

Terminology

Output process – a module inside the simulator that collects and outputs some statistic.

Replication (run) – one execution of a simulator.

Batch simulation – a replication in which the output process(es) produces data at several times (e.g. every 1000 seconds of simulated time).

R replications of a batch simulator yield R sequences of auto-correlated data items. These can be aggregated (i.e. averaged) within each replication or across replications.

The typical terms associated with performance are:

Mean

Median

Percentile = Quantile which equals $Pr(X \leq x) = \mathbf{p}$ for a given \mathbf{p} (x is the p^{th} quantile of X). This measure is of use when one wants to know, for example, what maximum delay will be experienced by, say, 95% of the customers. Note that the median is the 50^{th} percentile (or the 0.5^{th} quantile).

Example copied from Banks et al.

Three replications of a batch simulation were performed, giving these results:

Output time	Batch	Replication		
		1	2	3
1000	1	3.61	2.91	7.67
2000	2	3.21	9.00	19.53
3000	3	2.18	16.15	20.36
4000	4	6.92	24.53	8.11
5000	5	2.82	25.19	12.62
		$\bar{X}_1 = 3.75$	$\bar{X}_2 = 15.56$	$\bar{X}_3 = 13.66$

The data points were aggregated within each replication, yielding 3 points for output analysis.

Estimation

Simulation produces a set of output data x_1, x_2, \dots, x_n .

The objective is to determine (as accurately as possible) what is the mean θ of the underlying random variable.

The obvious **point estimate** of θ is:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$$

If the expectation $E(\hat{\theta}) = \theta$, the estimator is **unbiased**; otherwise the value $E(\hat{\theta}) - \theta$ is called the **bias** of the estimator.

Note that the above formula assumes that all points are equally meaningful. If (for unclear reasons) the measurements are not taken at equal intervals, the estimation becomes:

$$\hat{\theta} = \frac{1}{T_{total}} \sum_{i=1}^n x_i \times t_i$$

where t_i is the (simulated) time that elapsed between taking measurement $i - 1$ and measurement i .

Confidence interval

Even if the estimator is unbiased, $\hat{\theta}$ seldom equals θ , and it is natural to claim that θ lies in some interval with some confidence level:

$$\hat{\theta} - \delta \leq \theta \leq \hat{\theta} + \delta$$

with a confidence level of p (obviously, δ is a function of p).

It is traditionally assumed that the data points produced by simulation are **normally distributed** variates. This assumption allows to express δ as a function of p and the sample variance.

If we have n data points $x_1 \dots x_n$ (typically from n replications), the sample mean is $\hat{\theta}$ and the variance of the sample is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\theta})^2$$

A sample of n elements has $n - 1$ degrees of freedom which can be explained ([very naively](#)) as saying that if you give me $\hat{\theta}$ and $n - 1$ points, I will be able to calculate the n^{th} point.

If the sample is normally distributed around $\hat{\theta}$, I can make a claim about the interval in which θ lies: the value of δ is:

$$t_{\alpha/2, n-1} \times \frac{S}{\sqrt{n}} = t_{\alpha/2, n-1} \times \sqrt{\frac{S^2}{n}}$$

where $t_{\alpha/2, n-1}$ is an entry in the table containing values for the **t-distribution**.

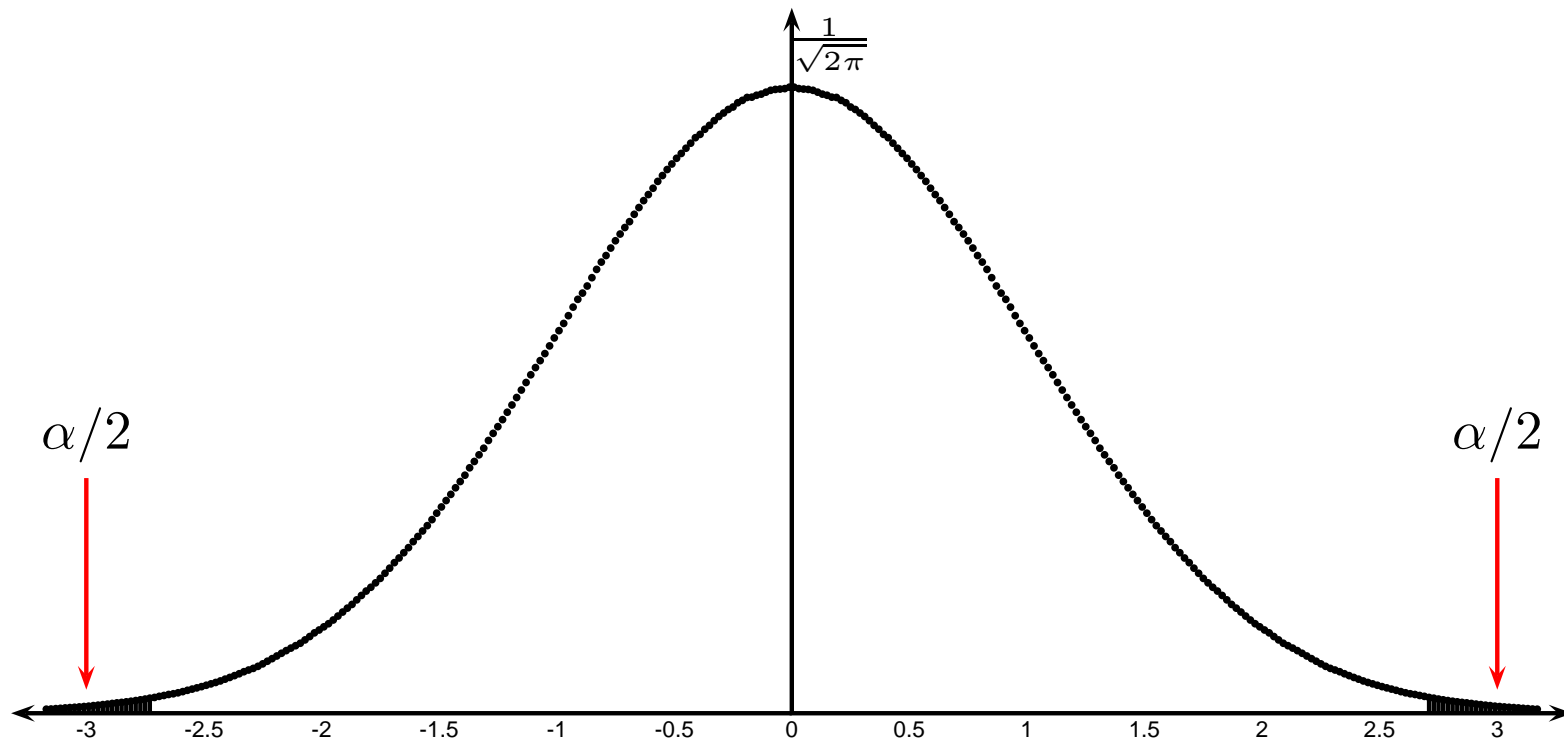
This claim can be made with a significance level of α (significance + confidence being equal to 1, I am making the claim **with confidence level $1 - \alpha$**).

t-distribution

It is an approximation of the normal distribution which takes into account the finite number of degrees of freedom (independent data points).

Its pdf approaches the pdf of the normal distribution as n increases and becomes equal to it when $n = \infty$.

The significance level α is expressed by cutting off the two tails of the distribution, such that their total probability equals $\alpha/2$ at each end.

t-Distribution

Example

There are $n = 120$ data points each representing the daily average time needed to manufacture a part. They have an average of $\hat{\theta} = 5.8$ hours and sample standard deviation $S = 1.6$ hours. The table gives $t_{0.025,119} = 1.98$ ($0.05/2 = 0.025$, i.e. a confidence level of 95% and $119 = 120 - 1$).

The confidence interval is:

$$5.8 \pm 1.98 \times \frac{1.6}{\sqrt{120}} = 5.8 \pm 0.29$$

which says that the overall average manufacturing time is somewhere between 5.51 and 6.09 hours.

Note that the above statement can be made with a confidence level of 95%, i.e. it will be false—on average—once every 20 times it is uttered.

Confidence intervals

The formula

$$\hat{\theta} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$$

gives the confidence interval for the range of the **overall mean**.

A similar formula gives the range of unit averages (i.e. the range in which lie actual values that each replication simulated):

$$\hat{\theta}_R \pm t_{\alpha/2, n-1} S \sqrt{\frac{n+1}{n}}$$

Note that

$$\lim_{n \rightarrow \infty} t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} = 0$$

$$\lim_{n \rightarrow \infty} t_{\alpha/2, n-1} S \sqrt{\frac{n+1}{n}} = \zeta_{\alpha/2} \times \sigma$$

where $\zeta_{\alpha/2}$ is a positive value derived from α using the Normal distribution.

Example continued

There are $n = 120$ data points each representing the daily average time needed to manufacture a part. They have an average of $\hat{\theta} = 5.8$ hours and sample standard deviation $S = 1.6$ hours. The table gives $t_{0.025,119} = 1.98$.

The range of the individual data points is given by the formula (for the confidence interval 95%):

$$5.8 \pm 1.98 \times 1.6 \times \sqrt{\frac{121}{120}} = 5.8 \pm 3.18$$

which means that a floor manager that sees a time to manufacture outside the range of $[2.62, 8.98]$ has a problem.

Continued

If a manufacturing time falls outside the interval $[2.62, 8.98]$, the floor manager can use one of two explanations:

1. A confidence level of 95% implies that we should expect a value outside the range 5% of the time.
2. The data points do not necessarily follow a normal distribution which makes the whole derivation invalid.

Confidence-related problems

The possibilities:

- Given a confidence level and a sample size, find the confidence interval (already covered).
- Given an interval range (“precision”) and a sample size, find the highest confidence level.
- Given a confidence level, find the minimum sample size to satisfy it for a given precision.
- Suppose we have a S and n ; they give a family of confidence intervals (one for each confidence level). Find the highest confidence level such that a given value is outside the corresponding interval.

Sample size and precision given

We need to know the maximum confidence level to satisfy a given precision requirement.

Suppose we have n and S and want to know the lowest value of α that makes δ less than the given precision Δ .

The answer is:

$$\{\min_{\alpha} : t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \leq \Delta\}$$

or

$$\{\min_{\alpha} : t_{\alpha/2, n-1} \leq \frac{\Delta\sqrt{n}}{S}\}$$

Confidence level and precision given

We need to know the minimum sample size to satisfy a given precision requirement.

We are given α and an upper bound on the acceptable length of the confidence interval Δ .

If somehow we knew S , the answer would be simple:

$$\left\{ \min_i : \Delta \geq t_{\alpha/2, i-1} \frac{S}{\sqrt{i}} \right\}$$

The problem is that S is derived from the sample itself, so a more exact formula would be:

$$\left\{ \min_i : \Delta \geq t_{\alpha/2, i-1} \frac{S_i}{\sqrt{i}} \right\}$$

here S_i is the derived from the sample of size i we currently have.

Value outside the interval

We need to know the maximum confidence level with which we can claim that a given value is outside a corresponding confidence interval.

Suppose we have $\hat{\theta}$, n , S and some value \mathcal{V} . The maximum confidence is the minimum significance, hence the question is:

$$\left\{ \min_{\alpha} : |\mathcal{V} - \hat{\theta}| \geq t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \right\}$$

Note that as α decreases, $t_{\alpha/2}$ gets larger.

Example

A simulation of the operation of a bank consisted of $n = 10$ replications. We are interested in a number of statistics:

\mathcal{X}_1 The average waiting time (delay) of a customer.

\mathcal{X}_2 The proportion of customers that had to wait for less than 5 minutes.

Waiting time

The sample yielded the values: $\widehat{X}_1 = 2.03$, $S^2 = 0.31$ (in minutes and minutes²).

For a confidence level of **90%** ($\alpha = 0.1$) the confidence interval is:

$$\widehat{X}_1 \pm t_{0.05,9} \times \sqrt{\frac{S^2}{10}}$$

or

$$\widehat{X}_1 \pm 1.83 \times \sqrt{0.031} = \mathbf{2.03 \pm 0.32}$$

For a confidence level of **99%** it would be

$$\widehat{X}_1 \pm 3.25 \times \sqrt{0.031} = \mathbf{2.03 \pm 0.57}$$

The lengths of the confidence intervals differ so little because the variance is very small.

Waiting less than 5 minutes

The sample yielded the values: $\widehat{X}_2 = 0.853$, $S^2 = 0.004$

If we choose a 90% confidence level ($\alpha = 0.1$), we get:

$$\delta = 1.83 \times \sqrt{\frac{0.004}{10}} = 0.036$$

so that the confidence interval for customers waiting no longer than 5 minutes is 0.853 ± 0.036 with a confidence level of 90% (it would also be 0.853 ± 0.044 with confidence 95%).

If we want to say that “no more than 10% of customers waited more than 5 minutes” we would find that we can say so with a confidence slightly above 95% (but below 96%).

Minimum number of replications

The boss wants an estimate average delay with a confidence level of 90% and a confidence interval of ± 0.25 minutes.

We have 10 replications so far, so the formula is:

$$n_{min} = \left\{ \min_i : t_{\alpha/2, i-1} \frac{S_{10}}{\sqrt{i}} \leq 0.25 \right\}$$

The subscript in S_{10}^2 indicates that different results would be obtained if a different number of replications were available.

With the given 10 samples, the estimate for the outcome is 16. If we had 15 samples (and S_{15} we might get the minimum number to be 30 (or 15).

Thus, this approach is very misleading if n is small (the quality of the estimators $\hat{\theta}$ and S being low).

Another way to deal with precision

This approach requires producing new replications until a satisfactory result is achieved.

Given are α and δ . We need to find the minimum number of replications n such that:

$$\hat{\theta} - \delta \leq \hat{\theta} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \leq \theta \leq \hat{\theta} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \leq \hat{\theta} + \delta$$

Define two functions of i : $\hat{\theta}_i$ and

$$\delta_i = t_{\alpha/2, i-1} \frac{S_i}{\sqrt{i}}$$

```
 $i = n_0 ;$   
generate  $i$  replications.  
compute  $\hat{\theta}_i$  and  $\delta_i$ .  
while(  $\delta_i > \delta$  ) {  
     $i++ ;$   
    generate another replication.  
    compute  $\hat{\theta}_i$  and  $\delta_i$ .  
}
```

When the algorithm stops, i is the minimum number of replications such that $\hat{\theta}_i$ and δ_i are acceptable estimators for θ and a confidence interval not larger than $\pm\delta$.

Relative error

α is given. Also given is a relative error requirement of ϵ' .
The formula for the minimum number of replications to satisfy the relative error requirement is:

$$n_{min} = \left\{ \min_i : \frac{t_{\alpha/2, i-1} \sqrt{\frac{S^2}{i}}}{\hat{\theta}} \leq \epsilon' \right\}$$

In bank operation example, we had $\hat{X} = 2.03$ and $S^2 = 0.31$ for 10 replications.

To get a minimum relative error $\epsilon' = 0.09$ with significance $\alpha = 0.1$, the minimum number of replications required is:

$$n_{min} = \left\{ \min_i : \frac{t_{0.05, i-1} \sqrt{\frac{0.31}{i}}}{2.03} \leq 0.09 \right\} = 27$$

The correct answer is 74 (27 was obtained incorrectly because n is too small compared to the correct result).

How many queues?

Suppose that there are 5 tellers and that the overall server (“teller”) utilisation $\rho = 0.8$. Is it better to have a single queue or five independent queues?

A simplistic analysis shows that there is no difference, since the average waiting times are the same. A more careful analysis shows that the five–queue system generates more unhappy customers:

Interval	5 queues	1 queue
0–10	80.8%	78.5%
10–20	12.3%	20.2%
20–30	5.1%	1.3%
30+	1.8%	0%

More obscure parameters

\mathcal{X} is a random variable and x_1, \dots, x_n form a sample drawn from it. We need to know what is the probability $p = Pr(\mathcal{X} \in B)$ where B is a set of numbers.

Let s be the size of the subset of the sample that falls in the set B . If n is sufficiently large, s has a binomial distribution with parameters n and p and a point estimator for p is:

$$\hat{p} = \frac{s}{n}$$

Likewise, we may want to determine the size of s for a given value of \hat{p} .

Example

The operation of a bank was simulated in 100 independent replications. The objective was to find how often the maximum (daily) queues size does not exceed 15. It turned out that $s = 77$, hence $\hat{p} = 0.77$.

Conversely, the data points were used to find out what the 0.95^{th} quantile is and $s_{0.95} = 20$. Likewise, $s_{0.99} = 23$.

This information may be useful when designing the size of a waiting area (lobby).

Bonferroni's inequality

If one conducts a simulation that produces k output variables μ_1, \dots, μ_k simultaneously and each of these variables has a confidence interval I_i with a confidence level of $1 - \alpha_i$, then the probability of all these these intervals being correct is equal or greater than:

$$1 - \sum_{i=1}^k \alpha_i$$

This inequality holds regardless of the correlation level among the individual measures.

Example

A simulation experiment determined that $v_1 \in [0.5, 0.7]$, $v_2 \in [2.7, 3.2]$ and $v_3 \in [17.7, 18.9]$ with confidence of 90%.

The confidence that all three claims are true is only 70% which is not much more than the “lack of confidence” percentage of 50%.

Alternative design configurations

This is the most useful part of simulation.

Consider the typical problem of one fast server vs. two slow ones. The most challenging case is: server \mathcal{A} costs twice as much as server \mathcal{B} and is exactly twice as fast (these could be ATM machines or soda dispensers).

Common sense hints that this is a non-issue: two \mathcal{B} have the same cost as one \mathcal{A} and deliver exactly the same service. Hence the average manager will choose \mathcal{A} if he gets a higher bribe from its manufacturer than from the maker of \mathcal{B} .

Consider the two scenarios:

- One \mathcal{A} machine with a utilisation factor of 0.9.
- Two \mathcal{B} machines with the utilisation of 0.9 each (but twice as slow). To make the field level, the customers form a single queue to the two machines.

The measure of performance is the waiting delay (before being served) $d_{\mathcal{A}}$ and $d_{\mathcal{B}}$; it is generally assumed that people are far more aggravated waiting than being taken care of.

A simulation with 100 replications of 100 customers yielded the results: $d_{\mathcal{A}} = 4.13$ and $d_{\mathcal{B}} = 3.70$ (a paper by Kelton&Law).

However, in only 48% of the 100 replications $d_{\mathcal{A}} > d_{\mathcal{B}}$, so that in 52% of the replications machine \mathcal{A} served faster.

The natural thing is to opt for the scenario that has lesser delays most of the time (i.e. \mathcal{A}) and to consider the higher mean to result from a black swan.

As usual, the answer is to run more tests.

Suppose we perform a more thorough test. If we perform 5, 10, 20 sets of 100–replications each we get the following proportion of experiments favouring the \mathcal{A} machine:

1	0.52
5	0.43
10	0.38
20	0.34

Clearly, machine \mathcal{B} is better.

Comparison of two systems

We have two alternative designs and want to compare their relative performance.

There are two methods of doing so.

- Calculate a t confidence interval for the difference of two samples.
- Compute a two-sample t confidence (Welch) interval.

Both methods have limitations which should be carefully checked before applying.

Difference of samples

- Two systems are simulated resulting in two samples of equal size n : X_{1i} and X_{2i} for $i = 1, \dots, n$. Note that each data point is the result of a whole replication.
- An artificial sample $Z_i = X_{1i} - X_{2i}$ is constructed. We want to find $\hat{\zeta} = \widehat{E}(Z)$ and a confidence interval for $\hat{\zeta}$.

$$\hat{\zeta} = \sum_{i=1}^n \frac{Z_i}{n}$$

$$S^2 = \sum_{i=1}^n \frac{(Z_i - \hat{\zeta})^2}{n-1}$$

- The confidence interval is

$$\hat{\zeta} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$$

This approach works for correlated data.

Example

The costs of two systems are being compared:

i	X_{1i}	X_{2i}	Z_i
1	126.97	118.21	8.76
2	124.31	120.22	4.09
3	126.68	122.45	4.23
4	122.66	122.68	- 0.02
5	127.23	119.40	7.83

The interval is $4.98 \pm t_{\alpha/2,4} \times 1.56$.

At a significance level of 0.1 (90%) the resulting interval is $[1.65, 8.31]$ and a claim can be made that system 2 is cheaper (with this confidence level).

Note that at significance 0.02 (98%) the interval would be $[-0.87, 10.83]$ and no claim could be made. Of course, $n = 5$ is not a suitable choice for the size of a sample.

Welch confidence interval

- Two systems are simulated resulting in two samples of sizes n_1 and n_2 : X_{1i} and X_{2i} for $i = 1, \dots, n$. Note that each data point is the result of a whole replication.
- The means and variances are computed for each sample:

$$\hat{\zeta}_i = \sum_{j=1}^{n_i} \frac{X_{ij}}{n_i}$$

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \hat{\zeta}_i)^2$$

- The Welch confidence interval is:

$$\zeta_1 - \zeta_2 \pm t_{\alpha/2, df} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

The degree of freedom df is a bit complicated. Denoting

$$W_i = \frac{S_i^2}{n_i}$$

$$df = \frac{(W_1 + W_2)^2}{\frac{W_1^2}{n_1 - 1} + \frac{W_2^2}{n_2 - 1}}$$

This obscure formula becomes

$$(n-1) \times \frac{(W_1 + W_2)^2}{W_1^2 + W_2^2} = n-1 + \frac{(n-1)W_1W_2}{W_1^2 + W_2^2} > n-1$$

when $n_1 = n_2 = n$.

When $n_1 \neq n_2$ the formula for df usually yields a non-integer value, requiring interpolation in the printed tables (linear interpolation will do).

Example

(Same example as above.)

$$n_1 = n_2 = 5, \zeta_1 = 125.57, \zeta_2 = 120.59, S_1^2 = 4.00, S_2^2 = 3.76 \text{ and } df = 7.99.$$

The Welch confidence interval is $[2.66, 7.30]$ at a confidence of 90% and $[0.78, 9.17]$ at a confidence level of 99% allowing to claim that system 2 is cheaper with 99% confidence.

But this claim is meaningful only if the two samples were independent (i.e. different pseudo-random streams were used).

More than two systems

When more than two systems are being compared, two approaches are possible:

- Choose one system as “standard” and compare all the other systems with it. Most commonly, the “standard” system is an existing one the performance of which is well known.
- Perform all–pairwise comparisons. This approach is practical if the number of systems is small (for 20 systems we need 190 pairwise comparisons).

In any case, one must remember the Bonferroni inequality about the overall confidence level of a set of simultaneously computed intervals.

$$P(\forall_i \mu_i \in I_i) \geq 1 - \sum_i \alpha_i$$

Comparing with a standard

Consider an existing inventory system X_1 and 4 alternative policies X_2, \dots, X_5 . We consider X_1 to be the “standard” and want to know which policies are better than the standard with an overall significance of 0.1 (confidence of 90%).

Since there are 4 confidence intervals to determine, they should be computed with a significance of 0.025

(Bonferroni); 5 replications were made for each X_2, \dots, X_5 .

i	$\mu_i - \mu_1$	Difference Interval	Welch Interval
2	-4.98	[-10.44,0.48]	[-8.52,-1.44]
3	-1.23	[-8.80,6.34]	[-7.44,4.97]
4	6.36	[0.27,12.46]	[1.82,10.91]
5	17.15	[13.20,81]	[14.07,20.22]

Analysis

i	$\mu_i - \mu_1$	Difference Interval	Welch Interval
2	-4.98	[-10.44,0.48]	[-8.52,-1.44]
3	-1.23	[-8.80,6.34]	[-7.44,4.97]
4	6.36	[0.27,12.46]	[1.82,10.91]
5	17.15	[13.20,81]	[14.07,20.22]

- In 3 of the 8 cases it is not clear whether there is a significant difference from the standard.
- Since only 5 replications were made, the length of the intervals can be reduced by making more replications. As a rule of thumb, one can reduce the interval width by half by performing 4 times as many replications. Note that in the case of X_3 it would take about 300 replications to determine whether it differs from X_1 .

All pairwise comparisons

We simulate k systems X_1, \dots, X_k and obtain for each of them \bar{X}_i and S_i^2 which are estimators of their mean μ_i and variance σ_i^2

To compare these systems we must create $\frac{k(k-1)}{2}$ confidence intervals for $\mu_j - \mu_i$ for all i and j such that $i > j$. The significance level for each interval must be $\frac{2\alpha}{k(k-1)}$ so that for $k = 5$ we need to build each interval with a confidence level of 99% in order to get an overall confidence of 90% (in the “comparison with a standard” we would need only 97.5%).

As usual, the abundance of experimental data makes definite statements difficult and the all-pairwise comparisons are not very useful unless there is a clear “best” system.

Example

Same example as before. The confidence levels are 99%.

Sample difference				
i	j			
	2	3	4	5
1	-4.98 ± 7.18	-1.23 ± 9.99	6.36 ± 8.01	17.15 ± 4.83
2		3.75 ± 9.58	11.34 ± 8.38	22.12 ± 3.80
3			7.60 ± 5.66	18.38 ± 7.73
4				10.78 ± 5.85

Welch

i	j			
	2	3	4	5
1	-4.98 ± 4.36	-1.23 ± 7.91	6.36 ± 5.60	17.15 ± 3.80
2		3.75 ± 7.86	11.34 ± 5.88	22.12 ± 3.72
3			7.60 ± 7.67	18.38 ± 8.51
4				10.78 ± 5.89

Analysis

- If the objective is to maximise the value of the variable, X_5 is a clear winner.
- If the objective is to minimise the value of the variable, X_5 is a clear loser, but it is not possible to tell the best system.

One can build a table of winners, ties and losers:

Difference				
	1	2	3	4
1		=	=	=
2	=		=	2
3	=	=		3
4	=	2	3	

Welch				
	1	2	3	4
1		2	=	1
2	2		=	2
3	=	=		=
4	1	2	=	

Even if the confidence interval width is reduced by half, this comparison will still give the conclusion that $\mu_1 = \mu_2 = \mu_3$ (“difference” method). The Welch method gives the contradictory conclusion that $\mu_1 = \mu_3, \mu_2 = \mu_3$ which would falsely imply $\mu_1 = \mu_2$.

Note that the t-test does not **prove** anything, because nothing is certain in stochastic processes; it only gives a bound on the degree of uncertainty.

To be able to claim that system 3 is worse than system 2, we would have to lower the confidence level to 70% (add to it Bonferroni’s inequality).

More on comparing systems

The main purpose of comparing alternative systems is to rank them based on the expected mean of some critical variable (cost, profit, time, etc.).

Considering the inherent uncertainty of the t-test (and all other statistical tests), it is practical to replace the traditional $\mu_1 = \mu_2$ with:

$$\mu_1 = \mu_2 \text{ if and only if } |\mu_1 - \mu_2| \leq \delta^*$$

for some small value δ^* .

This reflects the fact that in the presence of uncertainty it is naive to insist, say, that $1234.5 > 1232.1$.

Hence, much work went into methods of selecting a subset of the systems that:

- Contains the best system (but does not identify it).
- Contains no system that is rejected as best with a given confidence level.
- Attempts to avoid all-pairwise comparisons by organising some form of tournament-style elimination.