Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

DETERMINATION OF THE "BEST" SYSTEM THAT MEETS A LIMIT STANDARD

Roy R. Creasey, Jr.

College of Business and Economics
Longwood University
201 High Street
Farmville, VA 23901, U.S.A.

K. Preston White, Jr.

University of Virginia
151 Engineers' Way
Charlottesville, VA 22904, U.S.A.

Melanie B. Marks
Bennie D. Waller

College of Business and Economics
Longwood University
201 High Street
Farmville, VA 23901, U.S.A.

ABSTRACT

This paper describes on-going research, where we compare, via simulation experiments, a stochastic system to a standard. We are particularly interested in a subset of standards we call *limit standards*. A limit standard is a maximum or minimum benchmark derived from requirements, another model, or the actual system. The problem is to determine if a system meets the limit standard at customer-defined probabilities. Then, for those systems that meet the limit standard, identify which system is the "best," or results in the lowest probability of reaching the standard. Current comparison methods are based on expected value and cannot solve this type of problem. We outline a two-step approach, using methods from acceptance sampling and ordered statistics, to solve this problem.

1 INTRODUCTION

A "standard" is an identified benchmark or source to which the output of a simulation is compared (Nelson and Goldsman 2001, Law and Kelton 2000). The most desired standard for comparison is the actual system. Many times, however, this standard is not available, and the analyst must look to other sources. These can range from the use of another simulation replication to the use of system requirements.

Current methods of comparing a simulation output to a standard only consider expected performance. (Nelson and Goldsman 2001, Law and Kelton 2000). The mean of the observations in an experiment is compared to the standard. Therefore, the standard must be defined as an expected value (mean) in order for these methods to provide a statistically valid conclusion. Miller, et al (2002) consider selecting the best system based on the number of times the observations within a system exceed the observations within other systems.

A limit standard is defined as a discrete maximum or minimum value, which the customer desires to not violate. Limit standards are present in many operational, economic, and design models. If the standard under investigation is a limit standard, use of the current comparison methods will result in the analyst making an incorrect conclusion about the model.

For example, when developing a model of parking patterns at a university, a limit standard is the number of parking spaces available. It is not possible to exceed this standard, although attempts to do so are always made by students.

The goals of this comparison are to determine if a system meets the limit standard and, if a group of systems meet the standard, to determine the "best" system. In other words, we want to find the system that meets the standard while providing the maximum probability of not exceeding the standard. In this paper, we define a "system" as a specific combination of input and process random variables in a model.

Returning to the parking model, the number of commuter students, faculty and staff vehicles authorized to park on campus exceeds the number of available parking spaces. From a resource allocation standpoint, it is not possible to have a parking space for every authorized vehicle. It is equally undesirable to regularly reach the limit standard. Therefore, the goal is to find the combination of vehicle arrivals and parking durations that result in the "best", or highest, probability of not reaching full capacity.

We identify a solution to the limit standard problem, using concepts from ordered statistics and acceptance sampling.

2 PROBLEM DEFINITION

For the purpose of brevity, we will only define the problem for a maximum limit standard, using a two-step approach. The problem and procedure for a minimum standard is the same as the maximum limit standard, but with reversal of sign in some equations.

The first step involves selecting systems which meet the standard. The second step determines, from the systems that meet the standard, which system is the "best."

Let ω_i denote the ith system under investigation, where $i = 0, 1, \ldots, k$. Let ω_0 denote the standard system. The other systems will be referred to as the alternative systems. Define X_{ij} as the jth observation of interest from the ith system, for $j = 1, 2, \ldots, n$. For the standard system, there is only one observation of interest, which is the limit standard. The limit standard is defined as a *known standard*, meaning it is a discrete value with no variation. (Nelson and Goldsman 2001) Let T denote this limit standard.

It is assumed that the number of observations for each alternative system, n_i, remains constant. While there is nothing to prevent a difference in n_i between systems, this assumption is used to provide a very simple selection and ordering procedure. For example, instead of the simple ordering method described in the second step, using different values of n_i would require the use of Equations (8) for each system and Equation (9) when comparing each system. Therefore, the rest of this paper will simply refer to the number of observations in a system as n. Also, in this paper, it is assumed that $X_{i1}, X_{i2}, \ldots, X_{in}$ are independent and identically distributed (IID).

Define d_i as the number of observations that exceed the standard in system i. In other words,

$$d_i = \sum_{j=1}^{n} b_j \ , \tag{1}$$

where $b_j = 1$ when $X_{ij} > T$ and $b_j = 0$ otherwise.

Define c as the number of observations in a sample of size n that are allowed to exceed the standard. The value of c is dependent on several factors and is discussed in the next section.

Compliance to the standard is then determined for each alternative system using the following rule:

$$\text{If } d_i \le c, \text{ accept } \omega_i \ .$$

In other words,

$$\Pr\,[\omega_i \text{ meets standard}] \ge 1 - \alpha, \text{ given } d_i < c. \tag{2}$$

In (2), α is defined as a Type I error, or the probability of rejecting the system when the number of observations exceeding the standard are c or less.

A rejection of (2) invokes the possibility of a Type II error, meaning ω_i is rejected when it should have been selected. That is,

$$\Pr\,[\omega_i \text{ meets standard}] \le \beta, \text{ given } d_i > c. \tag{3}$$

If no alternative system meets the limit standard, then the standard system should be accepted, or

$$\Pr\,[\text{selecting } \omega_0] \ge 1 - \alpha \text{ when } d_i > c \tag{4}$$

for $i = 1, 2, \ldots, k$.

In other words, if an alternative system is not selected, the standard system should be selected as the "best" system with probability $1 - \alpha$. (Nelson and Goldsman 2001)

If only one alternative system meets the standard after this first step, that system is designated the "best" system. However, if more than one alternative system meets the standard, we must proceed to the second step.

Each remaining system is then ordered, based on its probability of acceptance, or the probability of not exceeding the standard. To simplify the ordering, we only need to consider the number of "failed" observations in each system, or

$$d_{(1)} < d_{(2)} < \ldots < d_{(m)}, \tag{5}$$

where m = number of remaining systems.

The ordered systems are designated $\omega_{(i)}$, which indicates the ith "best" system based on the probability of acceptance (P_a) of the system.

In the event of a tie, or $d_{(i)} = d_{(i+1)}$, order those systems based on the one-tailed t-statistic for the system, or

$$t_{(i)} = \frac{T - \overline{X}_{(i)}}{s_{(i)}}, \tag{6}$$

where $\overline{X}_{(i)}$ is the sample mean for the system and $s_{(i)}$ is the sample standard deviation for the system.

Define δ as a percent t-statistic indifference-zone selected by the customer. The system with the largest t-statistic is designated the "best" of those systems if

$$t_{(i)} \ge t_{(i+1)}(1+\delta)\,. \tag{7}$$

Otherwise, it is possible for any of those systems to be considered the "best".

Alternative system $\omega_{(1)}$ is designated the "best" system. This designation indicates that $\omega_{(1)}$ has the smallest probability of exceeding the standard. Note that it is possible for other alternative systems (such as $\omega_{(2)}$) to have a P_a very close to that of $\omega_{(1)}$. However, for a sufficiently large sample size n, the ordering system specified above assures that $\omega_{(1)}$ will be as good as or better than these other systems in meeting the limit standard.

3 PROCEDURE

For each system of n observations, we are interested in only one characteristic of each observation – whether or not the observation exceeds the standard. We can therefore treat this characteristic as a binomial distribution, irregardless of the underlying distribution of each observation. (Conover 1980)

This is the principle applied to acceptance sampling. The distribution of the measurement and the severity of the defect (how much the defect was out of tolerance) is not considered. The lot (or system) is accepted or rejected based on the sample size and the number of defectives observed in the sample. Some quality control analysts prefer to use a Poisson distribution. However, because we are assuming the samples come from a very large 'lot' (essentially of infinite size), "the binomial distribution is the exact probability distribution for calculating the probability of lot acceptance." (Montgomery 2001)

The analyst must establish a critical region α and a proportion p_1 of defectives that is acceptable at $P_a = 1 - \alpha$. If the sample size is known, these values can be used to identify c, where $\Pr[Y > c] = \alpha$ and Y is a binomial random variable with parameters p_1 and n. (Conover 1980)

However, when conducting simulation experiments, we usually do not know the acceptable sample size, or number of necessary replications, in advance of conducting experiments. Several methods to determine n for simulation experiments have been previously defined. (Law and Kelton 2000, Nelson and Goldsman 2001) Most of these methods require a two-stage approach using the sample mean and standard deviation. We recommend a method from acceptance sampling, using two points from a Type-B Operational Characteristic (OC) curve to determine n and c. (Montgomery 2001) One point has already been defined above as $(p_1, 1 - \alpha)$. The analyst must now establish a second point, corresponding to a critical region β and a proportion p_2 of defectives that is acceptable at $P_a = \beta$.

For example, in the parking problem above, we may wish to accept a system, with $P_a = 0.95$, if the proportion of defectives is $p_1 = 0.01$. We wish to accept the system, with $P_a = 0.10$, if the proportion of defectives is $p_2 = 0.02$.

The selection of α, p_1, β, and p_2 are critical to the comparison. Logically, there is very little interest in having observations exceed a limit standard. Therefore, these values should be kept very small. However, to maintain smaller values requires very large sample sizes. Also, the separation between p_1 and p_2 should be very small – typically no more than one or two units of significant measurement. A large separation will result in a large indifference zone.

Once these points are established, n and c can be determined by solving two simultaneous Binomial cumulative distribution functions (cdf).

$$
\begin{aligned}
1 - a &\le \sum_{h=0}^{c} \frac{n!}{h!(n-h)!} p_1^h (1 - p_1)^{n-h} \\
\beta &\ge \sum_{h=0}^{c} \frac{n!}{h!(n-h)!} p_2^h (1 - p_2)^{n-h}
\end{aligned}
\tag{8}
$$

(Montgomery 2001).

Solving these equations requires the use of a spreadsheet, which can produce several combinations of n and c. It is noted that n and c must be positive integers and that $n > c \geq 0$.

Using the values from the parking example above, the combinations of c and n shown in Table 1 can be used.

Table 1: Combinations of c and n for $p_1 = 0.01, \alpha = 0.05$, $p_2 = 0.02$, $\beta = 0.10$.

c	n
0	299
1	473
2	628
3	773
4	913
5	1049

In other words, at least 299 observations are required. If none of the observations exceed the standard, then that system meets the standard. If the analyst is willing to accept some observations exceeding the standard, then larger of values of n are required.

The null hypothesis to test is that the parameter ρ, defined as the proportion of defectives in the system, is less than or equal to p_1. The alternate hypothesis is that ρ is greater than p_1.

$$
\begin{aligned}
H_0&: \rho \leq p_1 \\
H_1&: \rho > p_1
\end{aligned}
$$

The test statistic is d, the number of defectives in a system of n observations. If $d \leq c$, then accept H_0; else, reject H_0. If we accept H_0, we are confident, with probability $P_a \geq 1 - \alpha$, that ρ is less than or equal to p_1.

If only one system is accepted, that system is considered the "best." If no systems are accepted, then the analyst must consider a change to the standard (with customer approval), a change in α, p_1, β, and p_2, or a re-evaluation of model assumptions.

If more than one system is accepted, analysis proceeds to Step 2 below.

4 STEP TWO - SELECTING THE "BEST"

After executing the above procedures, what remains are systems which have a proportion of defectives less than or equal to p_1. Any of these systems can be used, and the standard will be met. Step two is to determine which of these systems is the "best", or has the highest P_a.

Since the Binomial cdf was used to establish the criteria for the critical region, it must be used to determine which of the remaining systems has the highest P_a. For each system, the probability of acceptance is determined by

$$Pa_i = 1 - \sum_{g=0}^{d_i} \frac{n!}{g!(n-g)!} p_1^g (1-p_1)^{n-g} \qquad (9)$$

for $i = 1, 2, \ldots, m$.

As n and p_1 are the same for each system, we only need to be concerned with the number of defectives, d_i, in each system. As d gets smaller, P_a gets larger. Therefore, the ordering system shown in (5) simplifies the calculations for this step.

In the event of a tie, where two or more systems have the same number of defectives, there must be a statistically valid method of determining the "best" from among these systems. So far, we have not considered the *severity* of the defective observations. We would expect that systems with observations significantly higher (or lower, for a minimum limit standard) than the standard are not as desirable as systems with observations close to the standard.

We suggest the use of the t-statistic for breaking ties, but are investigating other methods. For large n, a lower t-statistic usually indicates a higher probability the system will produce observations that will exceed the standard. There are, however, several concerns with this method. First, use of the t-statistic assumes the observations are from a normally-distributed population. Second, since these systems already meet the standard, the difference in the ρ and t-statistic between these systems will be small. As a result, we have the possibility of a system not selected being as good as the system selected. We include an indifference-zone variable to reduce this possibility. We continue to research several indifference-zone procedures.

5 PARKING EXAMPLE

Using Arena®, we modeled the parking patterns of faculty, staff and commuter students at Longwood University. (Kelton, et al 2004) The characteristic of interest was the maximum number of vehicles in the parking model, either parked or looking for a parking space, in a 10-hour day. Specifically, we were interested in the number of times (days) the maximum number of spaces was exceeded. Each system, or model, consisted of a set of arrival rates and durations for each type of driver.

Twenty systems were investigated. For each system, the arrival rates and durations of faculty and commuter students were (manually!) modified to reflect expected differences in course schedules. For example, one system reflected 300-level courses in the late morning and 400-level courses in the afternoon, since most of the commuter students are seniors taking 300- and 400-level courses. The arrival rates and durations for staff members were not changed, since the working hours for staff members are usually not affected by changes in course schedule.

The number of parking spaces was set at 687. We also used $p_1 = 0.03$, $\alpha = 0.10$, $p_2 = 0.04$ $\beta = 0.10$ and $\delta = 0.01$. From (8), we selected $n = 265$ (days) and $c = 4$.

Of the twenty systems, six met the criteria established above, and are listed in Table 2.

Table 2: Systems Where $d \leq c = 4$

System	d	mean	s.d.
3	4	680.91	3.54
4	2	678.25	3.17
10	3	677.89	3.66
12	2	680.23	3.06
18	2	678.62	2.99
19	3	680.11	2.96

The three systems with $d = 2$ provide us with a $P_a \geq 0.987$. Of these, System 18 had the largest t-statistic of 2.80. However, System 4 was a close second with a t-statistic of 2.76. Using the rule defined in (7), System 18 is designated the "best" of the systems in meeting the limit statistic. Ten additional replications of Systems 4 and 18 were conducted, using different random number streams for each replication set. The results indicate that both systems maintain similar statistics, with System 18 as the "best" in eight of the ten replications.

6 CONTINUED RESEARCH

We continue to research many aspects of comparison to a limit standard.

- As always, a model is only as good as its input data. We continue to gather more data on traffic patterns to refine the models.
- Except when conducting additional replications of Systems 4 and 18, we used Arena's ® default Common Random Number (CRN) stream. Miller, et al (2002) note that the use CRNs may result in a lack of independence across systems. We will investigate the impact in using CRNs.
- The tie-breaking system described above provides a good indication of the "best" system from a group of systems determined to meet the standard. We indicate that errors are inherent in this ap-

proach. It is therefore possible for one system not selected as the "best" to exhibit characteristics similar to the selected system. We will continue to research and refine the selection methods.

- Develop a set of optimization algorithms, to determine the optimal value of input and process variables that minimizes the probability of exceeding the standard.
- Research the use of other quality control techniques in the comparison of standards. Specifically, we will look at sequential sampling methods to handle indifference zones. This method may allow for a smaller number of observations necessary to determine acceptance.
- Research the use of optimization techniques from the multi-armed bandit problem to quickly eliminate systems that will not meet the standard. (Lai and Robbins 1985) These techniques may also reduce the necessary sample size.

REFERENCES

Conover, W.J. 1980. *Practical Nonparametric Statistics*, 2nd ed., Wiley, New York.

Kelton, W., R. Sadowski and D. Sturrock. 2004. *Simulation with Arena®*, 3rd ed., McGraw-Hill, Boston.

Lai, T. and H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied. Mathematics*, 6: 4-22.

Law, A. and W. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd ed. Boston: McGraw-Hill.

Miller, J., B. Nelson and C. Reilly. 2002. Estimating the probability that a simulated system will be the best. *Naval Research Logistics*, 49 (4): 341-358.

Montgomery, D.C. 2001. *Introduction to Statistical Quality Control*, 4th ed., New York: Wiley.

Nelson, B. and D. Goldsman. 2001. Comparisons with a Standard in Simulation Experiments. *Management Science*, 47 (3): 449-463.

AUTHOR BIOGRAPHIES

ROY R. CREASEY, JR. is a Lecturer in Management at Longwood University, and a Ph.D. student in Systems and Information Engineering at the University of Virginia. He received the B.S. from the Virginia Military Institute and the M.E. from the University of Virginia. He has seventeen years prior experience as a systems and test engineer in the defense industry. He is a student member of INFORMS. His e-mail is creaseyrr@longwood.edu

K. PRESTON WHITE, JR. is a Professor of Systems and Information Engineering at the University of Virginia. He received the B.S.E., M.S., and Ph.D. degrees from Duke University. He has held faculty appointments at Polytechnic University and Carnegie-Mellon University and served as Distinguished Visiting Professor at Newport News Shipbuilding and at SEMATECH. He is U.S. Editor for *International Abstracts in Operations Research* and Associate Editor for *International Journal of Intelligent Automation* and *IEEE Transactions on Electronics Packaging Manufacturing Technology.* He is a member of INFORMS, SCS, and INCOSE and a senior member of IEEE and IIE. He sits on the Advisory Board of VMASC and represents IEEE/SMC on the Winter Simulation Conference Board. His e-mail is kpwhite@virginia.edu.

MELANIE B. MARKS is an Associate Professor in Economics at Longwood University. She received the B.S., M.S., and Ph.D. in Economics from Texas A&M University. She is the Director of The Center for Economic Education and a member of the National Council on Economic Education. Her e-mail is marksmb@longwood.edu.

BENNIE D. WALLER is an Assistant Professor in Finance at Longwood University. He received his B.S. from Longwood University, M.B.A. from UNC-Wilmington, and Ph.D. from the University of Mississippi. His research interests include optimization, financial information systems and artificial intelligence. His e-mail is wallerbd@longwood.edu.