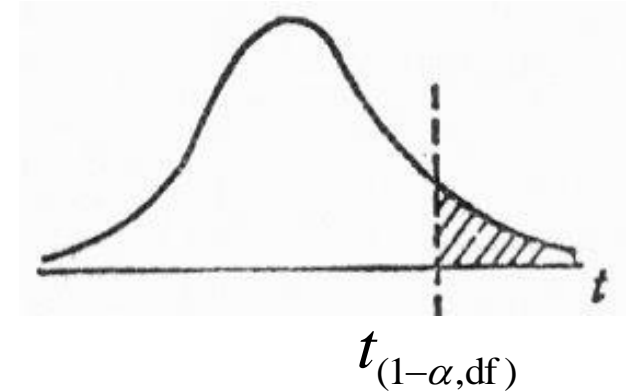
- Parametric hypothesis testing (Z-test)
- The parametric tests have specific assumptions
 - Underlying distributions are Normal
 - Population variance is known
 - Variances of the populations are assumed equal

Student t test

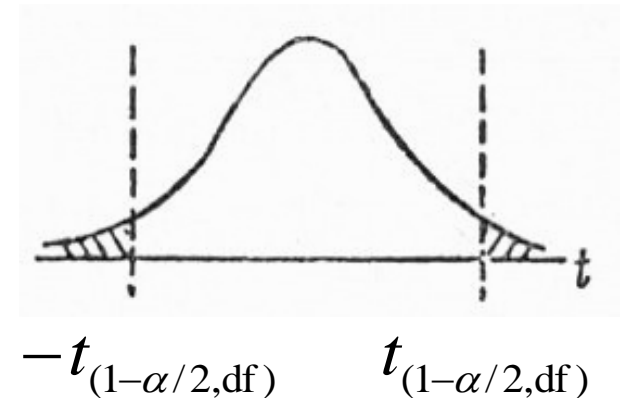
- A test of the null hypothesis that the means of two normally distributed populations are equal
- **Useful when the number of samples is small**, which is usually the case when comparing different meta-heuristic algorithms

One-tailed and Two-tailed tests

- One-tailed:
 - Reject H_0 ($H_0: \mu_A - \mu_B \leq 0$) if
$$t > t_{(1-\alpha, df)}$$



- Two-tailed:
 - Reject H_0 ($H_0: \mu_A - \mu_B = 0$) if
$$t > t_{(1-\alpha/2, df)} \text{ or } t < -t_{(1-\alpha/2, df)}$$



Student t test (equal variance)

**t STATISTIC FOR SMALL INDEPENDENT
SAMPLES (EQUAL VARIANCES)**

$$t = \frac{(\bar{X}_A - \bar{X}_B) - D_0}{s_D}$$

using

$$s_D = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

with $df = n_A + n_B - 2$.

\bar{X}_A : sample mean of variable A

$D_0 : \mu_A - \mu_B$

s_A^2 : sample variance of variable A

n_A : number of samples for variable A

Student t test (unequal variance)

t STATISTIC FOR SMALL INDEPENDENT
SAMPLES (UNEQUAL VARIANCES)

$$t = \frac{(\bar{X}_A - \bar{X}_B) - D_0}{s_D}$$

using

$$s_D = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

with

$$df = \frac{(s_A^2/n_A + s_B^2/n_B)^2}{(s_A^2/n_A)^2/(n_A - 1) + (s_B^2/n_B)^2/(n_B - 1)}$$

(rounded to nearest integer)

t distribution

Degrees of Freedom	Upper-Tail Area α									
	.4	.25	.1	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

How do we know if variances are equal?

- We need to know whether the variance is statistically different before t-test
- F-test: compare the variances of two samples

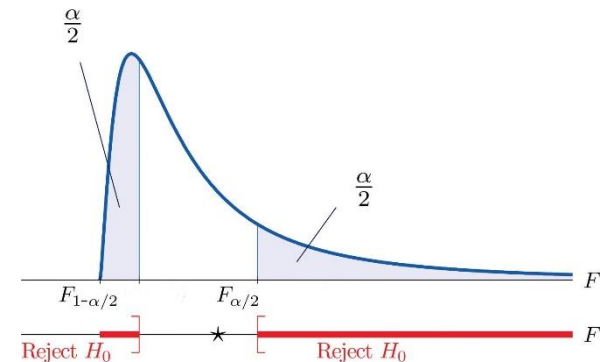
- $H_0: \sigma_A = \sigma_B$

- $H_1: \sigma_A \neq \sigma_B$

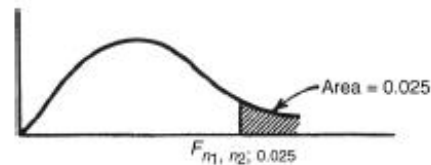
- F statistics: $F = \frac{s_A^2}{s_B^2}$

- Reject H_0 if

$$F \geq F_{(\alpha/2; n_1-1, n_2-1)} \text{ or } F \leq F_{(1-\alpha/2; n_1-1, n_2-1)}$$

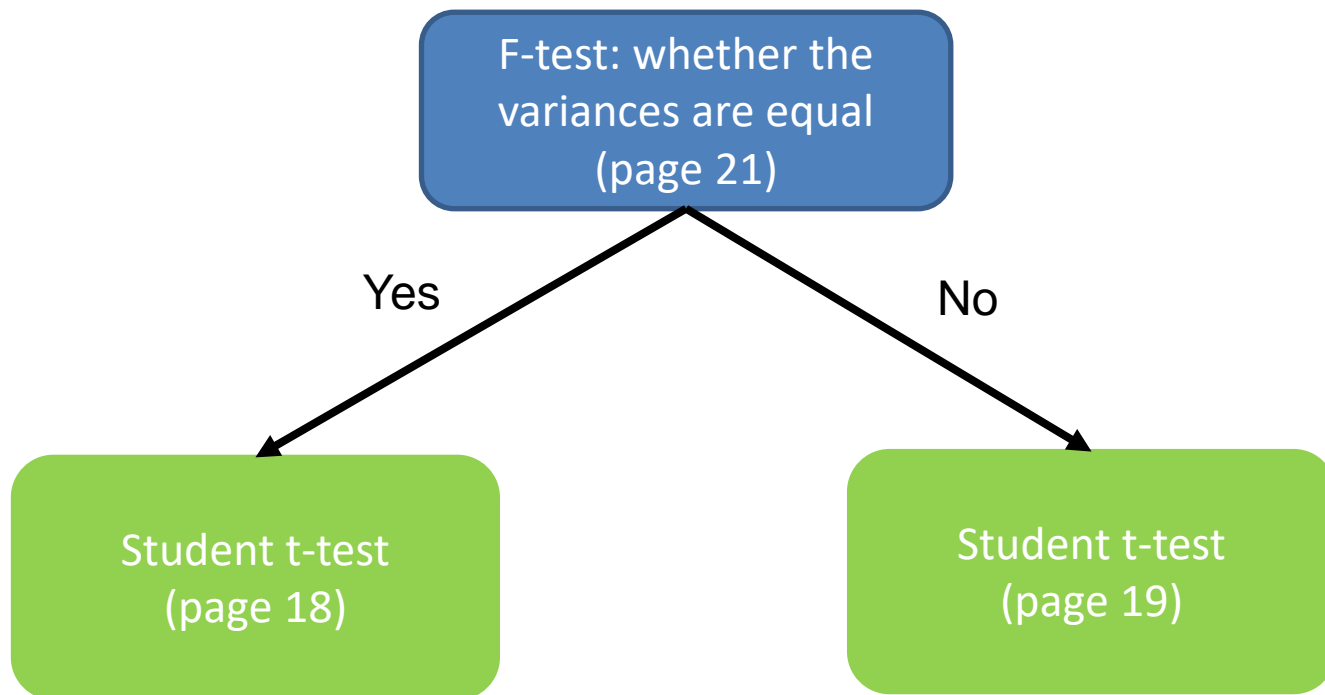


Values of $F_{n_1, n_2, 0.025}$ such that $\text{Prob}[F_{n_1, n_2} > F_{n_1, n_2, 0.025}] = 0.025$



$n_1 \backslash n_2$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16
1	648	800	864	900	922	937	948	957	963	969	973	977	980	983	987
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.4	14.3	14.3	14.3	14.2
4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.72	8.69	8.64
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.49	6.46	6.41
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.33	5.30	5.25
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.63	4.60	4.54
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.16	4.13	4.08
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.83	3.60	3.74
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.58	3.55	3.50
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.39	3.36	3.30
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.24	3.21	3.15
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.12	3.08	3.03
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	3.01	2.98	2.92
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.85	2.82	2.76
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.73	2.70	2.64
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.64	2.60	2.55
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.56	2.53	2.47
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.50	2.47	2.41
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.54	2.49	2.45	2.42	2.36
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.49	2.45	2.41	2.37	2.32
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.37	2.34	2.28
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.25	2.21	2.15
50	5.34	3.98	3.39	3.06	2.83	2.67	2.55	2.46	2.38	2.32	2.26	2.22	2.18	2.14	2.08
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.13	2.09	2.03
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.36	2.38	2.21	2.16	2.11	2.07	2.03	1.97
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	2.12	2.08	2.04	2.00	1.94
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	2.06	2.01	1.97	1.93	1.87
500	5.05	3.72	3.14	2.81	2.59	2.43	2.31	2.22	2.14	2.07	2.02	1.97	1.93	1.89	1.83
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.99	1.94	1.90	1.87	1.80

Flowchart of t-test



Student t test (Example)

- Two algorithms are used to solve a maximization problem
- Among 12 trials, the solutions of algorithm A has mean = 85, variance = 16;
- Among 10 trials, the solutions of algorithm B has mean = 81; variance = 25.
- Suppose both solutions are normally distributed, check whether algorithm A is superior to B at $\alpha = 0.05$

Student t test (Example Solution)

- First, check if the variances are equal
 - $H_0: \sigma_A = \sigma_B$
 - $H_1: \sigma_A \neq \sigma_B$
 - Two-tailed: $F = 16/25 = 0.64$

$$F_{\underline{(0.025;11,9)}} = \boxed{3.91} = 1/F_{(0.975;9,11)} \quad F_{(0.975;11,9)} = 0.28$$

Two-tailed (Page 22)

- $0.28 < 0.64 < \boxed{3.91}$
- Cannot reject H_0 ; variances are equal

Student t test (Example Solution)

Continued

- Then, check if the solutions of A is greater than those of B, assuming **variances are equal**

- $H_0: \mu_A - \mu_B \leq 0$

12 trials, A has mean = 85, variance = 16;

- $H_1: \mu_A - \mu_B > 0$

10 trials, B has mean = 81; variance = 25.

$$s_D^2 = \frac{(12-1)*16 + (10-1)*25}{12+10-2} \left(\frac{1}{12} + \frac{1}{10} \right) = 3.676$$

$$t = \frac{(85-81)-0}{\sqrt{3.676}} = 2.086 > t_{(0.95, 12+10-2)} = 1.725 \quad (\text{Page 20})$$

One-tailed

- Reject H_0 ; **solutions of A are significantly greater than B**

Non-parametric methods

- If **normality may not be assumed**, non-parametric methods can be used
 - Wilcoxon signed rank test (samples are related)
 - Mann-Whitney test (samples are independent)
- If normality is acceptable, use **student t-test** on the ranks

Wilcoxon signed rank test

Procedure (1)

- Suppose we collect $2n$ observations, two observations of each of the n subjects
- Let i denote the particular subject that is being referred to
- The first observation measured on subject i be denoted by x_i and second observation be y_i .
- Paired data means that the values in the two groups being compared are naturally linked

Wilcoxon signed rank test

Procedure (2)


- Calculate the difference between each pair of observations.
- **Rank the differences** by the absolute value, ignoring the sign, giving 1 for the smallest difference, 2 for the next smallest and so on.
- **Sum the ranks** of the positive differences and sum the ranks of the negative differences.

Wilcoxon signed rank test

Procedure (3)

- The test statistic T is **the lesser** of these two sums (in absolute value).
- If the null hypothesis were true and there was no difference, we would expect the rank sums for positive and negative ranks **to be close**. That is, **reject the null hypothesis if T is too small** (smaller than the critical value)
- $n \leq 20$, use Wilcoxon table

Wilcoxon signed rank test (Table)

 More significant

n	One-side Alpha		
	0.01	0.025	0.05
5			1
6		1	2
7	0	2	4
8	2	4	6
9	3	6	8
10	5	8	11
11	7	14	14
12	10	17	17
15	20	25	30
20	43	52	60

Wilcoxon signed rank test (Example)

Obser.	X	Y	Diff=X-Y	Rank	Signed Rank
1	54.5	48.6	5.9	8	8
2	54.3	47.2	7.1	9	9
3	53.7	50.1	3.6	5	5
4	47.8	49.3	-1.5	1	-1
5	64.5	56.1	8.4	10	10
6	60.8	55.6	5.2	7	7
7	45.6	47.2	-1.6	2	-2
8	51.4	49.2	2.2	3.5	3.5
9	53.8	49.4	4.4	6	6
10	45.9	48.1	-2.2	3.5	-3.5

$H_0: X \leq Y$; $H_1: X > Y$ (One-tailed test)

Sum of positive ranks = 48.5; sum of negative ranks = -6.5; So, $T_s = 6.5$.

According to the Table, critical value at $\alpha = 0.05$ $T_{crit} = 11$. (Page 30)

Since $T_s < T_{crit}$, **reject the null hypothesis**

Conclusion: **X finds larger solutions than Y** at $\alpha = 0.05$

Wilcoxon signed rank test (Example)

- If we change

critical value at $\alpha = 0.01$, $T_{crit} = 5$ (Page 30)

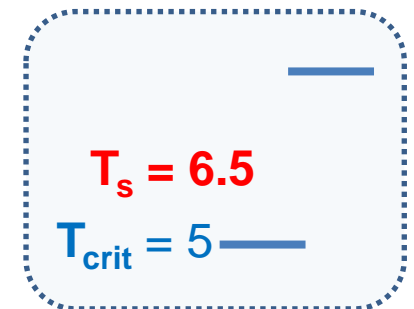
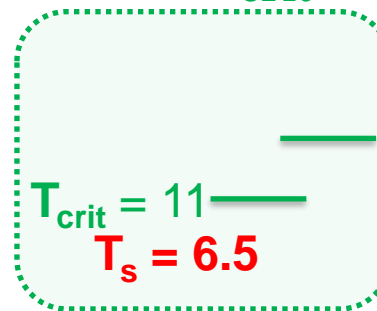
$T_s > T_{crit}$, cannot reject the null hypothesis

Not enough evidence to show that X is better than Y at $\alpha = 0.01$

- In such a situation, the condition for us to reject the null hypothesis is that T_s must be smaller than 5

- $T_{crit} = 5$ is more restrictive than $T_{crit} = 11$

$T_s = 6.5$ can prove superiority
for $T_{crit} = 11$, not $T_{crit} = 5$
 $\alpha = 0.05$ $\alpha = 0.01$



Mann-Whitney test

- If the samples are not paired, use Mann-Whitney test
- It can be used to perform
 - One-tailed test: Whether one variable tends to be higher (lower) than the other
 - Two-tailed test: Whether one variable is significantly different from the other
 - In our applications, we tend to use the one-tailed test

Mann-Whitney test procedure (1)

- Place all the values together in rank order (i.e. from lowest to highest).
- Inspect each B sample in turn and count the number of A's which precede (smaller than) it. Add up the total to get a U value (U_1).
- Inspect each A in turn and count the number of B's which precede it. Add up the total to get a second U value (U_2).

Mann-Whitney test procedure (2)

- Take the smaller of the two U values
- Look up the critical value in the table below.
- If $P(U \leq u \mid H_0 \text{ is true}) < \alpha$, reject H_0
- U_1 and U_2 can be conveniently computed as

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - W_2$$

W_1 and W_2 are sums of ranks

$$\text{Check } U_1 + U_2 = n_1 n_2$$

Mann-Whitney test (Table_1)

$$P(U \leq u \mid H_0 \text{ is true})$$

$n_2 = 4$

u	n_1			
	1	2	3	4
0	0.200	0.067	0.028	0.014
1	0.400	0.133	0.057	0.029
2	0.600	0.267	0.114	0.057
3		0.400	0.200	0.100
4		0.600	0.314	0.171
5			0.429	0.243
6			0.571	0.343
7				0.443
8				0.557

$n_2 = 5$

u	n_1				
	1	2	3	4	5
0	0.167	0.047	0.018	0.008	0.004
1	0.333	0.095	0.036	0.016	0.008
2	0.500	0.190	0.071	0.032	0.016
3	0.667	0.286	0.125	0.056	0.028
4		0.429	0.196	0.095	0.048
5		0.571	0.286	0.143	0.075
6			0.393	0.206	0.111
7			0.500	0.278	0.155
8			0.607	0.365	0.210
9				0.452	0.274
10				0.548	0.345
11					0.421
12					0.500
13					0.579

Mann-Whitney test (Table_2)

$n_2 = 6$

u	n_1					
	1	2	3	4	5	6
0	0.143	0.036	0.012	0.005	0.002	0.001
1	0.286	0.071	0.024	0.010	0.004	0.002
2	0.428	0.143	0.048	0.019	0.009	0.004
3	0.571	0.214	0.083	0.033	0.015	0.008
4		0.321	0.131	0.057	0.026	0.013
5		0.429	0.190	0.086	0.041	0.021
6		0.571	0.274	0.129	0.063	0.032
7			0.357	0.176	0.089	0.047
8			0.452	0.238	0.123	0.066
9			0.548	0.305	0.165	0.090
10				0.381	0.214	0.120
11				0.457	0.268	0.155
12				0.545	0.331	0.197
13					0.396	0.242
14					0.465	0.294
15					0.535	0.350
16						0.409
17						0.469
18						0.531

$n_2 = 7$

u	n_1						
	1	2	3	4	5	6	7
0	0.125	0.028	0.008	0.003	0.001	0.001	0.000
1	0.250	0.056	0.017	0.006	0.003	0.001	0.001
2	0.375	0.111	0.033	0.012	0.005	0.002	0.001
3	0.500	0.167	0.058	0.021	0.009	0.004	0.002
4	0.625	0.250	0.092	0.036	0.015	0.007	0.003
5		0.333	0.133	0.055	0.024	0.011	0.006
6		0.444	0.192	0.082	0.037	0.017	0.009
7		0.556	0.258	0.115	0.053	0.026	0.013
8			0.333	0.158	0.074	0.037	0.019
9			0.417	0.206	0.101	0.051	0.027
10			0.500	0.264	0.134	0.069	0.036
11			0.583	0.324	0.172	0.090	0.049
12				0.394	0.216	0.117	0.064
13				0.464	0.265	0.147	0.082
14				0.538	0.319	0.183	0.104
15					0.378	0.223	0.130
16					0.438	0.267	0.159
17					0.500	0.314	0.191
18					0.562	0.365	0.228
19						0.418	0.267
20						0.473	0.310
21						0.527	0.355
22							0.402
23							0.451
24							0.500
25							0.549

Mann-Whitney test (Example)

Algorithm A	180	195	232	185	200	186
Algorithm B	219	233	196	202	188	286

- Two algorithms are used to solve an maximization problem. We would like to check **whether algorithm B is superior to A** at $\alpha=0.05$;
NO evidence can support the solutions are normally distributed
- H_0 : Algorithms A and B are not different
- H_1 : Algorithms A and B perform differently

Mann-Whitney test (Example Solution)

Algorithm A	1	5	10	2	7	3
Algorithm B	9	11	6	8	4	12

- Rank the samples; W_1 and W_2 are 28 and 50

- $U_1 = 6*6 + (6*7)/2 - 28 = 29$ (5+6+4+5+3+6)

- $U_2 = 6*6 + (6*7)/2 - 50 = 7$ (0+1+4+0+2+0)

- $U = 7$

- Check the Table

$P(U \leq 7 \mid H_0 \text{ is true}) = 0.047 < 0.05$; reject H_0 (Page 38)

Algorithm B performs statistically better than A

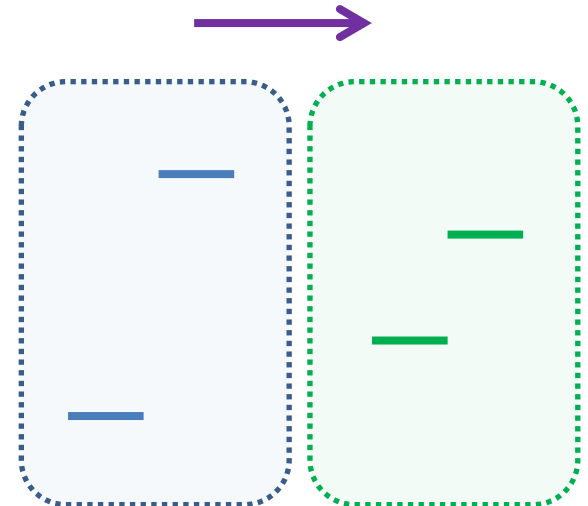
Concept of Mann-Whitney test

- Concept: If U is greater, U_1 and U_2 are closer, more likely to prove “difference is less significant”

$n_2 = 6$

u	n_1					
	1	2	3	4	5	6
0	0.143	0.036	0.012	0.005	0.002	0.001
1	0.286	0.071	0.024	0.010	0.004	0.002
2	0.428	0.143	0.048	0.019	0.009	0.004
3	0.571	0.214	0.083	0.033	0.015	0.008
4		0.321	0.131	0.057	0.026	0.013
5		0.429	0.190	0.086	0.041	0.021
6		0.571	0.274	0.129	0.063	0.032
7			0.357	0.176	0.089	0.047
8			0.452	0.238	0.123	0.066
9			0.548	0.305	0.165	0.090
10				0.381	0.214	0.120
11				0.457	0.268	0.155
12				0.545	0.331	0.197
13					0.396	0.242
14					0.465	0.294
15					0.535	0.350
16						0.409
17						0.469
18						0.531

$\alpha = 0.05$



Consistency

- We may wish to test the consistency of an algorithm against others
- **Variance among performance** samples is a good measurement for consistency

Test Variance

- If normality assumption is valid, we may use F-test (Page 21)

Optional (Pages 44-46)

- Mann-Whitney test can also be used to test whether two variables have the same variance
 - H_0 : Variables have the same variance
 - H_1 : Variances are significantly different
- Assumption:
mean values are close to each other

Mann-Whitney test procedure (Test Variance)

- Place all the values together in rank order (i.e. from lowest to highest).
- Start from the smallest value, rank it 1, jump to the largest value, rank it 2; go back to the second smallest, then to the second largest, and so on.
- Calculate U_1 and U_2
- $U = \min(U_1, U_2)$
- If $P(U \leq u \mid H_0 \text{ is true}) < \alpha$, reject H_0

Mann-Whitney test (Example)

Algorithm A	720	733	726	735	739
Algorithm B	723	727	731	729	

- Two algorithms are used to solve an maximization problem. We would like to check whether the solutions of both algorithms have the same variance at $\alpha=0.05$
 - H_0 : Solutions of algorithms A and B have the same variance
 - H_1 : Variances are significantly different

Mann-Whitney test (Example Solution)

Algorithm A	1	6	5	4	2
Algorithm B	3	7	8	9	

- $U_1 = 2+5+5+5 = 17$
- $U_2 = 0+1+1+1+0 = 3$
- $U = 3$
- Check the Table

$P(U \leq 3 \mid H_0 \text{ is true}) = 0.056 > 0.05$ (Page 37)

Do not reject H_0 ; Variances are NOT significantly different

We cannot conclude that B is more consistent

Conclusions

- Validation of a meta-heuristics algorithm is usually done by comparing it to others
- **Statistics** offer a systematic way to evaluate and compare the performance of metaheuristics
- Common comparison criteria include
 1. **Solution quality**
 2. **Computational efficiency**
 3. **Consistency**