The Exponential Model

If a random variable is exponentially distributed with mean T*a*,then the probability that the variable will exceed any specific value *t* is given by e-t/Ta, where *e* = 2.718282 is the base of the natural logarithm. This probability can be calculated by using the EXP function in Microsoft Excel.

**Example 8.5**

Suppose the time between consecutive arrivals is exponentially distributed Ta= 6 secs n🡺 *Ra* is thus 1/6 per sec (or 10 per min). The probability that the time between two consecutive arrivals will exceed 10 seconds is given by *e*–10/6= EXP(–1.667) = 0.1888.

Suppose the time required to scan one customer’s bags is exponentially distributed Tp= 5 secs. The probability that it will take no more than 3 secs is 1 – *e*–3/5= 0.451188.

If the interarrival time is exponentially distributed with mean Tp,then the number of arrivals in any interval of duration *t* turns out to have Poissondistribution with mean *t/Tp or tRa*.

Since Ca=Cp=1 then

🡺 ====



T= Ti+Tp

T=

That is, a customer’s average flow time is inversely proportional to the safety capacity, so increasing safety capacity decreases average flow time.

With multiple servers (*c* ≥ 2), exact formulas for computing the queue length and wait are also available but complicated, even with the exponential distribution assumption. These formulas are provided in Appendix II for reference and are programmed in an Microsoft Excel spreadsheet called **Performance.xls**, which was developed by Professor John O. McCain of Cornell University and can be downloaded from the Prentice Hall’s Web site at [www.prenhall.com/anupindi](http://www.prenhall.com/anupindi)

The spreadsheet computes performance characteristics of two types of processes: those with (1) finite and (2) infinite input buffer capacities. It requires specification of four inputs: the number of servers (*c*), the average arrival rate (*Ri*), the average processing rate of each server (1/*Tp*) and the buffer capacity (*K*).

**8.6 Performance Variability and Promise**

Our entire discussion of performance measurement and improvement has focused on the average queue length and the averagewaiting time. In this section, we study why considering only the average values of these performance measures may not be sufficient.

 In a service process, the average waiting time includes both customers with very long waits and those with short or no waits. Now a customer who in fact had to wait 30 minutes for service is not likely to be comforted to know that the average waiting time of all customers was only ten minutes, and in fact 20 percent of all customers did not have to wait at all! Typically, customers’ tolerance for waiting decreases with the duration of the wait. Those who have to wait for longer times are disproportionately more dissatisfied than those who have to wait for shorter times. Ideally, we would like to look at the entire probability distribution of the waiting time across all customers, not just its average. At least, we would like to know the probability that a customer may have to wait longer than a specified duration that we consider acceptable. It is important to know, therefore, what fraction of customers may experience an extraordinarily long waits, because that would highlight problematic or unacceptable service. Thus, we need to focus on the upper tail of the probability distribution of the waiting time, not just its average value.

**In a single-phase single server service process** and **exponentially distributed interarrival and processing** times, it turns out that the actual total time that a customer spends in the process is also exponentially distributed with mean *T.*

Prob(Total time in process > *t*) = e− *t/T*

is the proportion of arrivals who will have to wait for more than *t* time units. To illustrate, consider the following example.

**Example 8.11**

Suppose Walgreendrug store’spharmacy is staffed by one pharmacist, Dave. Suppose on average *Ra*= 20 customers come to the pharmacy every hour to get their prescriptions filled. Suppose Dave takes an average of *Tp* = 2.5 minutes to fill a prescription, so his processing rate is *Rp* = 24 customers per hour. If we assume exponentially distributed interarrival and processing

times, we have a single-phase, single-server exponential model of Section 8.3.2, so by Equation 8.10, the average total time in the process is given by T = 1 / (Rp – Ri) = 1/ (24 – 20) = 0.25 hours, or 15 minutes. The fraction of customers who will spend more than t = 15 minutes is e– t/T = e– 15/15 = e– 1 = 0.3679, so about 37 percent of customers will need more than 15 minutes to have their prescriptions filled. More seriously, the fraction of customers who will spend more than t = 45 minutes will be e– t/T = e– 45/15 = e– 3 = 0.0498, so about 5 percent of customers will spend more than 45 minutes in the drug store! For these 5 percent of customers, it is cold comfort to know that an average customer spends only 15 minutes in the system. These 5 percent of customers will also be the ones who will complain most bitterly about the delay and affect Walgreen’s business. Thus, in addition to the average value, the 95th percentile of the distribution of the actual total time in process provides important information about the process performance. It qualifies the average total time of 15 minutes with a caution that 5 percent of customers will experience delays of 40 minutes or more.

Since a large fraction of customers may experience delays longer than the average *T,* the process manager may not wish to announce *T* as what most customers will experience (even though that is what a typical customer is expectedto spend in the process). As a **service promise**, we may wish to quote that customers will be served within some conservatively high duration, say *Td*, so that we will meet that promise most of the time. This promised duration *Td* refers to a promised time period within which the product or service will be delivered with a high probability*.* For example, if Dave promises to fill prescriptions within half an hour, we have *Td =* 30, even though average is *T* = 15 minutes*.* In that case, the proportion of customers who will have to wait more than the promised time is

e– t/T = e– 30/15 = e– 2 = 0.1353

Thus, Dave will not be able to keep his promise on 13.53% of customers, that is, he will deliver on promise 86.47% of the time.

Analogous to Chapter 7, we may define the proportion of customers that will be served during the promised duration *Td* ascustomer service level*SL*. Thus, by promising 30 minute turnaround time, Dave is providing 86.47% customer service.

Conversely, we may set a service level (*SL*), and derive the corresponding due date *Td* to promise such that

 P(Total time in process > *Td* ) = 1 – *SL* **(Equation 8.13)**

e– Td/T = 1– *SL*

*Td* = – *T* *ln* (1–*SL*)

where *ln* denotes the natural logarithm with *.*

We can then define safety time (analogous to safety inventory and safety capacity) as

 *Ts* = *Td* – *T 🡺 Td* = *T* + *Ts*

Thus, **safety time** *is the time margin that we should allow over and above the expected time to deliver service in order to ensure that we will be able to meet the promised duration with high probability.* Clearly, the larger the safety time, the greater the probability of fulfilling the promise.

**Example 8.12.** Suppose Dave wishes to promise time Td such that he will be able to keep his promise on 90 percent of the customers. Thus he is choosing SL = 0.90, when the average time required is T = 15 minutes. The necessary due date will be

Td = – (15) ln(0.10) = 34.54 minutes.

Therefore, Dave should announce that 90 percent of his customers will get their prescriptions filled within 35 minutes of arrival. Even though he expects to fill prescriptions within 15 minutes of submission on average, he is allowing additional 20 minutes of safety time to ensure serving 90 percent of his customers within the promised time.