Meeting Web Application Performance Service Level Requirements Head-on

James F Brady
Capacity Planner for the State of Nevada
jfbrady@admin.nv.gov

Performance service level requirements are often included in web application development contracts and SLAs without much thought given to the measurements and analysis needed to demonstrate compliance with those requirements. Many times specifications are so vague they lead to customer and vendor disagreeing over what data should be used to determine compliance, or agreeing on a data collection and analysis scheme that is inconsistent with fundamental statistical principles. This paper looks at what constitutes a good web application performance service level specification and how compliance with that requirement can be demonstrated with a measurement mechanism which mirrors it as well as conforms to standard statistical inference methods.

1.0 Introduction

Performance service level requirements are a standard component of a software development contract or a Service Level Agreement (SLA). Often they are specified in general terms as a single average value that is void of any measurement context e.g., "The average web page response time shall not exceed 1 second." Broad requirements like this are subject to a wide range of interpretations which can be a major source of contention between customers and vendors.

The usual approach to this open ended requirement is to record response times for all web page queries over a convenient time interval and compute their average. If that average does not exceed 1 second, the system is declared to be producing an acceptable response time service level. Although straight forward, this method for determining compliance is a statistical mine field and easily challenged. For example, it does not indicate which queries are in the measurement mix, whether or not multiple measurement intervals are needed, or if those intervals are busy periods. If performance requirements are to be met head-on, parameters like these need to be spelled out so that standard statistical methods that are clear to all parties can be applied to the measurements.

The first step in this process is to choose metrics wisely, so the discussion begins in **Section 2** by defining the characteristics of a good performance metric. A metric is selected in **Section 3** and implemented within the framework of a defendable service level specification which identifies the measurements needed and the statistical methods that apply. An illustration based on a real web site puts these measurements and statistical methods into context. A sizing model is introduced in **Section 4** and an inconsistency between the service levels specified by the model and measurements taken is identified. **Section 5** encapsulates the information contained in the first four sections and emphasizes how important this subject is to **cloud computing**. A glossary of terms is provided for definitional clarity and to aid those less familiar with the statistical terminology used. Text in **red** is hyperlinked to a definition or a document location. Return to the text using the Alt + left arrow keys.

The measurement and analysis approach in what follows is modeled after the one AT&T Bell Laboratories used during the last century to develop telephone equipment resource sizing methods. They applied fundamental traffic flow principles and sound statistical sampling and analysis techniques to produce credible results which clearly indicated if required service levels were being met. During AT&T's 1984 divestiture this author became very familiar with these techniques while working for a major telecommunications company writing contributions for the T1Q1.1 Traffic/Availability Telecommunications Standards Committee [BRAD86].

2.0 Good Performance Metric

Three fundamental characteristics that a good computer performance metric should possess are that it be:

1. Logical,
2. Robust, and
3. Measureable.

These three characteristics are discussed in order.

2.1 Logical

The metric chosen to measure acceptable service levels should be a logical performance indicator such as response time for determining how quickly the system reacts to user queries or throughput to evaluate how much work the system is capable of accomplishing per unit time. Whatever metric is chosen, it needs to be a logical performance indicator for the resources being scrutinized.

2.2 Robust

The specific metric selected ought to be robust in the sense that it possesses what Dr. Buzen calls “Extended Applicability”. This means that it applies in many cases when the assumptions of model parameters are not satisfied exactly, making it more generally applicable and, therefore, less risky [BUZE16].

Examples of these “Extended Applicability” measurements are:

1. Utilization & Throughput
2. Average response time
3. Average queue length

Examples of more risky measurements, or what Dr. Buzen refers to as having “Standard Applicability”, are:

1. Response time percentiles (97.5% < 1 sec)
2. Response time distributions

The “Extended Applicability” metrics tend to be indicators of central tendency, which in a probability distribution sense, is usually the **mean** or first **moment** about the origin. The “Standard Applicability” metrics, on the other hand, are normally probability statements implying the probability distribution is known and, therefore, a complete set of **moments** are available. There is clearly a great deal less uncertainty, or risk, associated with using an indicator of central tendency than assuming knowledge of an entire probability distribution.

2.3 Measureable

Determining whether or not a given responsiveness or capacity is being achieved requires measurements to be taken which can be put into statistical context. In fact, the measurement mechanism ought to be stated within the service level requirement and that requirement should answer the following three questions:

1. How long is each measurement interval?
2. When are these measurements taken?
3. How many measurements are required?

If these questions are answered within the specification, measurement ambiguity is dramatically reduced.

3.0 Service Level Requirement

If the service level metric is logical, robust, and, above all, measureable the requirement, “The average web page response time shall not exceed 1 second”, could be made more complete and defendable by stating it as follows:

Service Level Requirement

One second must be greater than or contained within a 95% confidence interval about the mean of the busy hour response times for the busiest twenty contiguous business days (M-F) of the calendar year excluding holidays. Each busy hour value is the average of the round trip time for Type 1 web event couplets flowing between the Web/App server and the Database server.

This later specification is far more complete and defendable than the previous one but contains a significant number of moving parts which need explanation and statistical context placement. The best way to unravel the key components of this requirement is through an illustration that has a real world foundation like the following.

3.1 Illustration

The illustration is based on a real State government web site, referred to here as Web.gov, where citizens and state employees make queries and perform updates to a database used to run the State's business operations. Figure 1 shows the computing environment with two Web/App servers and a Database server. This figure also indicates that round trip response time is captured on the Web/App Servers by a custom designed internal logging mechanism which creates daily files and produces timestamped records at the HTTP GET and POST level.

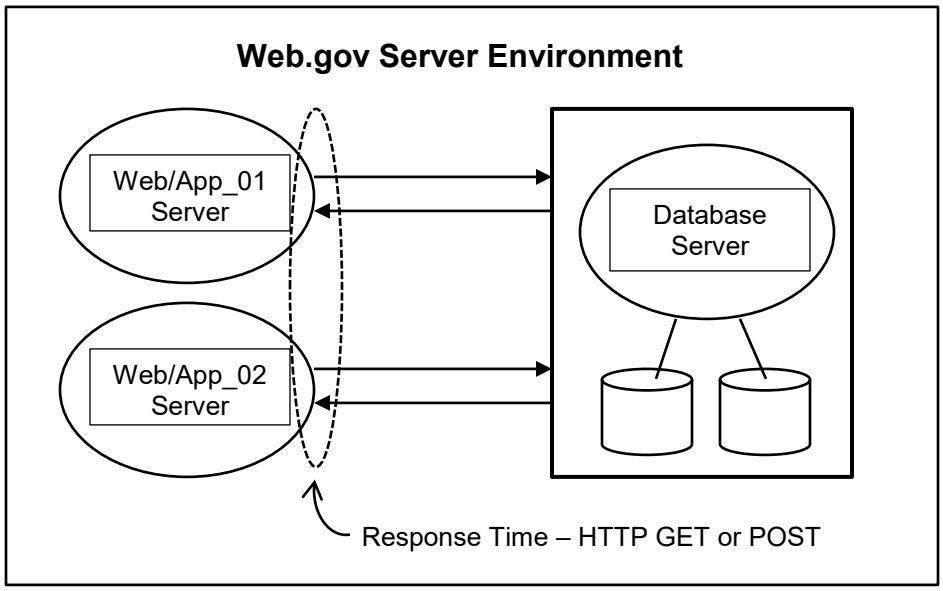


Figure 1: Web.gov Server Environment

The Type 1 web events mentioned in the **service level requirement** and their associated HTTP methods are shown in Table 1 using generic event labels for security reasons. This author created a set of Perl language scripts [Perl15] which capture the latency information contained in the logs and produce count and response time statistics by event label on an hourly basis for each Web/App Server.

Web.gov Type 1 Web Events		
Number	**Event Label**	**HTTP Method**
1	View_Citizen_Data	GET
2	Add_Citizen_Data	POST
3	Delete_Citizen_Data	POST
4	Update_Citizen_Data	POST
5	View_Agency_Data	GET
6	Add_Agency_Data	POST
7	Delete_Agency_Data	POST
8	Update_Agency_Data	POST
9	View_State_Employee_Data	GET
10	Add_State_Employee_Data	POST
11	Delete_State_Employee_Data	POST
12	Update_State_Employee_Data	POST

Table 1: Web.gov Type 1 Web Events

The Web/App_01 and Web/App_02 server log records are merged as the initial step in implementing the busy period selection process illustrated in Figure 2. This figure lists the days of the year on the left, the twenty contiguous business days (excluding holidays) having the highest transaction rate in the middle, and the hour for each of these days with the highest transaction rate as a hashed bar on the right. These business days are referred to here as the **Busy Season (BS)** and their high traffic hour is the **Busy Hour (BH)**.

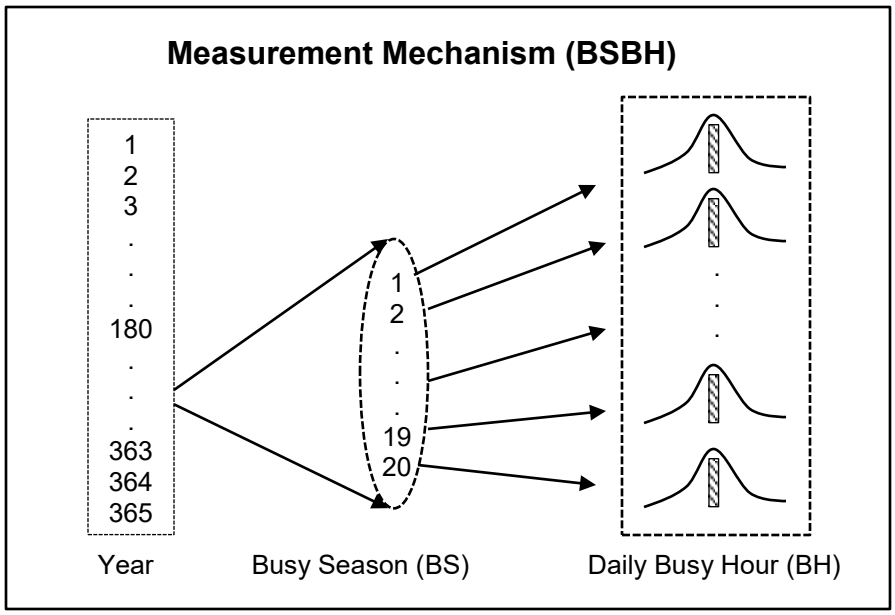


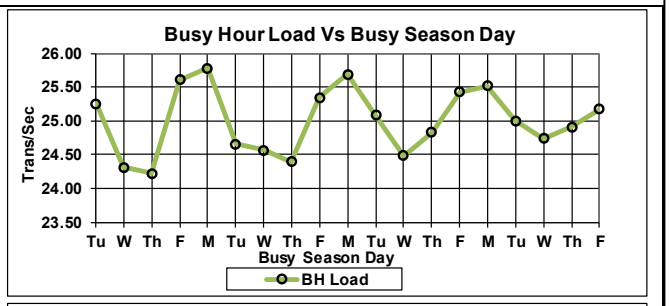
Figure 2: Busy Season (BS) Daily Busy Hour (BH) Time Period Selection

When this time period selection process is complete the results look like the **time series** data shown in Figure 3. This graph and table represent Type 1 **Busy Hour (BH)** workloads and corresponding **mean** response times for the **Busy Season (BS)** where the first Monday in the series is a holiday and therefore excluded from the analysis. The **mean** and **standard deviation (Sdev)** for these 19 workloads and response times are computed and listed at the bottom of the table.

Busy Season BH Loads And Response Times - Time Series

M/M/1 Serv Rate Trans/Sec= 26.00

Count	DoW	Busy Season Day	BH Load Trans/Sec	Response Time (Sec)
0	M	00_Holiday	- - -	- - -
1	Tu	01_Tuesday	25.26	1.35
2	W	02_Wednesday	24.31	0.59
3	Th	03_Thursday	24.22	0.56
4	F	04_Friday	25.60	2.52
5	M	05_Monday	25.78	4.49
6	Tu	06_Tuesday	24.66	0.74
7	W	07_Wednesday	24.57	0.70
8	Th	08_Thursday	24.40	0.62
9	F	09_Friday	25.34	1.53
10	M	10_Monday	25.69	3.23
11	Tu	11_Tuesday	25.09	1.09
12	W	12_Wednesday	24.48	0.66
13	Th	13_Thursday	24.83	0.85
14	F	14_Friday	25.43	1.76
15	M	15_Monday	25.52	2.07
16	Tu	16_Tuesday	25.00	1.00
17	W	17_Wednesday	24.74	0.79
18	Th	18_Thursday	24.91	0.92
19	F	19_Friday	25.17	1.21
		Mean	25.00	1.40
		Sdev	0.49	1.04



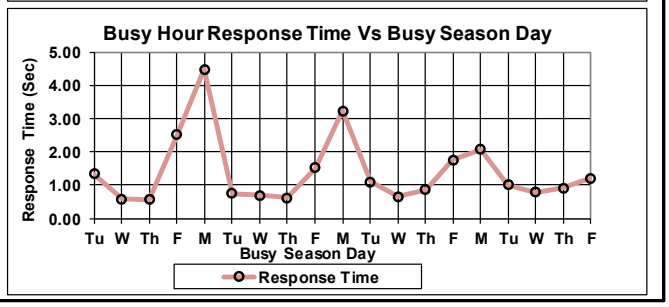


Figure 3: Busy Hour (BH) Load and Response Time over the Busy Season (BS)

In order to mitigate security concerns and illustrate the analysis steps required in a clear and concise manner, Figure 3 contains model data instead of data from the Web/App servers. The model used to produce the data in this figure consists of Type 1 **Busy Hour (BH)** workloads in green that are drawn from a Normal distribution, [HAHN68] and [WIKI17f], with a **mean** of 25 Trans/Sec and response times in red generated using an M/M/1 queueing model, [ALLE78] and [GIFF78], with a service rate of 26 Trans/Sec.

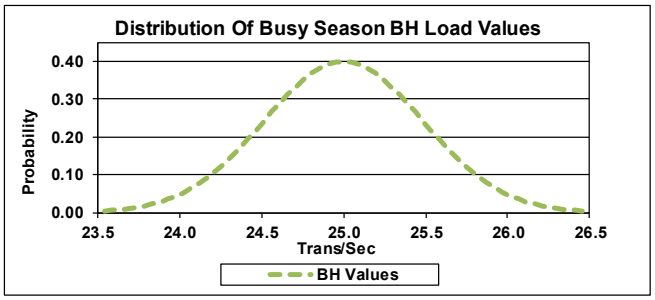
3.2 Distribution of Sample Values

The first step toward creating the **confidence interval** for the **service level requirement** is to transform the **time series** data in Figure 3 into the **probability density functions** in Figure 4. This is accomplished by using the **mean** and **Sdev** values shown to create parameters for the illustrative probability distributions. Microsoft Excel functions are invoked to produce these **probability density functions** which consist of Normally distributed load quantities, [EXCE17b], and Gamma distributed response time values, [EXCE17a]. Normally distributed fluctuations around a **mean** value not close to zero, 25 Trans/Sec, is a plausible representation for the load distribution and the Gamma distributed, [HAHN68] and [WIKI17c], response times are reasonable because the majority of them are short, less than the 1.4 second average, but a few of them are relatively long.

Busy Season BH Load And Response Time - Probability Density Functions

M/M/1 Serv Rate Trans/Sec= 26.00

Count	DoW	Busy Season Day	BH Load Trans/Sec	Response Time (Sec)
0	M	00_Holiday	- - -	- - -
1	Tu	01_Tuesday	25.26	1.35
2	W	02_Wednesday	24.31	0.59
3	Th	03_Thursday	24.22	0.56
4	F	04_Friday	25.60	2.52
5	M	05_Monday	25.78	4.49
6	Tu	06_Tuesday	24.66	0.74
7	W	07_Wednesday	24.57	0.70
8	Th	08_Thursday	24.40	0.62
9	F	09_Friday	25.34	1.53
10	M	10_Monday	25.69	3.23
11	Tu	11_Tuesday	25.09	1.09
12	W	12_Wednesday	24.48	0.66
13	Th	13_Thursday	24.83	0.85
14	F	14_Friday	25.43	1.76
15	M	15_Monday	25.52	2.07
16	Tu	16_Tuesday	25.00	1.00
17	W	17_Wednesday	24.74	0.79
18	Th	18_Thursday	24.91	0.92
19	F	19_Friday	25.17	1.21
		Mean	25.00	1.40
		Sdev	0.49	1.04



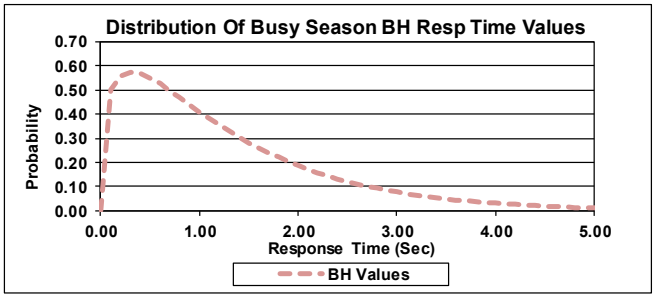


Figure 4: Probability Density Function of BH Load and Response Time over the Busy Season

No matter what the distribution of individual values, however, the **Central Limit Theorem** says that, as sample size increases for a set of independent samples drawn from a population, the distribution of sample averages becomes Normally distributed. This is important because a **confidence interval** is a probability statement and a probability statement cannot be made without specifying a probability distribution. The **Central Limit Theorem** defines that distribution as the Normal distribution for the average **BSBH** response time metric under consideration.

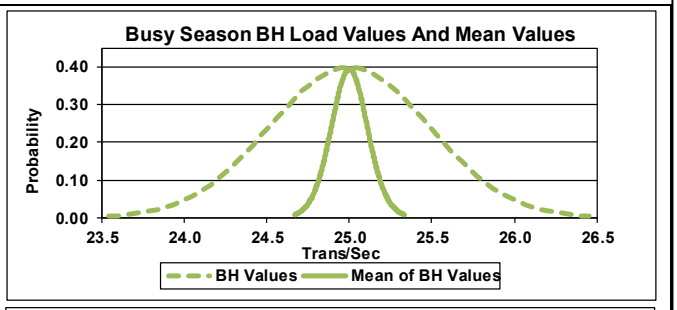
3.3 Distribution of Sample Means

Figure 5 shows both the distribution of individual **BSBH** values and their averages on the same set of charts to demonstrate how much they differ from each other. The two probability distributions representing individual values can vary in shape and have a diverse set of parameters, but the average of the samples becomes Normally distributed as the number of them used to compute the **mean** increase. The individual **BSBH** load values are Normally distributed so the distribution of their average has the same shape but tapers off more quickly in the tails. The individual **BSBH** response times are Gamma distributed with a high degree of **skewness** but their average becomes Normally distributed as well.

Busy Season BH Load And Response Time - Distribution Of Values And Their Mean Values

M/M/1 Serv Rate Trans/Sec= 26.00

Count	DoW	Busy Season Day	BH Load Trans/Sec	Response Time (Sec)
0	M	00_Holiday	- - -	- - -
1	Tu	01_Tuesday	25.26	1.35
2	W	02_Wednesday	24.31	0.59
3	Th	03_Thursday	24.22	0.56
4	F	04_Friday	25.60	2.52
5	M	05_Monday	25.78	4.49
6	Tu	06_Tuesday	24.66	0.74
7	W	07_Wednesday	24.57	0.70
8	Th	08_Thursday	24.40	0.62
9	F	09_Friday	25.34	1.53
10	M	10_Monday	25.69	3.23
11	Tu	11_Tuesday	25.09	1.09
12	W	12_Wednesday	24.48	0.66
13	Th	13_Thursday	24.83	0.85
14	F	14_Friday	25.43	1.76
15	M	15_Monday	25.52	2.07
16	Tu	16_Tuesday	25.00	1.00
17	W	17_Wednesday	24.74	0.79
18	Th	18_Thursday	24.91	0.92
19	F	19_Friday	25.17	1.21
		Mean	25.00	1.40
		Sdev	0.49	1.04



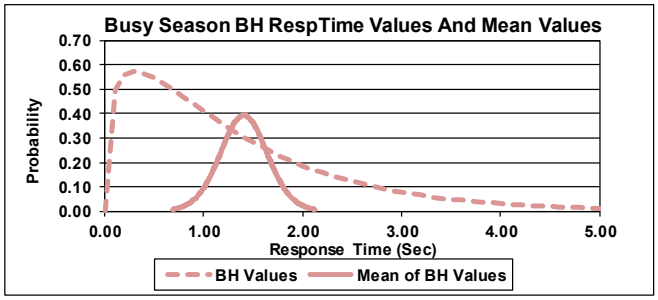


Figure 5: Busy Season Distributions and the Distribution of Their Mean Values

Even though the distributions of sample values and sample averages can have widely dispersed statistical properties, there is a direct relationship between their means and standard deviations [HOEL62] which is:

$Let;$
$x_i = i^{th} sample\ value$
$\bar{x} = mean\ of\ sample\ values$
$\bar{\bar{x}} = mean\ of\ sample\ mean\ values$
$s = standard\ deviation\ of sample\ values$
$s_{\bar{x}} = standard\ deviation\ of sample\ mean\ values$
$n = number\ in\ the\ sample$

$Then;$

$$\bar{\bar{x}} = \bar{x} \quad (1)$$

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \quad (2)$$

$Where;$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (3)$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \quad (4)$$

These relationships are applied in **Section 3.5**.

3.4 Student's t Distribution

The convergence of the distribution of sample means to the Normal distribution as $n \to \infty$ only applies for large sample sizes so are the 19 values in the **Busy Season** considered a large number of samples? Statistical inference says that for samples less than 30 the Student's t distribution, [HOEL62] and [WIKI17i], is a better fit than the Normal distribution. The Student's t **probability density function** looks bell shaped like the Normal but is more spread out at its base and its curvature depends not only on the **standard deviation** but the number of samples, referred to as its **Degrees of Freedom (DoF)**, where $DoF = n - 1$.

Figure 6 is a comparison of a **Normal(0,1)** Vs a **Student's t(0,1)** with **DoF** = 4. The Student's t distribution is

symmetric like the Normal distribution but its peak is lower and it tapers off more slowly in the tails. The shape of the Student's t [EXCE17c] becomes more like the Normal as its **DoF** increase. This is an important characteristic that is discussed when the **confidence interval** probability statement is formulated in **Section 3.5**.

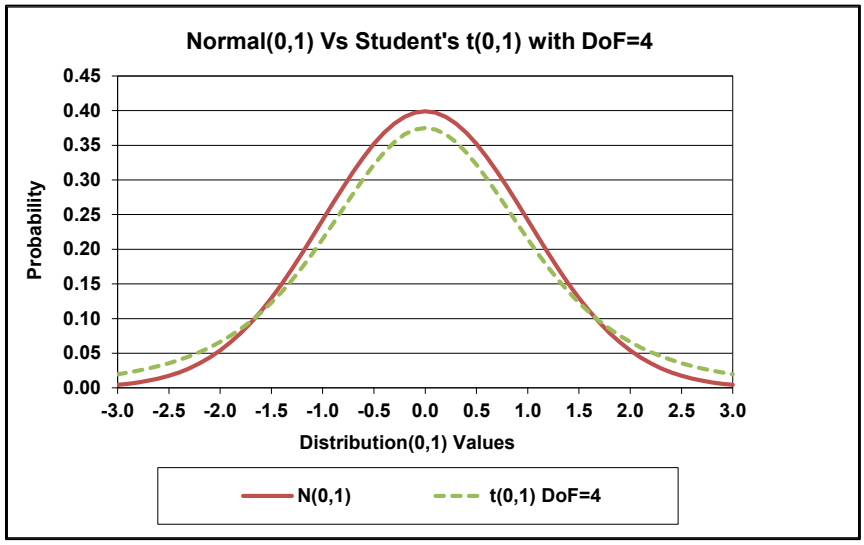


Figure 6: Normal N(0,1) Distribution Vs Student's t(0,1) DoF=4 Distribution

The Student's t distribution has another property that makes it attractive for these smaller sample sizes. It does not depend on any unknown population parameters. Hence, there is no need to replace parameter values by questionable sample estimates as there is with the large sample Normal distribution [HOEL62].

3.5 Confidence Interval

How does the fact that the **mean** of the **BSBH** response time measurements is Student's t distributed apply to the 1 second service level requirement? If 1 second is greater than or lies within the 95% **confidence interval** the **service level requirement** is being satisfied with reasonable probability but, if not, the assumption is made that the target response time is not being achieved and performance improvements are needed.

Eq. 5 symbolically represents the Student's t distributed 95% **confidence interval** which is constructed with **Eq. 1**, **Eq. 2**, and a **Student's t(0,1)** distribution whose **DoF** = 18. When the response time **mean** and **Sdev** from Figure 5 are substituted into **Eq. 5** the result is **Eq. 6** which, when simplified, yields the **CI** in **Eq. 7** .

$$\bar{x} + t_{.025}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{.975}\frac{s}{\sqrt{n}} \tag{5}$$

$Where$:
$\mu = Confidence\ Interval\ variable$
$\bar{x} = mean\ of\ sample\ values = 1.40$
$s = standard\ deviation\ of sample\ values = 1.04$
$n = sample\ size = 19$
$t_p = Student's\ t\ Distribution\ value\ for\ probability\ p\ and\ n-1\ degrees\ of\ freedom$
$t_{.025} = T.INV(.025,18) = -2.10$
$t_{.975} = T.INV(.975,18) = 2.10$

$Then$:

$$1.40 - 2.10\frac{1.04}{\sqrt{19}} < \mu < 1.40 + 2.10\frac{1.04}{\sqrt{19}} \tag{6}$$

Confidence Interval (CI):

$$0.90 < \mu < 1.90 \tag{7}$$

Figure 7 shows the **confidence interval** of **Eq. 7** pictorially. Because the dotted arrow in this illustration representing the 1 second **mean** limit falls inside this **confidence interval**, the assertion that the system is meeting its response time **service level requirement** for Type 1 web requests cannot be rejected. If the **mean** limit is 0.8 seconds instead of 1.0 seconds the **confidence interval** is above the limit and the assumption is made the system is not achieving its Type 1 web request response time requirement so performance improvements are necessary.

Busy Season Mean Response Time 95% Confidence Interval (+-.025)

M/M/1 Serv Rate Trans/Sec= 26.00

Count	DoW	Busy Season Day	BH Load Trans/Sec	Response Time (Sec)
0	M	00_Holiday	- - -	- - -
1	Tu	01_Tuesday	25.26	1.35
2	W	02_Wednesday	24.31	0.59
3	Th	03_Thursday	24.22	0.56
4	F	04_Friday	25.60	2.52
5	M	05_Monday	25.78	4.49
6	Tu	06_Tuesday	24.66	0.74
7	W	07_Wednesday	24.57	0.70
8	Th	08_Thursday	24.40	0.62
9	F	09_Friday	25.34	1.53
10	M	10_Monday	25.69	3.23
11	Tu	11_Tuesday	25.09	1.09
12	W	12_Wednesday	24.48	0.66
13	Th	13_Thursday	24.83	0.85
14	F	14_Friday	25.43	1.76
15	M	15_Monday	25.52	2.07
16	Tu	16_Tuesday	25.00	1.00
17	W	17_Wednesday	24.74	0.79
18	Th	18_Thursday	24.91	0.92
19	F	19_Friday	25.17	1.21
		Mean	25.00	1.40
		Sdev	0.49	1.04
		LCL		0.90
		UCL		1.90

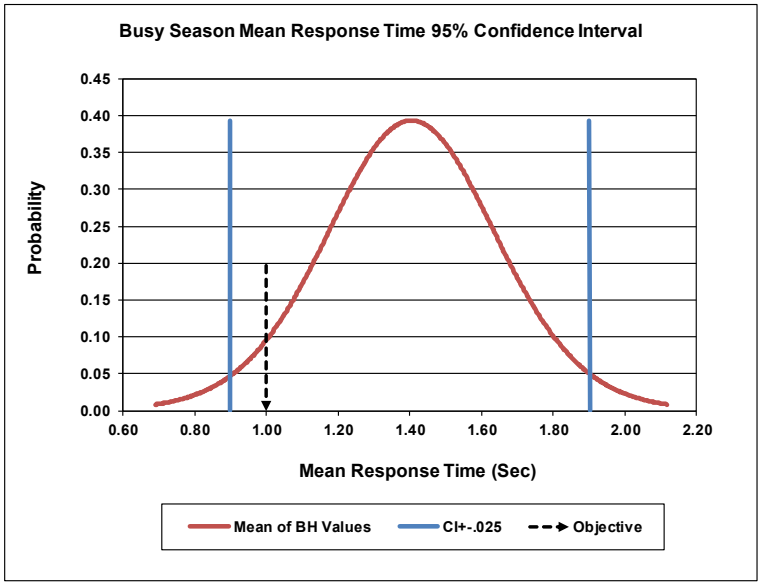


Figure 7: Busy Season Mean Response Time 95% Confidence Interval (+-.025)

In the example just discussed there were 19 busy hour samples used to produce the **mean** value statistic and associated **confidence interval** so how sensitive are these calculations to sample size? Figure 8 provides the answer using a **Student's t(0,1)** distribution by comparing a 5 Vs 19 sample environment.

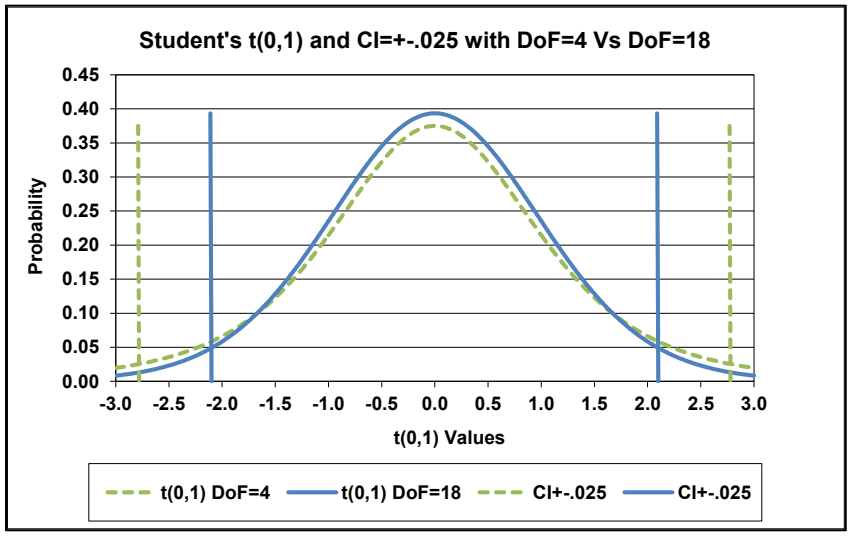


Figure 8: Student's t(0,1) with CI+-.025 for DoF=4 Vs DoF=18

Clearly, the Student's t distribution with 5 samples is shorter and wider, creating a significantly broader **confidence interval**. This broader confidence band, ± 2.78, introduces greater uncertainty regarding whether or not the required service level is being met than the larger 19 sample band, ± 2.10. This result is intuitive because larger samples provide more information than smaller ones, reducing the uncertainty for a given probability level.

4.0 Sizing Model Vs Measurement Mechanism

In **Section 3.1** an M/M/1 queueing model is applied to a set of Normally distributed **BH** loads to produce the response times listed in Figure 3. In this section the problem is turned around and the queueing model is used to determine how fast a single processor system needs to be to achieve an average **BSBH** response time of 1 second for the Type 1 web events.

The standard approach to this sizing problem is to plug the average load, 25 Trans/Sec, and the 1 second response time into the sizing model to compute the required service rate.

Let:
$R = response\ time\ requirement = 1\ second$
$\lambda = arrival\ rate = 25\ Trans/Sec$
$X = service\ rate\ needed$

$For\ an\ M/M/1\ queueing\ model;$

$$R = \frac{1}{(X-\lambda)} \tag{8}$$

$Solving\ for\ X$:

$$X = \frac{1}{R} + \lambda \tag{9}$$

$Service\ Rate$:

$$X = \frac{1}{1} + 25 = 26\ Trans/Sec \tag{10}$$

The service rate in **Eq. 10** satisfies the 1 second response time requirement for the average load, 25 Trans/Sec. However, the sizing rule is not based on the average load's response time but on the average response time over the **Busy Season**, which Figure 7 indicates is 1.4 seconds for the 26 Trans/Sec service rate in **Eq. 10**. Therefore, this single time period sizing formula yields a service rate that is too small to achieve the response time goal. This is because the formula uses the average load as input which does not account for how variations in load over the 19 day **Busy Season** impact average response time. When the service rate is increased to 26.21 Trans/Sec, as in Figure 9, the average **BSBH** response time goal of 1 is reached. The average load is seldom observed in the real world but is daily entry 16_Tuesday in this illustration. This row in the table has a 0.83 second response time in Figure 9 Vs 1 second in Figure 7.

Sizing Model Service Objective

M/M/1 Serv Rate Trans/Sec= 26.21

Count	DoW	Busy Season Day	BH Load Trans/Sec	Response Time (Sec)
0	M	00_Holiday	- - -	- - -
1	Tu	01_Tuesday	25.26	1.05
2	W	02_Wednesday	24.31	0.53
3	Th	03_Thursday	24.22	0.50
4	F	04_Friday	25.60	1.65
5	M	05_Monday	25.78	2.31
6	Tu	06_Tuesday	24.66	0.64
7	W	07_Wednesday	24.57	0.61
8	Th	08_Thursday	24.40	0.55
9	F	09_Friday	25.34	1.16
10	M	10_Monday	25.69	1.92
11	Tu	11_Tuesday	25.09	0.89
12	W	12_Wednesday	24.48	0.58
13	Th	13_Thursday	24.83	0.72
14	F	14_Friday	25.43	1.28
15	M	15_Monday	25.52	1.44
16	Tu	16_Tuesday	25.00	0.83
17	W	17_Wednesday	24.74	0.68
18	Th	18_Thursday	24.91	0.77
19	F	19_Friday	25.17	0.96
		Mean	25.00	1.00
		Sdev	0.49	0.51

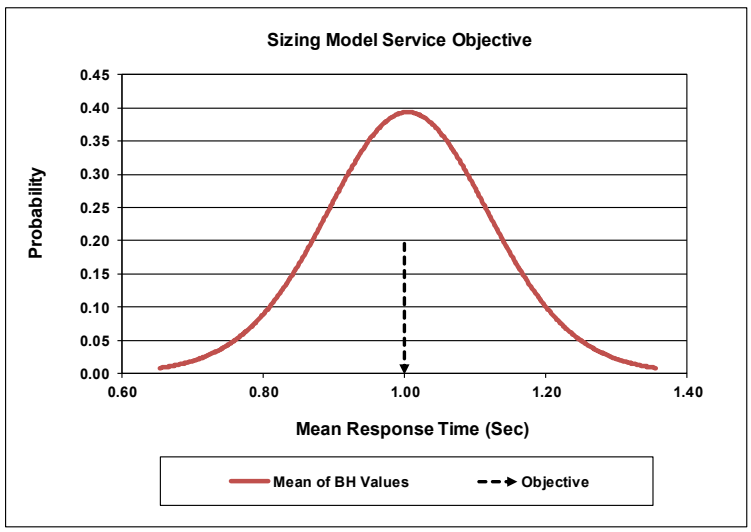


Figure 9: Sizing Model Service Objective 95% Confidence Interval (+-.025)

This service rate increase is not a straightforward calculation because it requires assumptions to be made about

the probability distribution of **BSBH** loads as well as it requires development of some complex mathematics that relates the single hour queueing model to the **BSBH** load distribution. Figure 9 does not have the benefit of these insights and is produced from Figure 7 through trial and error by systematically increasing the M/M/1 queueing model service rate.

The inconsistency between the single time period sizing model and the average **BSBH** performance criteria was a very big challenge for AT&T Bell Laboratories during the last century [HILL76]. They devoted a lot of time and effort to developing telephone equipment resource sizing methods that were consistent with the measurement mechanisms used [HAYW83]. For example, the tables developed for sizing trunk groups between switches are based on queueing models but adjusted to accommodate, what they called, day-to-day variation. The average busy hour load during the busy season and the percent of calls blocked service objective are used as input into the sizing table but, unlike **Eq. 8** thru **Eq. 10**, the trunk quantity specified reflects the day-to-day variation adjustment [WILK70].

Why not solve the day-to-day variation problem by taking a sample of size one instead of nineteen? The reason is that measurements are samples from an unknown population and a sample of size one cannot be used to fully represent that population with any degree of credibility. Statistical inference demands that multiple measurements be taken and the uncertainty associated with those measurements made explicit by constructing a **confidence interval** which, for this particular mean value statistic, takes advantage of the **Central Limit Theorem** for small samples. From a Student's t distribution perspective, a sample of size one has zero **Degrees of Freedom** and produces a **confidence interval** that is infinitely wide.

It is nearly always the case that the single time period model under-sizes the resource, which in this example, is processor throughput required. Figure 10 graphically shows why this is true for queueing model based resource sizing. This figure has the M/M/1 queueing model service rate set to 26 Trans/Sec with load and response time data sorted in increasing load level order. Right of the data is a graph with three plots. In green is a **BSBH** load column graph with its vertical axis on the left. In red and blue are response time line charts having their vertical axis on the right. The red response time curve is a plot of the red data as a function of the green data and the blue horizontal response time line is the average of the red response time values.

Busy Season Average Response Time Vs Response Time For Average Load

M/M/1 Serv Rate Trans/Sec= 26.00

Count	Busy Season Day	BH Load Trans/Sec	Response Time (Sec)
0	00_Holiday	---	---
1	03_Thursday	24.22	0.56
2	02_Wednesday	24.31	0.59
3	08_Thursday	24.40	0.62
4	12_Wednesday	24.48	0.66
5	07_Wednesday	24.57	0.70
6	06_Tuesday	24.66	0.74
7	17_Wednesday	24.74	0.79
8	13_Thursday	24.83	0.85
9	18_Thursday	24.91	0.92
10	16_Tuesday	25.00	1.00
11	11_Tuesday	25.09	1.09
12	19_Friday	25.17	1.21
13	01_Tuesday	25.26	1.35
14	09_Friday	25.34	1.53
15	14_Friday	25.43	1.76
16	15_Monday	25.52	2.07
17	04_Friday	25.60	2.52
18	10_Monday	25.69	3.23
19	05_Monday	25.78	4.49
	Mean	25.00	1.40
	% Diff	0%	40%
	Sdev	0.49	1.04

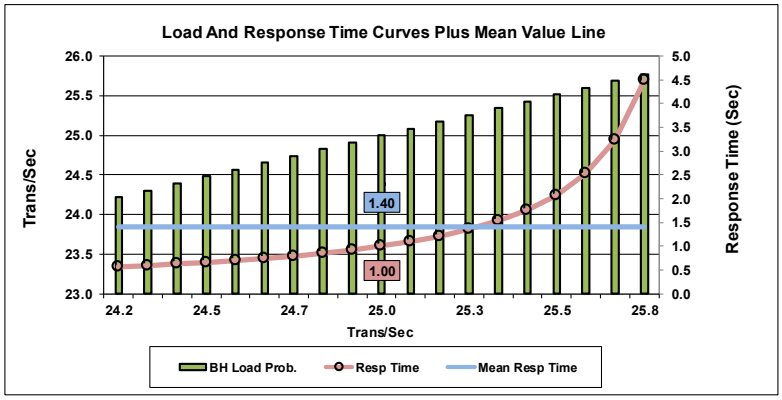


Figure 10: Busy Season Average Response Time Vs Response Time for Average Load

The load levels are increasing in a linear fashion with a **mean**, which is also the **median**, equal to 25 Trans/Sec. The response time for this middle load column, 1.00 seconds, is labeled in red while the average response time across the 19 load values, 1.40 seconds, is in blue. Even though the busy season loads are increasing linearly, the corresponding response times are very non-linear with a larger number of them below the blue line and the few above that average possessing disproportionally large values. Another perspective of this diversity is the response time **probability density function** in Figure 4. That density function is skewed to the right with a very long tail and that long tail is the reason the **mean** (1.40 seconds) is significantly greater than the **median** (1.00 seconds).

5.0 Summary

Application development contracts and SLA's often contain performance service level requirements that are vague and easily challenged. Such requirements are subject to a wide range of interpretations which can be a major source of contention between customer and vendor. These potential challenges can be mitigated if the metrics selected are chosen wisely, the statistical inference techniques employed are credible, and the measurement mechanisms implemented mirror the service level specification. The introduction of **cloud computing** puts a greater emphasis than ever on clear, concise, and measureable **service level requirements**. Computer professionals who implement these unambiguous requirements will be on top of their **clouds**.

The approach taken here is modeled after the methodology developed by AT&T Bell Laboratories during the latter half of the 1900s to dimension telecommunications network resources. The techniques they developed were statistically sound, clearly understood by regulators, and yielded satisfactory customer service levels. The intent in this paper is to leverage the wealth of knowledge they produced and apply it to the computing industry. The measurement term **Busy Season (BS),** introduced in **Section 3.1**, and the sizing model Vs measurement mechanism inconsistency discussed in **Section 4**, are extracted from this knowledge base.

The Web.gov example in **Section 3** is a conceptual illustration abstracted from a real web server environment intended to show what constitutes a defendable performance requirement and provides the analysis steps necessary to determine if that requirement is being achieved. This illustration involves response time service levels for a production web application that is being monitored but the statistical principles used apply to virtually any application environment where service level requirements are specified and measurements produced.

There are several aspects of the **service level requirement** in **Section 3.0** that make it difficult to challenge. First, it is based on a metric possessing "Extended Applicability", average response time. Second, the measurement mechanism, **Busy Hour (BH)** and **Busy Season (BS)**, is spelled out as part of the requirement. Third, the stated service objective, 1 second, is put into context with the 95% **confidence interval** which is a probability statement that can be made because the metric of choice, average **BSBH** response time, is a mean value statistic subject to the **Central Limit Theorem**. Fourth, because there are only 19 measurements involved, the small sample theorem applies and the Student's t distribution is used, instead of the Normal distribution. Finally, response time measurements are clearly specified as the Type 1 web events listed in Table 1.

The **Eq. 7 confidence interval** graphically illustrated in Figure 7 is a rather weak statement by design and emphasizes the fact that a sample mean greater than the service objective, 1.40 seconds Vs 1.0 seconds, is not an automatic indication the system is failing to meet the **service level requirement**. The service objective in that example needs to be less than 0.90 seconds to make the claim that performance requirements are not being met.

Section 4 contains the most subtle and thought provoking material. This author has found no discussions in the computing industry literature regarding the inconsistency described in this section between sizing models and their associated measurement mechanisms. This is not surprising because the mathematicians deriving the queueing formulas for resource sizing focus on tractable solutions that have **time-invariant** properties, while the statisticians developing data analysis techniques for measuring the resources being sized are looking for models that mirror credible sampling plans.

In the 1970's this problem was important to AT&T, a very large and heavily regulated company that periodically faced legal challenges. They sized resources on a massive scale, monitored network performance with a vast set of measurement tools, and routinely reported results to federal and state regulators. The computing industry is a large set of independent business entities with no such infrastructure or compliance organization so the issue isn't even on the radar, but it is a potential credibility problem for Capacity Planners and Performance Analysts.

Hopefully, this paper has met web application **service level requirements** head-on by arming the reader with the tools necessary to establish credible performance service levels and clearly determine if they are being achieved, while leaving little room for a challenge by either custom or vendor.

Glossary

Busy Hour (BH): The highest traffic volume hour of the 24-hour day. Within the context of this paper it is the wall clock hour (e.g., 10:00 AM – 10:59:59 AM) with the highest transaction rate in a 24-hour day.

Busy Season (BS): The set of days with the highest traffic volume. Within the context of this paper it is the twenty contiguous business days (excluding holidays) with the highest transaction rate.

Busy Season Busy Hour (BSBH): The busy hours that occur during the busy season. There are twenty busy hours within the context of this paper but one is a holiday and excluded from the analysis. Also see Figure 2.

Central Limit Theorem (For Sample Mean): The distribution of the mean of $\boldsymbol{n}$ independent observations from any distribution, or even from up to $\boldsymbol{n}$ different distributions, with finite mean and variance approaches a Normal distribution as the number of observations in the sample become large, i.e., as $\boldsymbol{n \to \infty}$**.** Also see [WIKI17a].

Cloud Computing: The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.

Confidence Interval (CI): An observed interval, which generally differs from sample to sample, that potentially includes the unobservable true parameter of interest. If confidence intervals are constructed in separate experiments on the same population following the same process, the proportion of such intervals that contain the true value of the parameter will match the given confidence level. Also see [WIKI17b].

Degrees of Freedom (DoF): The number of values in the final calculation of a statistic that are free to vary.

Kurtosis: From the Greek: *kurtos*, meaning "curved, arching". In probability theory it is a measure of the "tailedness" of the probability distribution of a real-valued random variable. Also see [WIKI17d].

Mean (arithmetic mean): The sum of a collection of numbers divided by the number of numbers collected. It is the first moment about the origin. In sampling terms let $\bar{\boldsymbol{x}} = \boldsymbol{mean}$, then $\bar{\boldsymbol{x}} = \frac{\sum_{i=1}^{n} \boldsymbol{x_i}}{\boldsymbol{n}}$ **,** where $\boldsymbol{x_i = i^{th} sample\ value}$ and $\boldsymbol{n = number\ in\ the\ sample}$.

Median: The middle value of a data set.

Moment: In mathematics, a moment is a specific quantitative measure, used in both mechanics and statistics, of the shape of a set of points. If the points represent probability density, then the zeroth moment is the total probability, the first moment is the mean, the second moment about the mean is the variance, the third moment about the mean is the skewness, and the fourth moment about the mean is the kurtosis. For a distribution of mass or probability on a bounded interval, the collection of all the moments (of all orders, from 0 to ∞) uniquely determines the distribution. Also see [WIKI17e].

Probability Density Function (PDF): a function of a continuous random variable, whose integral across an interval gives the probability that the value of the variable lies within the same interval. Also see [WIKI17g].

Skewness: In probability theory, it is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. Also see [WIKI17h].

Standard Deviation (Sdev): A measure of dispersion which is the square root of the second moment about the mean. In sampling terms *let* $\boldsymbol{s = standard\ deviation}$, then $\boldsymbol{s} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ where $\bar{\boldsymbol{x}} = \frac{\sum_{i=1}^{n} \boldsymbol{x_i}}{\boldsymbol{n}}$**,** $\boldsymbol{x_i = i^{th} sample\ value,}$ and $\boldsymbol{n = number\ in\ the\ sample}$.

Standard(0,1) distribution: A probability distribution with $\boldsymbol{\mu = 0}$ and $\boldsymbol{\sigma^2 = 1,}$ where $\boldsymbol{\mu = mean}$ and $\boldsymbol{\sigma = standard\ deviation}$ Values for Normal(0,1) and Student's t(0,1) are tabulated in statistics books such as **[**HOEL62]. For any random variable $\boldsymbol{x}$ with specified $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$**,** the relationship to the tabulated value $\boldsymbol{z}$ is: $\boldsymbol{z = \frac{x-\mu}{\sigma}}$**.**

Time-invariant: A system whose output does not depend explicitly on time. Within the context of this paper it is a stationary process in steady state where the process being analyzed is a queueing system.

Time Series: A series of data points listed in time order. Also see [WIKI17j].

References

[ALLE78] A.O. Allen, "Probability, Statistics, and Queueing Theory", Academic Press, Inc., Orlando, Florida, 1978.

[BRAD10] J.F. Brady, "Making Statistical Sense out of Time Series Data with 'Home Grown' Perl Scripts", CMG MeasureIT, July 2010. *https://www.semanticscholar.org/paper/Making-Statistical-Sense-out-of-Time-Series-Data-Brady-Brady/a2f82519ccd3b1d1707962e6ca2c69e2184775b7*

[BRAD13] J. F. Brady, “Transforming Time Series Data into Capacity Planning Information,” CMG Conf., 2013.

[BRAD86] J.F. Brady, “Fundamental Components of a Network Performance Standard”, T1 Standards Forum – T1Q1.1 Traffic/Availability Working Group Contribution, May 1986.

[BUZE16] J.P. Buzen, “Risky Business – Modeling and Predicting System Performance”, CMG Conference 2016.

[EXCE17a] Microsoft Office Support, “Excel GAMMA.DIST function”, 2017, *https://support.office.com/en-us/article/GAMMA-DIST-function-9b6f1538-d11c-4d5f-8966-21f6a2201def*

[EXCE17b] Microsoft Office Support, “Excel NORM.S.DIST function”, 2017 *https://support.office.com/en-us/article/NORM-S-DIST-function-1e787282-3832-4520-a9ae-bd2a8d99ba88*

[EXCE17c] Microsoft Office Support, “Excel T.DIST function”, 2017 *https://support.office.com/en-ie/article/T-DIST-function-4329459f-ae91-48c2-bba8-1ead1c6c21b2*

[GIFF78] W.C. Giffin, “Queueing: Basic Theory and Applications”, Grid, Inc, Columbus, Ohio, 1978.

[GTE85] GTE Service Corporation Telephone Operations, “Traffic Grade of Service Standards”, April, 1985.

[HAHN68] G.J. Hahn and S.S. Shapiro, “Statistical Models in Engineering”, John Wiley & Sons, New York, 1968.

[HAYW83] W.S. Hayward and P.J. Moreland, “Theoretical and Engineering Foundations”, The Bell System Technical Journal, Volume 62, Number 7, Sept, 1983.

[HILL76] D.W. Hill and S.R. Neal, “Traffic Capacity of a Probability-Engineered Trunk Group”, The Bell System Technical Journal, Volume 55, Number 7, Sept, 1976.

[HOEL62] P.G. Hoel, “Introduction to Mathematical Statistics”, John Wiley & Sons, New York, N.Y., 1962.

[Perl15] L. Wall, T. Christiansen, “Programming Perl”, Dec, 2015**,** *http://en.wikipedia.org/wiki/Programming_Perl*

[WIKI17a] Wikipedia, “Central Limit Theorem”, 2017, *https://en.wikipedia.org/wiki/Central_limit_theorem*

[WIKI17b] Wikipedia, “Confidence Interval”, 2017**,** *https://en.wikipedia.org/wiki/Confidence_interval*

[WIKI17c] Wikipedia, “Gamma distribution”, 2017, *https://en.wikipedia.org/wiki/Gamma_distribution*

[WIKI17d] Wikipedia, “Kurtosis”, 2017**,** *https://en.wikipedia.org/wiki/Kurtosis*

[WIKI17e] Wikipedia, “Moments”, 2017**,** *https://en.wikipedia.org/wiki/Moment_(mathematics)*

[WIKI17f] Wikipedia, “Normal distribution”, 2017, *https://en.wikipedia.org/wiki/Normal_distribution*

[WIKI17g] Wikipedia, “Probability Density Function”,2017,*https://en.wikipedia.org/wiki/Probability_density_function*

[WIKI17h] Wikipedia, “Skewness”, 2017**,** *https://en.wikipedia.org/wiki/Skewness*

[WIKI17i] Wikipedia, “Student’s t distribution”, 2017, *https://en.wikipedia.org/wiki/Student’s_t-distribution*

[WIKI17j] Wikipedia, “Time Series”, 2017**,** *https://en.wikipedia.org/wiki/Time_series*

[WILK70] R.I. Wilkinson, “Non Random Traffic Curves and Tables for Engineering and Administration Purposes”, Traffic Studies Center, Bell Telephone Laboratories, 1970.

Copyrights and Trademarks