A Cost-Based Model for Improving Customer Waiting Times

Alfred L. Guiffrida                        Michael J. Messina
Kent State University                    Gannon University
aguiffri@kent.edu                          messina@gannon.edu

Abstract

The effective management of a service organization requires measuring system performance on many attributes. Customer waiting time is inherent to nearly all service operations and represents a system attribute that is of high concern to both the customers who engage service and the managers of the service operation who are responsible for providing services. Effective performance measures of all system attributes are required to improve service quality. In this paper we present a cost-based, performance model for evaluating the expected cost of untimely customer waiting times for services within a service organization. The present worth of the expected costs due to untimely waiting that accrue over a finite time horizon provide management with input for justifying financial investment to support a continuous improvement program to improve service quality through the reduction of the variability in customer waiting time.

1. Introduction

In today’s competitive business environment, consumers demand increasing higher levels of quality in the services that they seek. In meeting this demand, service organizations must measure their level of quality to benchmark their current level of performance and to support the long-term improvement of service quality. Traditional quality assurance techniques from the manufacturing sector such as Pareto Analysis, statistical control charting, process-flow analysis, and cause-and-effect diagrams have been recommended as tools for improving service quality (see for example [1], [2], [3]). Measuring service quality is complex due to the many dimensions that define service quality [4].

Several researchers have developed specialized tools for measuring service quality based on customer feedback. An inventory of these service measurement tools includes: (i) SERVQUAL [5], (ii) SSQSC [6], (iii) SERVPERF [7], and (iv) FAIRSERV [8].

In addition to measurement issues, service providers are also faced with the task of implementing process improvements to improve service quality [9]. Improvements in service quality may involve the enhancement of an existing service process, or the redesign of a new service process. Break-even analysis has been used to evaluate alternative process designs for improving service quality [10].

In this paper we focus on one aspect of service quality, customer waiting time. Customer waiting time is inherent to nearly all service processes and is of high concern to both customers and managers of service operations [11], [12]. It has been well documented in the literature that the duration of waiting time, actual and perceived, is a crucial factor in a customer’s level of satisfaction with a service experience [13], [14]. Many empirical studies have addressed customer concerns with waiting times in specific industry environments such as banking, public transportation, restaurants, and hotels. A detailed review of such studies is found in [15].

We address the management and control of customer waiting time from a systems theory perspective. General systems theory has been applied to the measurement of service quality [16], [17]. Under systems theory, the service process is viewed as an overall operation that consists of a series of smaller integrated subunits. Customer demand for service acts as a load on the system and the response of the system to this load contributes to the quality of the service received by the customer. Management of the service system is therefore charged with the decision of how best to balance the input and output of the system subject to operational performance goals. Feedback mechanisms within the service system subunits and between the service system and the operating
environment provide management with real time information for monitoring service quality.

The major contributions of this paper are as follows. First, a model is developed that incorporates the variability in customer waiting time at the individual subunit (stage) level of the service process into an overall measure of system-wide customer waiting time. Customer waiting times at the stage level are defined by probability density functions with the resulting probability density function of system-wide customer waiting time defined by the convolution of the stage densities. This approach is in union with contemporary management theories which advocate the reduction of variability as the key step towards improving service system performance as well as overall system performance [18], [19], [20]. However, we extend the modeling of variability in service systems to the subunit stage level.

Second, our use of systems theory in services analysis is unique in that we address the issue of how to financially justify process improvements to support improvement in service quality. We define a systems framework that can be used to assist management in justifying funds for investment in process improvements to improve service quality through the reduction of customer waiting time. The model developed herein fits within the capital budgeting process of the organization and translates the uncertainty of customer waiting time into an expected cost metric. The expected cost metric provides a benchmark for justifying the capital investment required to improve overall system quality.

This paper is organized as follows. In Section 2, a model for evaluating the financial impact of customer waiting time is presented for four candidate probability density functions. In Section 3, the learning-based model framework is presented to support a long term program for improving service quality. The modeling effort is demonstrated for the case of normally distributed customer waiting times. A summary and conclusion for future research is presented in Section 4.

2. Model Development

In this section we present the modeling approach used to define a cost-based model for evaluating service quality through customer waiting. In Section 2.1 we model overall customer waiting time for a service process. In Section 2.2 we present a model for translating the variability of customer waiting time into an expected cost of poor quality.

2.1 Defining the Total System Customer Waiting Time

Consider an $n$-stage serial operation where the duration of customer waiting time at stage $i$, $W_i$, is defined by a continuous probability density function $f_{W_i}(w;\theta)$ with parameter set $\theta$. The total customer waiting time experienced in entire operation is defined by $X = \sum_{i=1}^{n} W_i$. Under the assumption of independence between stages, the form of the probability density function $f_X(x)$ defining overall customer waiting time is defined by the following convolutions:

$$f_{W_i+W_j}(x) = \int f_{W_i}(x-w_j)f_{W_j}(w_j)dw_j$$
$$f_{W_i+W_j+W_k}(x) = \int f_{W_i+W_j}(x-w_k)f_{W_k}(w_k)dw_k$$
$$\vdots$$
$$f_{W_i+W_j+\ldots+W_n}(x) = \int f_{W_i+W_j+\ldots+W_{n-1}}(x-w_n)f_{W_n}(w_n)dw_n$$

Our assumption of independence between stages is not unreasonable as it has been commonly adopted in the literature in modeling multi-stage stochastic systems [21], [22]. When independence cannot be assumed, a correlation structure should be introduced into the model.

The critical computational issue is to determine the analytical form of the density of overall customer waiting time, $f_X(x)$. The mathematics defined in (1) are vastly simplified when the summand density functions $W_i$ are reproductive under addition. For density functions that are reproductive under addition such as the Gaussian (normal) and gamma $f_X(x)$ is relatively easy to define and the evaluation of (1) is straightforward. However, when $f_X(x)$ is composed of the sum non-reproductive densities or the mixture of reproductive and non-reproductive densities, the evaluation of $f_X(x)$ using (1) can often lead to intractable results even for small values of $n$. An effective procedure which uses a discrete convolution algorithm to determine an approximate form for $f_X(x)$ when reproductive under addition does not exist is found in [23].
2.2. Evaluating the Expected Cost of Customer Waiting Time.

The expected cost for customer waiting time is

\[ W = K \int_{c}^{\infty} (x - c) f_X(x) dx \]  

(2)

where

- \( K \) = penalty cost per unit time waiting
- \( c \) = the critical wait time after which \( K \) applies
- \( f_X(x) \) = the probability density function of customer waiting time.

In the definition of (2), the value of the critical waiting duration is set by management. The magnitude of \( c \) can be set in two different ways. Externally, management may formally define and communicate \( c \) as a part of a service guarantee to customers ([24], [25]). For example, Jet Blue now explicitly states in their Customer Bill of Rights that monetary reimbursements will be given for delays. The reimbursements vary according to the length of the delay. Internally, a value of \( c \) based on customer input or expert opinion may be adopted and used to self-examine the current level of "expected waiting cost" being incurred under the current dynamics of the service process. The penalty cost per unit waiting time, \( K \), represents an opportunity cost due to waiting. A value of \( K \) may be assigned based on customer input and management opinion.

By direct integration of (2), expressions for the expected cost of customer waiting time for the Gaussian (normal), gamma, logistic and Weibull probability densities can be found (see Appendix I). Clearly, additional densities exist for defining the distribution of customer delivery times. The densities illustrated in the Appendix represent a mix of probability densities which exhibit varying shapes and tail thicknesses as defined by their coefficients of skewness and kurtosis. Software packages for determining the form of the best fitting distribution for an empirical data set of customer waiting times can be employed to aid the analyst in selecting an appropriate probability density.

3. Modeling Continuous Improvement of Customer Waiting Time

In this section we demonstrate how with knowledge of the form of \( f_X(x) \), the model defined by (2) can be used to support the continuous improvement of customer waiting time. Customer waiting time is modeled as a cost-based, time-dependent function of the waiting time variance. The financial benefit of reducing the variability in customer waiting time is demonstrated within the context of a continuous improvement program.

Let \( W(v, t) \) represent the expected cost of customer waiting as a function of the variance \( v \) of the probability density function defining customer waiting times at time period \( t \). Ideally the expected cost for customer waiting should be equal to zero. This implies that all customer waiting times for the service operation are less than the benchmark waiting duration of \( c \) times to which cost penalty \( K \) applies.

The present worth of the expected customer waiting cost stream over time horizon \( T \) provides an estimate in current dollars of costs incurred due to untimely customer waiting. Under continuous compounding (with nominal interest rate \( r \)), the present worth of the cost flow defined by (2) can be evaluated over the \( m \) equally spaced improvement cycles of length \( i = T/m \) which define time horizon \( T \) and defined as

\[ W_{NPV}(v, t) = \sum_{i=1}^{m} W(v, t) \exp(-r(i)) \]  

(3)

Initiating improvements in a service operation requires capital investment. The present worth estimate of the expected cost stream defined by (3), \( W_{NPV}(v, t) \), over the time horizon \( T \) provides a benchmark from which management can justify the capital investment required to improve customer waiting times. Thus, management should be willing to invest an amount equal to the present worth of the expected penalty cost over time horizon \( T \) in order to improve customer waiting time.

By adopting a systems perspective, management can study customer waiting time in the service operation and determine the assignable cause(s) for the untimely customer waiting. Corrective actions can be initiated, and as a result of the “learning” gained from studying customer waiting, process improvements can be implemented to remove the cause(s) of untimely waiting. Learning curve theory is widely accepted as a framework for modeling improvement in the performance of a process (see for example [26], [27]).
Under the widely-adopted log-linear learning curve model \[28\], the variance of customer waiting time is defined to be

\[
\nu(t) = \nu(0) t^d
\]

where

\[
\nu(0) = \text{initial variance of } f(x)
\]

\[t = \text{cumulative improvement period (} t = 1, 2, \ldots, m\)]

\[
\nu(t) = \text{variance of } f(x) \text{ for the } t^\text{th} \text{ improvement period}
\]

\[d = (\ln \theta)/(\ln 2)\]

\[\theta = \text{learning rate (} 0.5 < \theta < 1\).

Improvement the variance of customer waiting time as defined in (4) takes the functional form where \(\nu'(t) < 0\) and \(\nu''(t) > 0\). This form implies that when improvements are implemented the variance of customer waiting time will decrease at a diminishing rate. This functional form has intuitive appeal since it generally becomes harder to gain additional, incremental process improvements once such enhancements have already been made. This form has been widely adopted in several process improvement studies (see for example \[29\], \[30\]).

Consider the case when \(f(x)\) is defined by a normal distribution. For ease of notation, we redefine the variance of the normal distribution as \(\nu = \sigma^2\). The expected cost of customer waiting time as a function of the variance under the normal model is

\[
W(\nu) = K \left[ \sqrt{\nu} \Phi \left( \frac{c - \mu}{\sqrt{\nu}} \right) - (c - \mu) \left( 1 - \Phi \left( \frac{c - \mu}{\sqrt{\nu}} \right) \right) \right]
\]

Substituting (4) in to (5) and simplifying (see Appendix II) yields

\[
W(\nu, t) = K \left[ \sqrt{\nu(0)t^d} \exp(-k) - (c - \mu)(1 - \Phi(z)) \right]
\]

where

\[
k = \frac{(c - \mu)^2}{2\nu(0)t^d}
\]

\[
z = (c - \mu) \left( \sqrt{\nu(0)t^d} \right).
\]

Let \(T\) equal the time horizon for a continuous improvement program to reduce customer waiting time. To aid in the implementation of the continuous improvement program, time horizon \(T\) is broken down into \(m\) equally spaced milestone evaluation periods of length \(i = T/m\). Figure 1 illustrates the improvement in the expected cost of customer waiting as a result of variance reduction over time horizon of length \(T\).

Figure 1. Reduction in Expected Customer Waiting Costs.

4. Summary and Conclusions

This paper addressed one aspect of service quality by modeling customer waiting time for service from the contemporary perspective of reducing variability. A cost-based model has been presented that financially evaluates the effects of reducing customer waiting time variability. The present worth of future expected costs associated with untimely customer waiting provided by the model provide managers with a bound for attempting to justify the resources needed for investing in a continuous improvement program to improve service quality by reducing customer waiting times.

The model was demonstrated for learning-based reduction of the variance of customer waiting times. Other functional forms for modeling variance reduction can be explored in a similar manner.

The opportunity cost of management neglecting to improve delivery performance was introduced and illustrated for a selected set of
parameters. Using the model, the detrimental financial effects of managerial neglect were demonstrated.

There are several aspects of this research that could be expanded. An optimization model could be used to determine and allocate variance reduction throughout the sub stages of the service operation. Second, the opportunity cost of management neglecting to improve service quality performance could be investigated due to disruptions or decay in the learning process. Lastly, an industrial case study utilizing the model could be conducted.

5. References


Appendices

I. Expressions for the Expected Cost of Customer Waiting Time.

Normal

\[ f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \]

\[ W = K \left[ \phi\left(\frac{x-\mu}{\sigma}\right) - \left(\frac{x-\mu}{\sigma}\right) \left(1 - \Phi\left(\frac{x-\mu}{\sigma}\right)\right) \right] \]

Remarks:
\( \phi(.) \) = standard normal density function
\( \Phi(.) \) = standard normal cumulative distribution function

Gamma

\[ f_X(x) = \frac{\alpha^k x^{k-1} e^{-\alpha x}}{\Gamma(k)} \]

\[ W = K \left[ \frac{\Gamma(k+1, \alpha x)}{\alpha \Gamma(k)} - \frac{c \Gamma(k, \alpha c)}{\Gamma(k)} \right] \]

Remarks:
\( \Gamma(.) \) = incomplete gamma function
\( \Gamma(0) = \Gamma(0,0) \)

Logistic

\[ f_X(x) = \frac{a \exp(-a(x-\mu))}{\left[1 + \exp(-a(x-\mu))\right]^2} \]

\[ W = K \left[ \frac{\sqrt{3}}{\pi} \ln \left(1 + \exp\left(-\frac{a \pi}{\sqrt{3}}\right)\right) \right] \]

Remarks:
\( a = \frac{\pi}{\sigma \sqrt{3}} \)

Weibull

\[ f_X(x) = \frac{b}{d} \left(\frac{x}{d}\right)^{b-1} \exp\left[-\left(\frac{x}{d}\right)^b\right] \]

\[ W = K \left[ \frac{\Gamma\left(b+1, \frac{x}{d}\right) - \left(cm + \Gamma\left(b+1, \frac{b+1}{b}\right)\right)}{m} \right] e^c \]

Remarks:
\( m = \left(\Gamma\left(b+2, \frac{x}{d}\right) - \Gamma^2\left(b+1, \frac{b+1}{b}\right)\right)^{\frac{1}{2}} \)

II. Normality Derivation.

Equation (6) is separable into two terms. The key steps of the derivation of (6) are...
Term 1:

\[
\sqrt{v(t)} \Phi \left( \frac{c - \mu}{\sqrt{v(t)}} \right) = \sqrt{\frac{v(0)}{2\pi}} \exp \left( \frac{-(c - \mu)^2}{2v(0)} \right)
\]

and, Term 2:

\[
\Phi \left( \frac{c - \mu}{\sqrt{v(t)}} \right) = \int_{-\infty}^{(c - \mu)/\sqrt{v(0)}} \phi(x) dx = \Phi \left( \frac{c - \mu}{\sqrt{v(0)}} \right).
\]