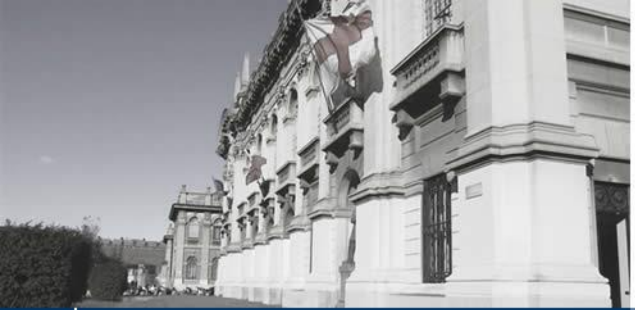
Performance Evaluation and Applications



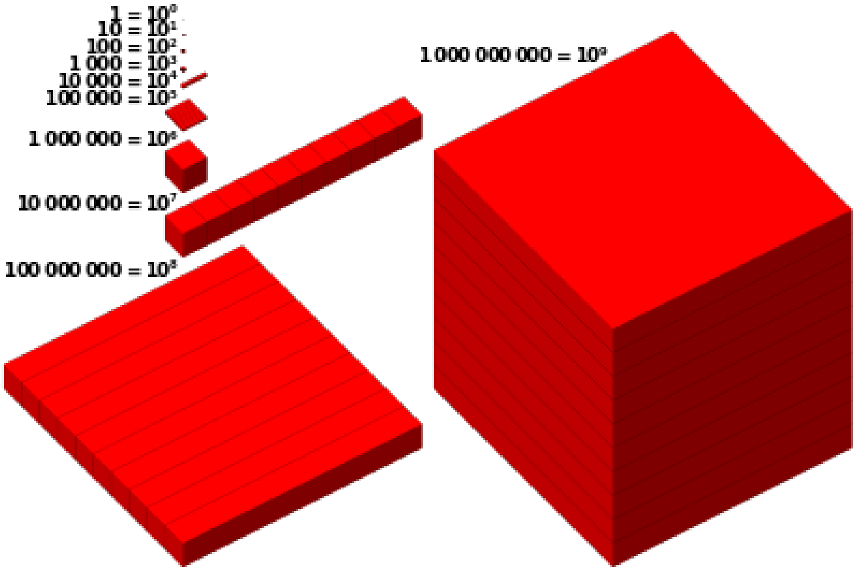
POLITECNICO DI MILANO

Confidence Intervals

Motivating example

We have seen that random variables and random numbers generation are tools to create realistic synthetic traces with which we can study a system.

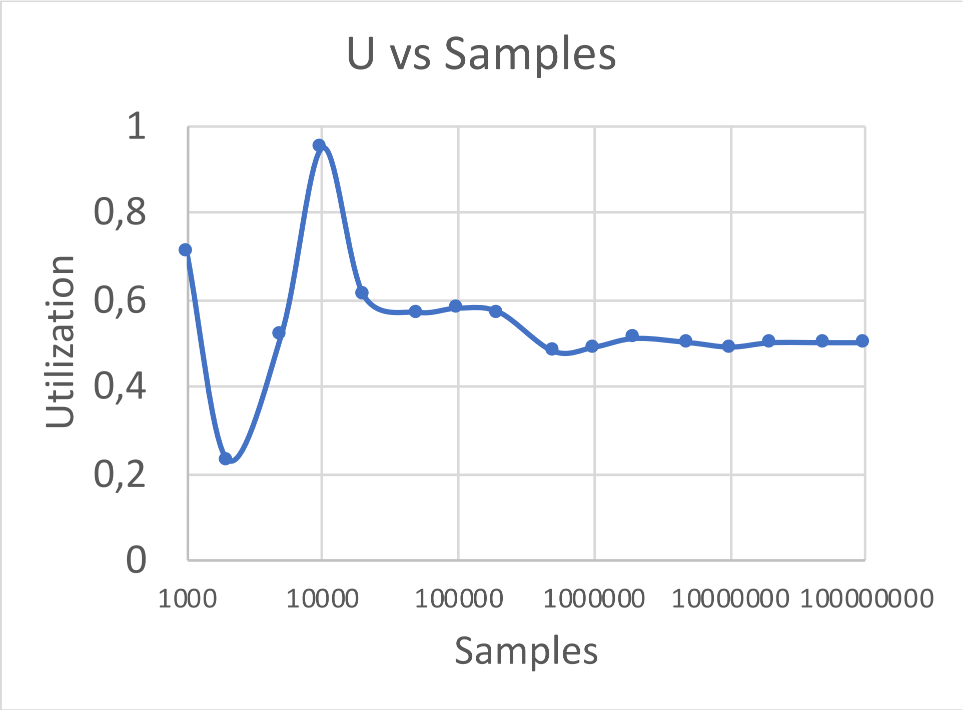
How big should these traces be? How many samples should I choose to have an accurate solution without wasting too much time and computational resources?



?

Confidence intervals

Whenever we are estimating the performance metrics of a system from a set samples, the size of the data set matters.



Confidence intervals

To properly consider the effect of the number of samples that are used to compute the performance indices, a *confidence interval* is generally determined instead of a single value.

The user specifies a *confidence level 0* < γ < *1*, and the measures is returned as an interval *(l*γ*, u*γ*)*. The actual performance index *a* falls in such interval with probability γ :

$$P\left(l_\gamma < \alpha < u_\gamma\right) = \gamma$$

Confidence intervals

The size of the interval *(l_γ, u_γ)* depends on both the confidence level γ and the number of samples *N*:

- It becomes lager with the increase of the confidence level γ
- It reduces with the population size *N*

Note that the definition still allows for the technique to compute a wrong performance indices estimate, i.e. a confidence interval *[l_γ, u_γ]* that does not include the actual value α, with probability 1- γ .

How a confidence interval is computed, depends on the requested index and on its statistical properties.

$$P\left(l_\gamma < \alpha < u_\gamma\right) = \gamma$$

Confidence intervals: average

Let us focus on the case of computing a measure expressed as the average of a set of *N* samples x_i : this can be applied, for instance, to the average *response time*, to the *service time* or to the *inter-arrival time*. As we have seen, the average can be computed as:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Example: compute the average service time S, from a set of *N* samples s_i.

$$E[S] = \bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i$$

Each sample of the measure can be considered as an instance of a random variable X. Moreover, $\bar{x}$ can be considered a random variable itself, whose distribution can be expressed as the sum of *N* independent identical distributed instances of *X*, multiplied by constant *1/N*.

$$\bar{X} \sim \frac{1}{N}\sum_{i=1}^{N} X$$

Here, $\bar{X}$ is a random variable, with a given distribution. To find an analogy, as an Erlang is the sum of *N* independent Exponential random variables, here $\bar{X}$ is the sum *N* random variables of type X.

Confidence intervals: average

Mathematicians have shown that, for a large amount of distributions X, if we call:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X \qquad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X-\bar{X})^2$$

And if we call μ the real average of the variable under study, i.e. $\mu = E[X]$, then the distribution of the following quantity follows a special distribution $T_{N\text{-}1}$ that depends only on *N*. This is called the *Student T distribution with* N *degrees of freedom*:

$$\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{N}}} = T_{N-1}$$

Confidence intervals: average

A set of samples x_i, would determine one value for both $\bar{X}$ and S^2, and it would be an instance t of the T_{N-1} distribution.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

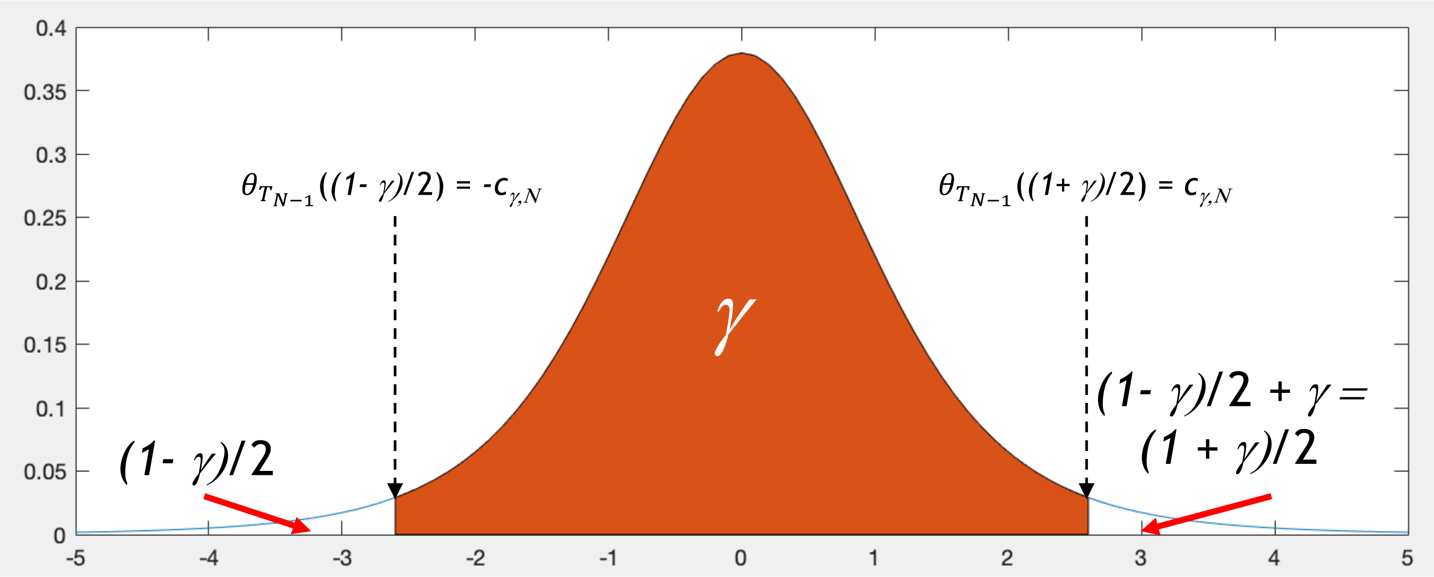
$$\boxed{t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}}$$

Please note that here we use small letters instead of capital letters to denote instances and not distributions.

Confidence intervals: average

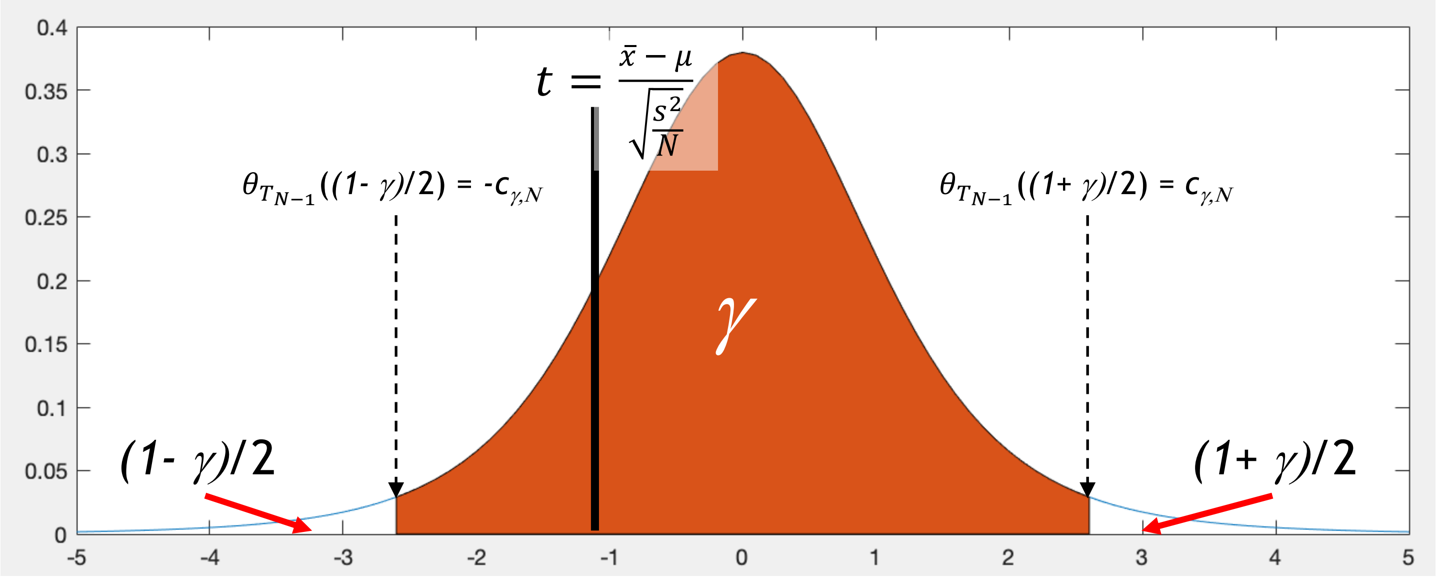
The shape of the T_{N-1} distribution is symmetrical along the origin. Let us remember that the area below a distribution is unitary, and let us search a fraction $\gamma < 1$ of this area. Due to the symmetry of the distribution, and the definition of the percentiles, we have:

$$\theta_{T_{N-1}}((1-\gamma)/2) = -\theta_{T_{N-1}}((1+\gamma)/2)$$



Confidence intervals: average

In particular, an instance *t* of T_{N-1} is inside its *(1- γ)*/2 and *(1 + γ)*/2 percentile with probability *γ*.



Let us call the percentiles $c_{\gamma,N} = -\theta_{T_{N-1}}((1-\gamma)/2) = \theta_{T_{N-1}}((1+\gamma)/2)$.
Then, by definition:

$$P\left(-c_{\gamma,N} \le t \le c_{\gamma,N}\right) = \gamma$$

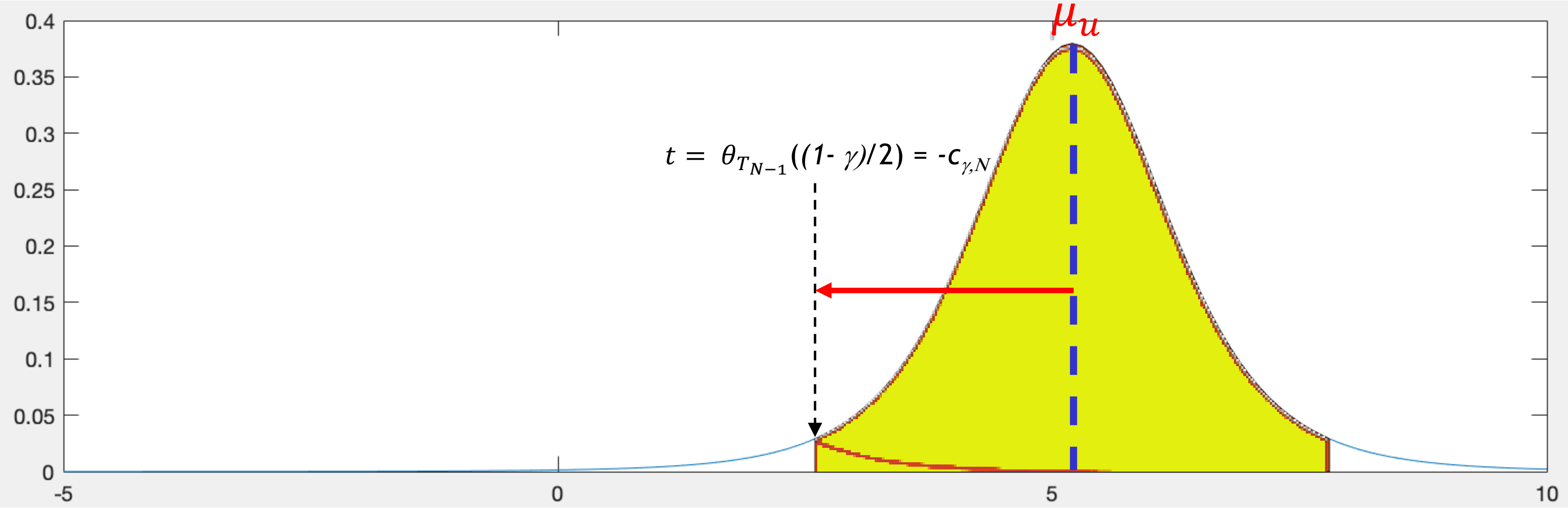
$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}}$$

$$P\left(-c_{\gamma,N} \le \frac{(\bar{x}-\mu)}{\sqrt{\frac{s^2}{N}}} \le c_{\gamma,N}\right) = \gamma$$

Confidence intervals: average

In the worst possible case when $\mu = \mu_u$ is high, then $t = -c_{\gamma,N}$.

$$-c_{\gamma,N} = \frac{\bar{x} - \mu_u}{\sqrt{\frac{s^2}{N}}} \qquad \mu_u = \bar{x} + c_{\gamma,N}\sqrt{\frac{s^2}{N}}$$

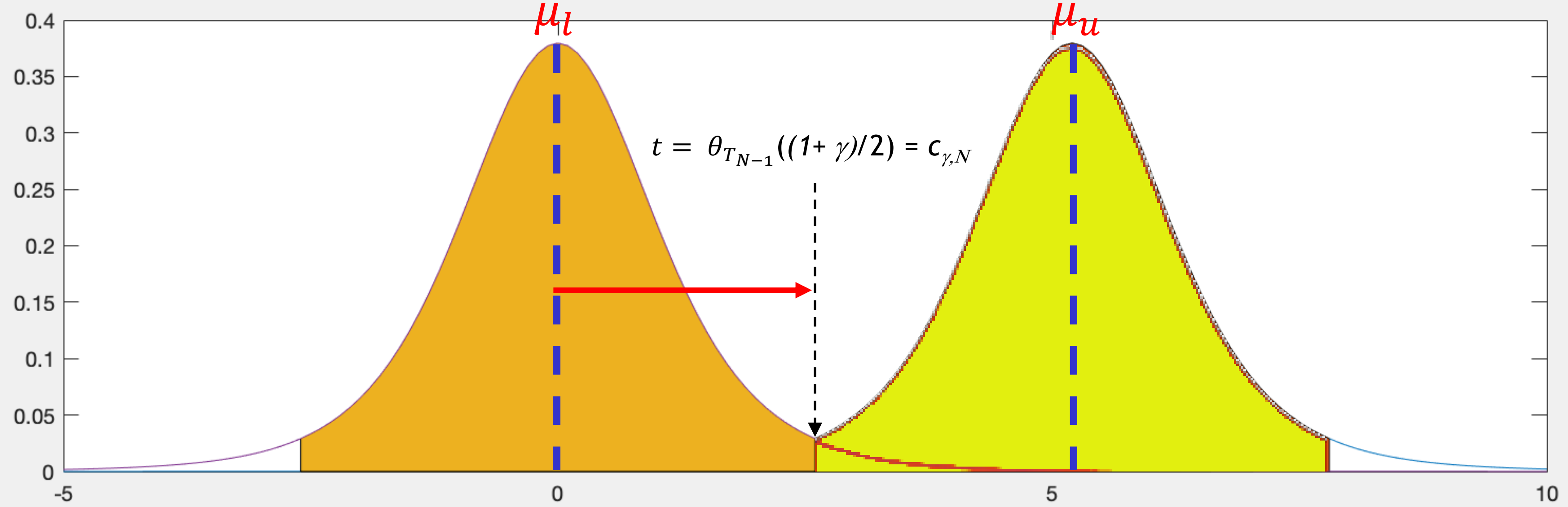


$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = t \qquad P\left(-c_{\gamma,N} \leq \frac{(\bar{x} - \mu)}{\sqrt{\frac{s^2}{N}}} \leq c_{\gamma,N}\right) = \gamma$$

Confidence intervals: average

Similarly, in the worst possible case when $\mu = \mu_l$ is low, then $t = c_{\gamma,N}$.

$$c_{\gamma,N} = \frac{\bar{x} - \mu_l}{\sqrt{\frac{s^2}{N}}} \qquad \mu_l = \bar{x} - c_{\gamma,N}\sqrt{\frac{s^2}{N}}$$

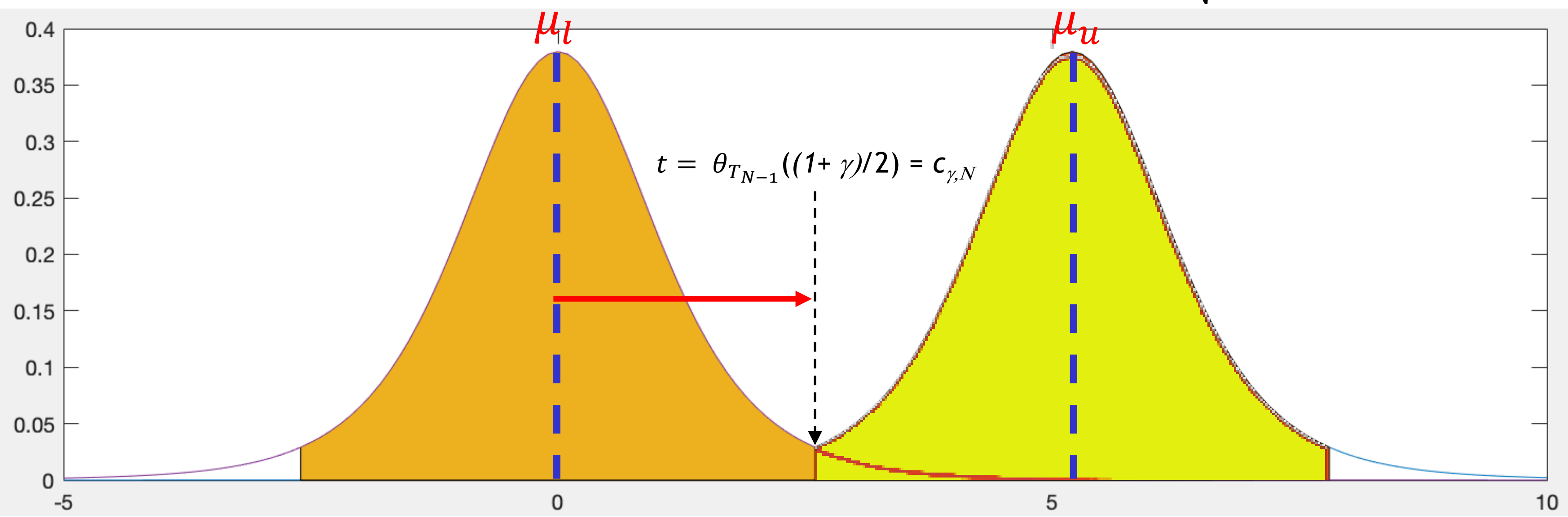


$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = t_{N-1} \qquad P\left(-c_{\gamma,N} \leq \frac{(\bar{x} - \mu)}{\sqrt{\frac{s^2}{N}}} \leq c_{\gamma,N}\right) = \gamma$$

Confidence intervals: average

Please don't be misled by the picture: although it would seem that the interval could be pretty large, it is inversely proportional to the square root of *N*. With a moderate number of samples *N* and a reasonable variance s^2, the interval becomes very tight.

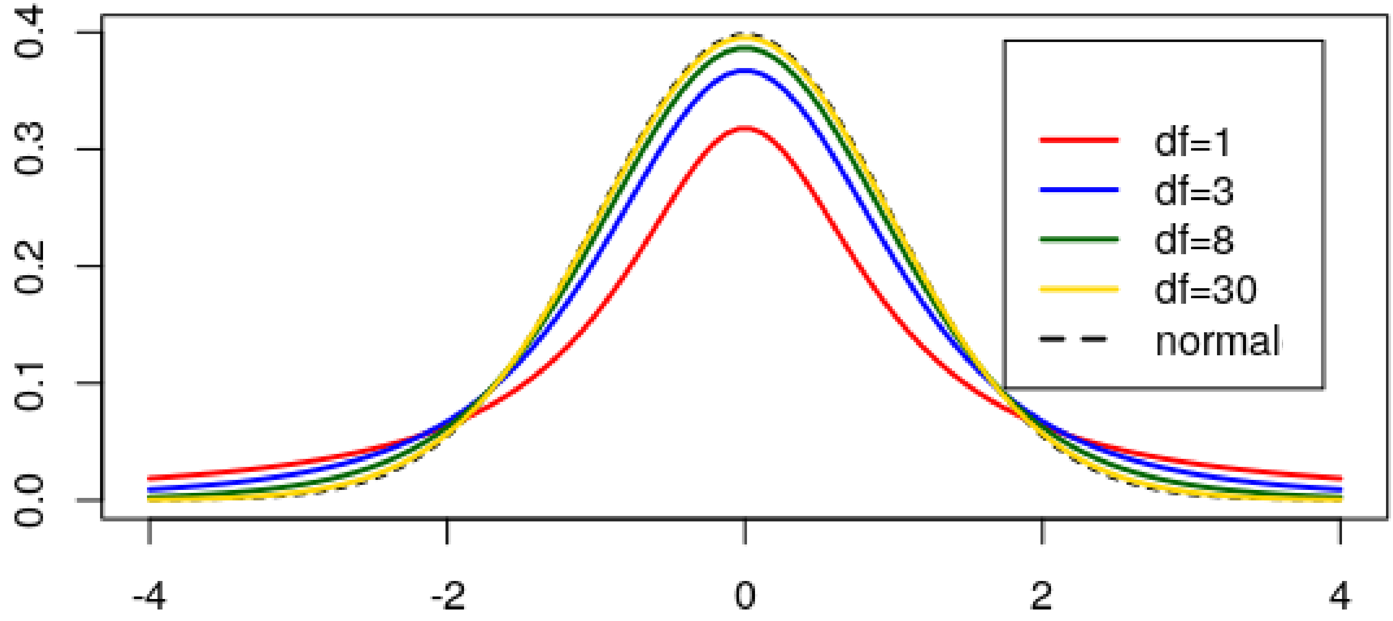
$$\mu_l = \bar{x} - c_{\gamma,N}\sqrt{\frac{s^2}{N}} \qquad \mu_u = \bar{x} + c_{\gamma,N}\sqrt{\frac{s^2}{N}}$$



Confidence intervals: average

As the number of degree of freedom N increases, the *Student T* distribution tends to the *standard normal distribution*.

For N > 30, they are basically identical.



Confidence intervals: average

To summarize, if we have few samples (*N* < *30*), and we call $c_{\gamma,N}$ = -$\theta_{T_{N-1}}$((1- γ)/2), the percentile of the Student T distribution with N-1 degree of freedom, we can express the γ confidence interval as:

$$\left[\bar{x} - c_{\gamma,N}\sqrt{\frac{s^2}{N}}, \bar{x} + c_{\gamma,N}\sqrt{\frac{s^2}{N}}\right]$$

If we call d_γ= -$\theta_{N<0,1>}$((1- γ)/2, the percentile of a standard Normal distribution, for a large number of samples (N >= 30), we have can compute the γ confidence interval as :

$$\left[\bar{x} - d_{\gamma}\sqrt{\frac{s^2}{N}}, \bar{x} + d_{\gamma}\sqrt{\frac{s^2}{N}}\right]$$

Confidence intervals for the average: discussion

The previous formulas are mathematically correct only when the considered samples follow a Normal distribution. For the *Central Limit Theorem*, however, they become a good approximation for large values of N for almost any distribution.

Student T values are meaningful in statistics, where the assumption of a Normal distributed population is quite common, samples have a small c.v. and it might be difficult to acquire a large number of measurements.

In *Performance Evaluation* it is quite common to have a large number of samples (i.e. N ~ [10^6 … 10^9]): for this reason the Normal distribution approximation is generally valid.

Confidence intervals for the average: discussion

Computation of the confidence interval uses a confidence level γ dependent constant which multiplies the standard deviation of the samples *s*, and that is divided by the square root of the number samples.

A table of the most commonly used values of constant d_γ is the following:

γ	d_γ
99%	2.576
98%	2.326
95%	1.96
90%	1.645

$$\left[\bar{x} - d_\gamma\sqrt{\frac{s^2}{N}}, \bar{x} + d_\gamma\sqrt{\frac{s^2}{N}}\right]$$

Confidence intervals for other measures

The technique we have just shown, can be directly applied only to very few measures:

- *Utilization*
- *Response time* → $R = \frac{W}{C} = \frac{\sum_{i=1}^{N} r_i}{N}$

  (with caution: later we will see why!)
- *Average queue length*
- *Throughput*

- *Arrival rate*
- *Average Service time* → $S = \frac{B}{C} = \frac{\sum_{i=1}^{N} s_i}{N}$

- *Average Inter-arrival time* → $\frac{\sum_{i=1}^{N} a_i}{N}$

Confidence intervals for other measures

For other measures, we can use a simple, yet effective technique:

- *Utilization* $\longrightarrow U = \frac{B}{T}$
- *Response time* (again, later we will see why!)
- *Average queue length* $\longrightarrow N = \frac{W}{T}$
- *Throughput* $\longrightarrow X = \frac{C}{T}$

- *Arrival rate* $\longrightarrow \lambda = \frac{A}{T}$
- *Average Service time*

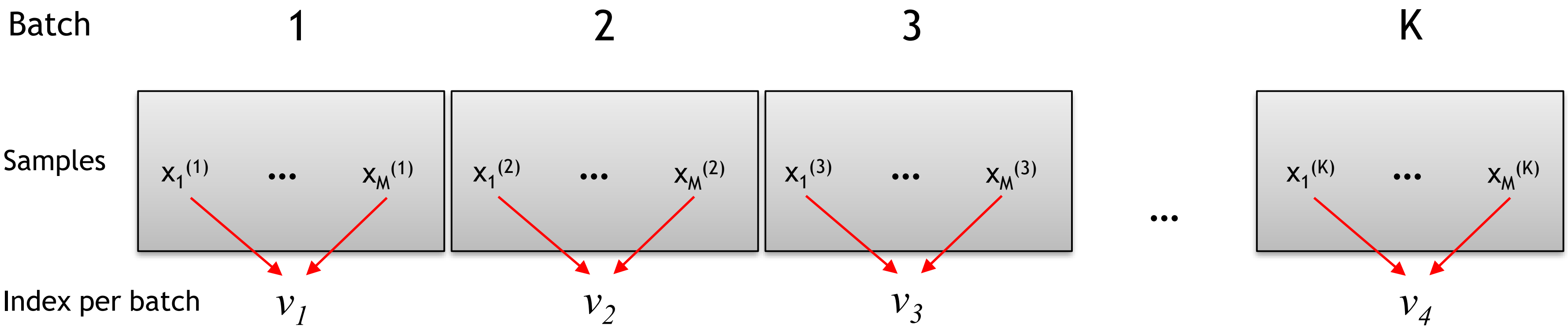
- *Average Inter-arrival time*

Confidence intervals for other measures

Instead of considering a large number of samples *N*, we consider *K* smaller *runs* of *M* events, so that *N* = *M* · *K* .

For each of the *K* runs, the desired measures are computed considering only its samples (*M* in total).

The confidence interval is then computed at end considering the value obtained in each batch as an instance of the measure.


Batch
1
2
3
K
Samples
$x_1^{(1)}$ ... $x_M^{(1)}$
$x_1^{(2)}$ ... $x_M^{(2)}$
$x_1^{(3)}$ ... $x_M^{(3)}$
...
$x_1^{(K)}$ ... $x_M^{(K)}$
Index per batch
v_1
v_2
v_3
v_4

Final result

$$\bar{v} = \frac{1}{K}\sum_{i=1}^{K} v_i$$

Confidence intervals for other measures

For example, considering the Utilization, and calling $s_j^{(i)}$ the service time of the jth job of the ith batch, and similarly $a_j^{(i)}$ and $c_j^{(i)}$ respectively the arrival and completion times:

$$B_i = \sum_{j=1}^{M} s_j^{(i)} \qquad T_i \cong c_M^{(i)} - a_0^{(i)}$$

$$U_i = \frac{B_i}{T_i} \qquad \forall\, 1 \leq i \leq K$$

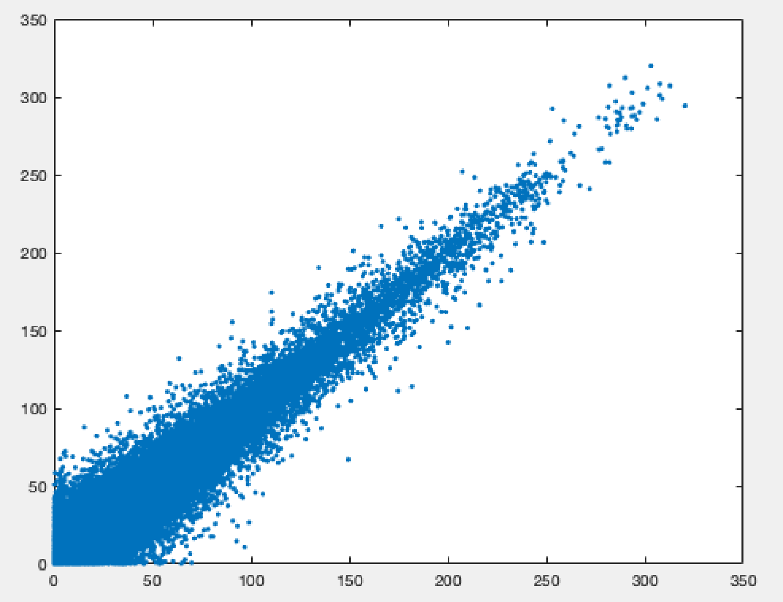
$$\bar{x} = \frac{1}{K}\sum_{i=1}^{K} U_i \qquad s^2 = \frac{1}{K-1}\sum_{i=1}^{K}(U_i - \bar{x})^2 \qquad d_\gamma = -\theta_{N<0,1>}((1-\gamma)/2)$$

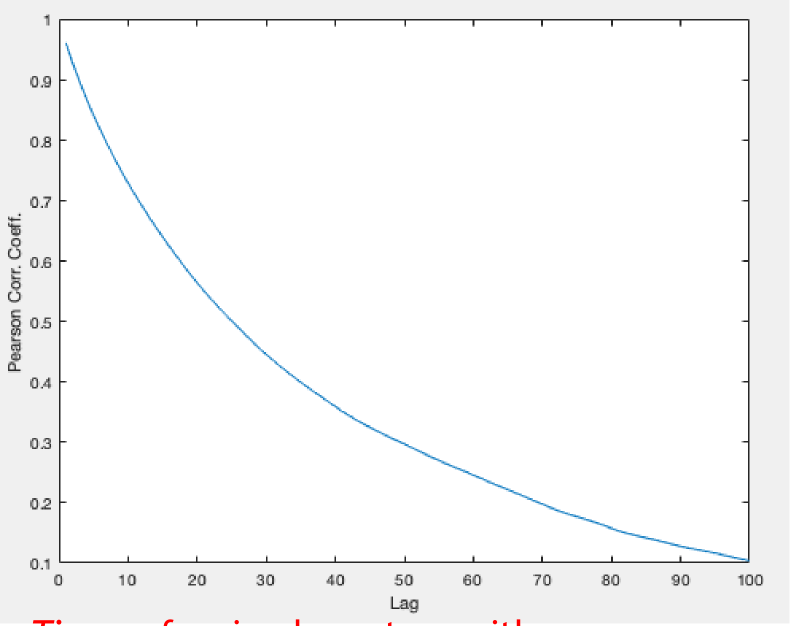
$$U = \left[\bar{x} - d_\gamma\sqrt{\frac{s^2}{K}}, \bar{x} + d_\gamma\sqrt{\frac{s^2}{K}}\right]$$

Confidence intervals for other measures

Some performance indices, such as the *Response Time*, produce measures that are very correlated.

The previous techniques seen for confidence intervals, requires samples to be independent: for this reason the *K batches* of *M samples* technique is usually used to compute intervals for the response time, since the aggregation reduces the correlation.





Correlation among successive *Response Times* of a simple system with exponential services and inter-arrival times

Stopping criteria

The confidence level determines the *reliability* of our results: the probability, which we are willing to take, that the actual solution will be outside the predicted range.

However, if the considered number of samples $N = K \cdot M$ gives a result whose confidence interval is too large for our purposes, we have to extend the number of runs.

Many techniques aims at looking for the minimum number of samples to reach a given accuracy.

Stopping criteria

The required accuracy is generally defined in term of the *Relative Error* α : for a confidence interval *(u_l, u_u)*, it is defined as the ratio between the size of the confidence interval u_u - u_l, and its average (u_u + u_l)/2 :

$$\alpha = \frac{u_u - u_l}{\frac{u_u + u_l}{2}} = 2\frac{u_u - u_l}{u_u + u_l}$$

The user specifies the *Maximum Relative Error* α : the relative error that the she seeks to obtain with the given confidence level.

Stopping criteria

A simple algorithm, declined in two versions depending on whether we consider sampled or batches, which starts from N_0 samples or K_0 batches, and increases every trial of ΔN samples or ΔK batches, are the following:

```
N = $N_0$
repeat
    [$u_l$, $u_u$]=AvgConfInt(N, $\gamma$)
    N = N + $\Delta$N
until $\frac{u_u - u_l}{\frac{u_u + u_l}{2}} > \alpha$
```

Please note that generally, improving the confidence interval with a larger data set, usually has a complexity that depends only on ΔN or ΔK, and not on the entire data-set N or K.

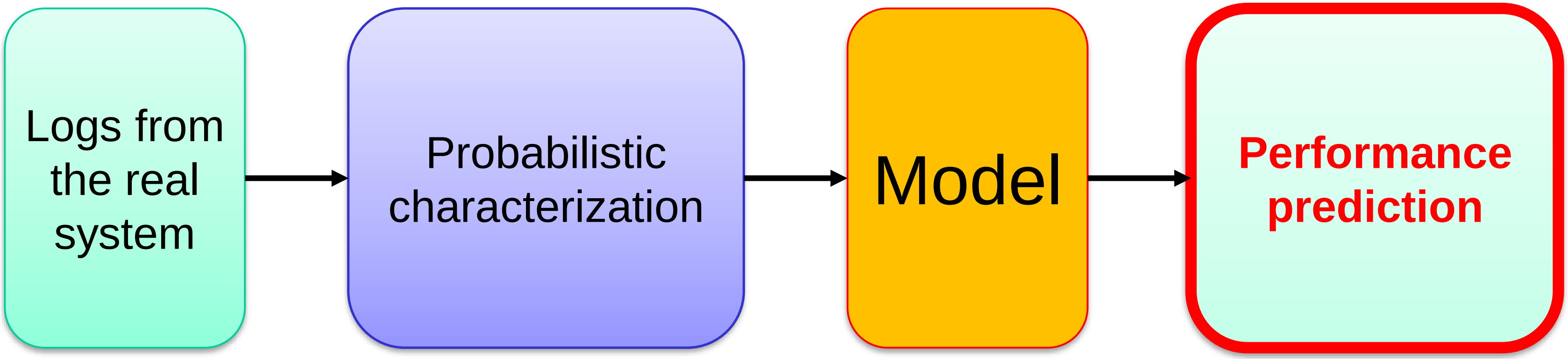
```
K = $K_0$
repeat
    [$u_l$, $u_u$]=BatchConfInt(K,M, $\gamma$)
    K = K + $\Delta$K
until $\frac{u_u - u_l}{\frac{u_u + u_l}{2}} > \alpha$
```

Usually, the algorithms fix a maximum N_{max} or K_{max} after which they stop in any case. Should this happen, the solution is marked as “bad”, but it is reported with the maximum accuracy it could have been reached in the limited number of iterations.

Finding meaningful values for N_0, ΔN, K_0, ΔK, and M, is usually black magic. Also, usually a maximum number of iterations or running time is set, since it is always better to have results that are not accurate, than waiting forever!

Motivation for distributions

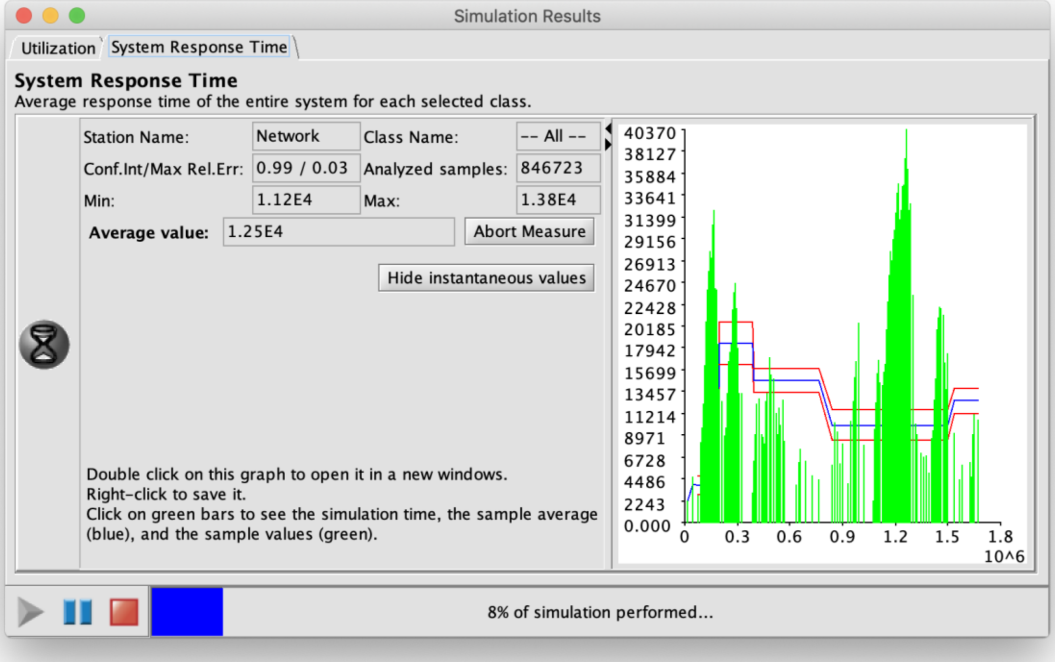
Please note that the main reason for which we prefer to firstcollect a trace, then fit the trace and generate the samples according to the obtained distribution, and finally use these samples for computing the performance prediction, is that *synthetic traces can be extended as much as required to obtain the required accuracy*.



Using a trace directly, it would generally result in a not significative number of samples, which is sufficient to obtain the required accuracy.

Analysis of Motivating Example

We can choose a confidence level γ, fix a desired accuracy α in terms of maximum relative error $\left(\frac{u_\gamma - l_\gamma}{u_\gamma + l_\gamma / 2}\right)$, and continue to increase the population size of our study until we reach our goal.



Simulation Results

System Response Time

Average response time of the entire system for each selected class.

Station Name: Network Class Name: -- All --

Conf.Int/Max Rel.Err: 0.99 / 0.03 Analyzed samples: 846723

Min: 1.12E4 Max: 1.38E4

Average value: 1.25E4

Double click on this graph to open it in a new windows.
Right-click to save it.
Click on green bars to see the simulation time, the sample average (blue), and the sample values (green).

8% of simulation performed...