CHAPTER 2

Renewal-Reward Processes

2.0 INTRODUCTION

The renewal-reward model is an extremely useful tool in the analysis of applied probability models for inventory, queueing and reliability applications, among others. Many stochastic processes are regenerative; that is, they regenerate themselves from time to time so that the behaviour of the process after the regeneration epoch is a probabilistic replica of the behaviour of the process starting at time zero. The time interval between two regeneration epochs is called a cycle. The sequence of regeneration cycles constitutes a so-called renewal process. The long-run behaviour of a regenerative stochastic process on which a reward structure is imposed can be studied in terms of the behaviour of the process during a single regeneration cycle. The simple and intuitively appealing renewal-reward model has numerous applications.

In Section 2.1 we first discuss some elementary results from renewal theory. A more detailed treatment of renewal theory will be given in Chapter 8. Section 2.2 deals with the renewal-reward model. It shows how to calculate long-run averages such as the long-run average reward per time unit and the long-run fraction of time the system spends in a given set of states. Illustrative examples will be given. Section 2.3 discusses the formula of Little. This formula is a kind of law of nature and relates among others the average queue size to the average waiting time in queueing systems. Another fundamental result that is frequently used in queueing and inventory applications is the property that Poisson arrivals see time averages (PASTA). This result is discussed in some detail in Section 2.4. The PASTA property is used in Section 2.5 to obtain the famous Pollaczek–Khintchine formula from queueing theory. The renewal-reward model is used in Section 2.6 to obtain a generalization of the Pollaczek–Khintchine formula in the framework of a controlled queue. Section 2.7 shows how renewal theory and an up- and down-crossing argument can be combined to derive a relation between time-average and customer-average probabilities in queues.

A First Course in Stochastic Models H.C. Tijms
© 2003 John Wiley & Sons, Ltd. ISBNs: 0-471-49880-7 (HB); 0-471-49881-5 (PB)

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

2.1 RENEWAL THEORY

As a generalization of the Poisson process, renewal theory concerns the study of stochastic processes counting the number of events that take place as a function of time. Here the interoccurrence times between successive events are independent and identically distributed random variables. For instance, the events could be the arrival of customers to a waiting line or the successive replacements of light bulbs. Although renewal theory originated from the analysis of replacement problems for components such as light bulbs, the theory has many applications to quite a wide range of practical probability problems. In inventory, queueing and reliability problems, the analysis is often based on an appropriate identification of embedded renewal processes for the specific problem considered. For example, in a queueing process the embedded events could be the arrival of customers who find the system empty, or in an inventory process the embedded events could be the replenishment of stock when the inventory position drops to the reorder point or below it.

Formally, let $X_1, X_2, \ldots$ be a sequence of non-negative, independent random variables having a common probability distribution function

$$F(x) = P\{X_k \leq x\}, \quad x \geq 0$$

for $k = 1, 2, \ldots$. Letting $\mu_1 = E(X_k)$, it is assumed that

$$0 < \mu_1 < \infty.$$

The random variable X_n denotes the interoccurrence time between the $(n-1)$th and nth event in some specific probability problem. Define

$$S_0 = 0 \quad \text{and} \quad S_n = \sum_{i=1}^{n} X_i, \quad n = 1, 2, \ldots .$$

Then S_n is the epoch at which the nth event occurs. For each $t \geq 0$, let

$$N(t) = \text{the largest integer } n \geq 0 \text{ for which } S_n \leq t.$$

Then the random variable $N(t)$ represents the number of events up to time t.

Definition 2.1.1 *The counting process $\{N(t), t \geq 0\}$ is called the renewal process generated by the interoccurrence times $X_1, X_2, \ldots$.*

It is said that a renewal occurs at time t if $S_n = t$ for some n. For each $t \geq 0$, the number of renewals up to time t is finite with probability 1. This is an immediate consequence of the strong law of large numbers stating that $S_n/n \to E(X_1)$ with probability 1 as $n \to \infty$ and thus $S_n \leq t$ only for finitely many n. The Poisson process is a special case of a renewal process. Here we give some other examples of a renewal process.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Example 2.1.1 A replacement problem

Suppose we have an infinite supply of electric bulbs, where the burning times of the bulbs are independent and identically distributed random variables. If the bulb in use fails, it is immediately replaced by a new bulb. Let X_i be the burning time of the ith bulb, $i = 1, 2, \ldots$. Then $N(t)$ is the total number of bulbs to be replaced up to time t.

Example 2.1.2 An inventory problem

Consider a periodic-review inventory system for which the demands for a single product in the successive weeks $t = 1, 2, \ldots$ are independent random variables having a common *continuous* distribution. Let X_i be the demand in the ith week, $i = 1, 2, \ldots$. Then $1 + N(u)$ is the number of weeks until depletion of the current stock u.

2.1.1 The Renewal Function

An important role in renewal theory is played by the *renewal function* $M(t)$ which is defined by

$$M(t) = E[N(t)], \quad t \geq 0. \tag{2.1.1}$$

For $n = 1, 2, \ldots$, define the probability distribution function

$$F_n(t) = P\{S_n \leq t\}, \quad t \geq 0.$$

Note that $F_1(t) = F(t)$. A basic relation is

$$N(t) \geq n \quad \text{if and only if} \quad S_n \leq t. \tag{2.1.2}$$

This relation implies that

$$P\{N(t) \geq n\} = F_n(t), \quad n = 1, 2, \ldots . \tag{2.1.3}$$

Lemma 2.1.1 *For any $t \geq 0$,*

$$M(t) = \sum_{n=1}^{\infty} F_n(t). \tag{2.1.4}$$

Proof Since for any non-negative integer-valued random variable N,

$$E(N) = \sum_{k=0}^{\infty} P\{N > k\} = \sum_{n=1}^{\infty} P\{N \geq n\},$$

the relation (2.1.4) is an immediate consequence of (2.1.3).

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

In Exercise 2.4 the reader is asked to prove that $M(t) < \infty$ for all $t \geq 0$. In Chapter 8 we will discuss how to compute the renewal function $M(t)$ in general. The infinite series (2.1.4) is in general not useful for computational purposes. An exception is the case in which the interoccurrence times $X_1, X_2, \ldots$ have a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$. Then the sum $X_1 + \cdots + X_n$ has a gamma distribution with shape parameter $n\alpha$ and scale parameter λ. In this case $F_n(t)$ is the so-called incomplete gamma integral for which efficient numerical procedures are available; see Appendix B. Let us explain this in more detail for the case that α is a positive integer r so that the interoccurrence times $X_1, X_2, \ldots$ have an Erlang (r, λ) distribution with scale parameter λ. Then $F_n(t)$ becomes the Erlang (nr, λ) distribution function

$$F_n(t) = 1 - \sum_{k=0}^{nr-1} e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad t \geq 0$$

and thus

$$M(t) = \sum_{n=1}^{\infty} \left[1 - \sum_{k=0}^{nr-1} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \right], \quad t \geq 0. \tag{2.1.5}$$

In this particular case $M(t)$ can be efficiently computed from a rapidly converging series. For the special case that the interoccurrence times are exponentially distributed ($r = 1$), the expression (2.1.5) reduces to the explicit formula

$$M(t) = \lambda t, \quad t \geq 0.$$

This finding is in agreement with earlier results for the Poisson process.

Remark 2.1.1 The phase method

A very useful interpretation of the renewal process $\{N(t)\}$ can be given when the interoccurrence times $X_1, X_2, \ldots$ have an Erlang distribution. Imagine that tokens arrive according to a Poisson process with rate λ and that the arrival of each rth token triggers the occurrence of an event. Then the events occur according to a renewal process in which the interoccurrence times have an Erlang (r, λ) distribution with scale parameter λ. The explanation is that the sum of r independent, exponentially distributed random variables with the same scale parameter λ has an Erlang (r, λ) distribution. The phase method enables us to give a tractable expression of the probability distribution of $N(t)$ when the interoccurrence times have an Erlang (r, λ) distribution. In this case $P\{N(t) \geq n\}$ is equal to the probability that nr or more arrivals occur in a Poisson arrival process with rate λ. You are asked to work out the equivalence in Exercise 2.5.

Asymptotic expansion

A very useful asymptotic expansion for the renewal function $M(t)$ can be given under a weak regularity condition on the interoccurrence times. This condition

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

will be formulated in Section 8.2. For the moment it is sufficient to assume that the interoccurrence times have a positive density on some interval. Further it is assumed that $\mu_2 = E(X_1^2)$ is finite. Then it will be shown in Theorem 8.2.3 that

$$\lim_{t\to\infty}\left[M(t) - \frac{t}{\mu_1}\right] = \frac{\mu_2}{2\mu_1^2} - 1. \tag{2.1.6}$$

The approximation

$$M(t) \approx \frac{t}{\mu_1} + \frac{\mu_2}{2\mu_1^2} - 1 \quad \text{for } t \text{ large}$$

is practically useful for already moderate values of t provided that the squared coefficient of variation of the interoccurrence times is not too large and not too close to zero.

2.1.2 The Excess Variable

In many practical probability problems an important quantity is the random variable γ_t defined as the time elapsed from epoch t until the next renewal after epoch t. More precisely, γ_t is defined as

$$\gamma_t = S_{N(t)+1} - t;$$

see also Figure 2.1.1 in which a renewal epoch is denoted by $\times$. Note that $S_{N(t)+1}$ is the epoch of the first renewal that occurs after time t. The random variable γ_t is called the *excess* or *residual life* at time t. For the replacement problem of Example 2.1.1 the random variable γ_t denotes the residual lifetime of the light bulb in use at time t.

Lemma 2.1.2 *For any $t \geq 0$,*

$$E(\gamma_t) = \mu_1[1 + M(t)] - t. \tag{2.1.7}$$

Proof Fix $t \geq 0$. To prove (2.1.7), we apply Wald's equation from Appendix A. To do so, note that $N(t) \leq n-1$ if and only if $X_1 + \cdots + X_n > t$. Hence the event $\{N(t) + 1 = n\}$ depends only on $X_1, \ldots, X_n$ and is thus independent of $X_{n+1}, X_{n+2}, \ldots$. Hence

$$E\left[\sum_{k=1}^{N(t)+1} X_k\right] = E(X_1)E[N(t)+1],$$

which gives (2.1.7).

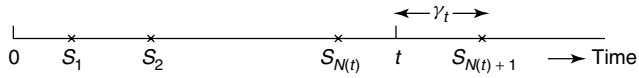

Figure 2.1.1 The excess life

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

In Corollary 8.2.4 it will be shown that

$$\lim_{t\to\infty} E(\gamma_t) = \frac{\mu_2}{2\mu_1} \quad \text{and} \quad \lim_{t\to\infty} E(\gamma_t^2) = \frac{\mu_3}{3\mu_1} \tag{2.1.8}$$

with $\mu_k = E(X_1^k)$ for $k = 1, 2, 3$, provided that the interoccurrence times have a positive density on some interval. An illustration of the usefulness of the concept of excess variable is provided by the next example.

Example 2.1.3 The average order size in an (s, S) ***inventory system***

Suppose a periodic-review inventory system for which the demands $X_1, X_2, \ldots$ for a single product in the successive weeks $1, 2, \ldots$ are independent random variables having a common probability density $f(x)$ with finite mean α and finite standard deviation σ. Any demand exceeding the current inventory is backlogged until inventory becomes available by the arrival of a replenishment order. The inventory position is reviewed at the beginning of each week and is controlled by an (s, S) rule with $0 \le s < S$. Under this control rule, a replenishment order of size $S - x$ is placed when the review reveals that the inventory level x is below the reorder point s; otherwise, no ordering is done. We assume instantaneous delivery of every replenishment order.

We are interested in the average order size. Since the inventory process starts from scratch each time the inventory position is ordered up to level S, the operating characteristics can be calculated by using a renewal model in which the weekly demand sizes $X_1, X_2, \ldots$ represent the interoccurrence times of renewals. The number of weeks between two consecutive orderings equals the number of weeks needed for a cumulative demand larger than $S - s$. The order size is the sum of $S - s$ and the undershoot of the reorder point s at the epoch of ordering (see Figure 2.1.2 in which a renewal occurrence is denoted by an $\times$). Denote by $\{N(t)\}$ the renewal process associated with the weekly demands $X_1, X_2, \ldots$. Then the number of weeks needed for a cumulative demand exceeding $S - s$ is given by $1 + N(S - s)$. The undershoot of the reorder point s is just the excess life γ_{S-s} of the renewal process. Hence

$$E[\text{order size}] = S - s + E(\gamma_{S-s}).$$

From (2.1.8) it follows that the average order size can be approximated by

$$E[\text{order size}] \approx S - s + \frac{\sigma^2 + \alpha^2}{2\alpha}$$

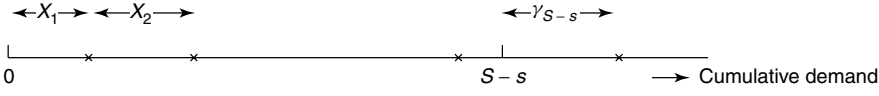

Figure 2.1.2 The inventory process modelled as a renewal process

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

provided that $S - s$ is sufficiently large compared with E(weekly demand). In practice this is a useful approximation for $S-s > \alpha$ when the weekly demand is not highly variable and has a squared coefficient of variation between 0.2 and 1 (say).

Another illustration of the importance of the excess variable is given by the famous waiting-time paradox.

Example 2.1.4 The waiting-time paradox

We have all experienced long waits at a bus stop when buses depart irregularly and we arrive at the bus stop at random. A theoretical explanation of this phenomenon is provided by the expression for $\lim_{t\to\infty} E(\gamma_t)$. Therefore it is convenient to rewrite (2.1.8) as

$$\lim_{t\to\infty} E(\gamma_t) = \frac{1}{2}(1 + c_X^2)\mu_1, \tag{2.1.9}$$

where

$$c_X^2 = \frac{\sigma^2(X_1)}{E^2(X_1)}$$

is the squared coefficient of variation of the interdeparture times $X_1, X_2, \ldots$. The equivalent expression (2.1.9) follows from (2.1.8) by noting that

$$1 + c_X^2 = 1 + \frac{\mu_2 - \mu_1^2}{\mu_1^2} = \frac{\mu_2}{\mu_1^2}. \tag{2.1.10}$$

The representation (2.1.9) makes clear that

$$\lim_{t\to\infty} E(\gamma_t) = \begin{cases} < \mu_1 & \text{if } c_X^2 < 1, \\ > \mu_1 & \text{if } c_X^2 > 1. \end{cases}$$

Thus the mean waiting time for the next bus depends on the regularity of the bus service and increases with the coefficient of variation of the interdeparture times. If we arrive at the bus stop at random, then for highly irregular service ($c_X^2 > 1$) the mean waiting time for the next bus is even larger than the mean interdeparture time. This surprising result is sometimes called the *waiting-time paradox*. A heuristic explanation is that it is more likely to hit a long interdeparture time than a short one when arriving at the bus stop at random. To illustrate this, consider the extreme situation in which the interdeparture time is 0 minutes with probability 9/10 and is 10 minutes with probability 1/10. Then the mean interdeparture time is 1 minute, but your mean waiting time for the next bus is 5 minutes when you arrive at the bus stop at random.

2.2 RENEWAL-REWARD PROCESSES

A powerful tool in the analysis of numerous applied probability models is the renewal-reward model. This model is also very useful for theoretical purposes. In

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Chapters 3 and 4, ergodic theorems for Markov chains will be proved by using the renewal-reward theorem. The renewal-reward model is a simple and intuitively appealing model that deals with a so-called regenerative process on which a cost or reward structure is imposed. Many stochastic processes have the property of regenerating themselves at certain points in time so that the behaviour of the process after the regeneration epoch is a probabilistic replica of the behaviour starting at time zero and is independent of the behaviour before the regeneration epoch.

A formal definition of a regenerative process is as follows.

Definition 2.2.1 *A stochastic process* $\{X(t), t \in T\}$ *with time-index set* T *is said to be regenerative if there exists a (random) epoch* S_1 *such that:*

(a) $\{X(t + S_1), t \in T\}$ *is independent of* $\{X(t), 0 \leq t < S_1\}$,

(b) $\{X(t + S_1), t \in T\}$ *has the same distribution as* $\{X(t), t \in T\}$.

It is assumed that the index set T is either the interval $T = [0, \infty)$ or the countable set $T = \{0, 1, \dots\}$. In the former case we have a continuous-time regenerative process and in the other case a discrete-time regenerative process. The state space of the process $\{X(t)\}$ is assumed to be a subset of some Euclidean space.

The existence of the regeneration epoch S_1 implies the existence of further regeneration epochs $S_2, S_3, \dots$ having the same property as S_1. Intuitively speaking, a regenerative process can be split into independent and identically distributed renewal cycles. A *cycle* is defined as the time interval between two consecutive regeneration epochs. Examples of regenerative processes are:

(i) The continuous-time process $\{X(t), t \geq 0\}$ with $X(t)$ denoting the number of customers present at time t in a single-server queue in which the customers arrive according to a renewal process and the service times are independent and identically distributed random variables. It is assumed that at epoch 0 a customer arrives at an empty system. The regeneration epochs $S_1, S_2, \dots$ are the epochs at which an arriving customer finds the system empty.

(ii) The discrete-time process $\{I_n, n = 0, 1, \dots\}$ with I_n denoting the inventory level at the beginning of the nth week in the (s, S) inventory model dealt with in Example 2.1.3. Assume that the inventory level equals S at epoch 0. The regeneration epochs are the beginnings of the weeks in which the inventory level is ordered up to the level S.

Let us define the random variables $C_n = S_n - S_{n-1}$, $n = 1, 2, \dots$, where $S_0 = 0$ by convention. The random variables $C_1, C_2, \dots$ are independent and identically distributed. In fact the sequence $\{C_1, C_2, \dots\}$ underlies a renewal process in which the events are the occurrences of the regeneration epochs. Hence we can interpret C_n as

$$C_n = \text{the length of the } n\text{th renewal cycle}, \quad n = 1, 2, \dots .$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Note that the cycle length C_n assumes values from the index set T. In the following it is assumed that

$$0 < E(C_1) < \infty.$$

In many practical situations a reward structure is imposed on the regenerative process $\{X(t), t \in T\}$. The reward structure usually consists of reward rates that are earned continuously over time and lump rewards that are only earned at certain state transitions. Let

$$R_n = \text{the total reward earned in the } n\text{th renewal cycle}, \quad n = 1, 2, \ldots .$$

It is assumed that $R_1, R_2, \ldots$ are independent and identically distributed random variables. In applications R_n typically depends on C_n. In case R_n can take on both positive and negative values, it is assumed that $E(|R_1|) < \infty$. Let

$$R(t) = \text{the cumulative reward earned up to time } t.$$

The process $\{R(t), t \geq 0\}$ is called a *renewal-reward process.* We are now ready to prove a theorem of utmost importance.

Theorem 2.2.1 (renewal-reward theorem)

$$\lim_{t\to\infty} \frac{R(t)}{t} = \frac{E(R_1)}{E(C_1)} \quad \textit{with probability } 1.$$

In other words, for almost any realization of the process, the long-run average reward per time unit is equal to the expected reward earned during one cycle divided by the expected length of one cycle.

To prove this theorem we first establish the following lemma.

Lemma 2.2.2 *For any $t \geq 0$, let $N(t)$ be the number of cycles completed up to time t. Then*

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{E(C_1)} \quad \textit{with probability } 1.$$

Proof By the definition of $N(t)$, we have

$$C_1 + \cdots + C_{N(t)} \leq t < C_1 + \cdots + C_{N(t)+1}.$$

Since $P\{C_1 + \cdots + C_n < \infty\} = 1$ for all $n \geq 1$, it is not difficult to verify that

$$\lim_{t\to\infty} N(t) = \infty \quad \text{with probability 1.}$$

The above inequality gives

$$\frac{C_1 + \cdots + C_{N(t)}}{N(t)} \leq \frac{t}{N(t)} < \frac{C_1 + \cdots + C_{N(t)+1}}{N(t)+1} \frac{N(t)+1}{N(t)}.$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

By the strong law of large numbers for a sequence of independent and identically distributed random variables, we have

$$\lim_{n\to\infty} \frac{C_1 + \cdots + C_n}{n} = E(C_1) \quad \text{with probability 1.}$$

Hence, by letting $t \to \infty$ in the above inequality, the desired result follows.

Lemma 2.2.2 is also valid when $E(C_1) = \infty$ provided that $P\{C_1 < \infty\} = 1$. The reason is that the strong law of large numbers for a sequence $\{C_n\}$ of *non-negative* random variables does not require that $E(C_1) < \infty$. Next we prove Theorem 2.2.1.

Proof of Theorem 2.2.1 For ease, let us first assume that the rewards are non-negative. Then, for any $t > 0$,

$$\sum_{i=1}^{N(t)} R_i \le R(t) \le \sum_{i=1}^{N(t)+1} R_i.$$

This gives

$$\frac{\sum_{i=1}^{N(t)} R_i}{N(t)} \times \frac{N(t)}{t} \le \frac{R(t)}{t} \le \frac{\sum_{i=1}^{N(t)+1} R_i}{N(t)+1} \times \frac{N(t)+1}{t}.$$

By the strong law of large numbers for the sequence $\{R_n\}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} R_i = E(R_1) \quad \text{with probability 1.}$$

As pointed out in the proof of Lemma 2.2.2, $N(t) \to \infty$ with probability 1 as $t \to \infty$. Letting $t \to \infty$ in the above inequality and using Lemma 2.2.2, the desired result next follows for the case that the rewards are non-negative. If the rewards can assume both positive and negative values, then the theorem is proved by treating the positive and negative parts of the rewards separately. We omit the details.

In a natural way Theorem 2.2.1 relates the behaviour of the renewal-reward process over time to the behaviour of the process over a single renewal cycle. It is noteworthy that the outcome of the long-run average actual reward per time unit can be predicted with probability 1. If we are going to run the process over an infinitely long period of time, then we can say beforehand that in the long run the average *actual* reward per time unit will be equal to the constant $E(R_1)/E(C_1)$ with probability 1. This is a much stronger and more useful statement than the statement that the long-run *expected* average reward per time unit equals $E(R_1)/E(C_1)$ (it indeed holds that $\lim_{t\to\infty} E[R(t)]/t = E(R_1)/E(C_1)$; this *expected-value version* of the renewal-reward theorem is a direct consequence of Theorem 2.2.1 when $R(t)/t$ is bounded in t but otherwise requires a hard proof). Also it is noted that for the case of non-negative rewards R_n the renewal-reward theorem is also valid when $E(R_1) = \infty$ (the assumption $E(C_1) < \infty$ cannot be dropped for Theorem 2.2.1).

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Example 2.2.1 Alternating up- and downtimes

Suppose a machine is alternately up and down. Denote by $U_1, U_2, \ldots$ the lengths of the successive up-periods and by $D_1, D_2, \ldots$ the lengths of the successive down-periods. It is assumed that both $\{U_n\}$ and $\{D_n\}$ are sequences of independent and identically distributed random variables with finite positive expectations. The sequences $\{U_n\}$ and $\{D_n\}$ are not required to be independent of each other. Assume that an up-period starts at epoch 0. What is the long-run fraction of time the machine is down? The answer is

$$\text{the long-run fraction of time the machine is down} = \frac{E(D_1)}{E(U_1) + E(D_1)} \quad \text{with probability 1.} \tag{2.2.1}$$

To verify this, define the continuous-time stochastic process $\{X(t), t \geq 0\}$ by

$$X(t) = \begin{cases} 1 & \text{if the machine is up at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

The process $\{X(t)\}$ is a regenerative process. The epochs at which an up-period starts can be taken as regeneration epochs. The long-run fraction of time the machine is down can be interpreted as a long-run average cost per time unit by assuming that a cost at rate 1 is incurred while the machine is down and a cost at rate 0 otherwise. A regeneration cycle consists of an up-period and a down-period. Hence

$$E(\text{length of one cycle}) = E(U_1 + D_1)$$

and

$$E(\text{cost incurred during one cycle}) = E(D_1).$$

By applying the renewal-reward theorem, it follows that the long-run average cost per time unit equals $E(D_1)/[E(U_1) + E(D_1)]$, proving the result (2.2.1).

The intermediate step of interpreting the long-run fraction of time that the process is in a certain state as a long-run average cost (reward) per time unit is very helpful in many situations.

Limit theorems for regenerative processes

An important application of the renewal-reward theorem is the characterization of the long-run fraction of time a regenerative process $\{X(t), t \in T\}$ spends in some given set B of states. For the set B of states, define for any $t \in T$ the indicator variable

$$I_B(t) = \begin{cases} 1 & \text{if } X(t) \in B, \\ 0 & \text{if } X(t) \notin B. \end{cases}$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Also, define the random variable

T_B = the amount of time the process spends in the set B of states during one cycle.

Note that $T_B = \int_0^{S_1} I_B(u)\,du$ for a continuous-time process $\{X(t)\}$; otherwise, T_B equals the number of indices $0 \leq k < S_1$ with $X(k) \in B$. The following theorem is an immediate consequence of the renewal-reward theorem.

Theorem 2.2.3 *For the regenerative process $\{X(t)\}$ it holds that the long-run fraction of time the process spends in the set B of states is $E(T_B)/E(C_1)$ with probability 1.*

That is,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t I_B(u)\,du = \frac{E(T_B)}{E(C_1)} \quad \text{with probability 1}$$

for a continuous-time process $\{X(t)\}$ and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n} I_B(k) = \frac{E(T_B)}{E(C_1)} \quad \text{with probability 1}$$

for a discrete-time process $\{X(n)\}$.

Proof The long-run fraction of time the process $\{X(t)\}$ spends in the set B of states can be interpreted as a long-run average reward per time unit by assuming that a reward at rate 1 is earned while the process is in the set B and a reward at rate 0 is earned otherwise. Then

$$E(\text{reward earned during one cycle}) = E(T_B).$$

The desired result next follows by applying the renewal-reward theorem.

Since $E(I_B(t)) = P\{X(t) \in B\}$, we have as consequence of Theorem 2.2.3 and the bounded convergence theorem that, for a continuous-time process,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t P\{X(u) \in B\}\,du = \frac{E(T_B)}{E(C_1)}.$$

Note that $(1/t)\int_0^t P\{X(u) \in B\}\,du$ can be interpreted as the probability that an outside observer arriving at a randomly chosen point in $(0, t)$ finds the process in the set B.

In many situations the ratio $E(T_B)/E(C_1)$ could be interpreted both as the long-run fraction of time the process $\{X(t)\}$ spends in the set B of states and as the probability of finding the process in the set B when the process has reached statistical equilibrium. This raises the question whether $\lim_{t\to\infty} P\{X(t) \in B\}$ always exists. This ordinary limit need not always exist. A counterexample is provided by

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

periodic discrete-time Markov chains; see Chapter 3. For completeness we state the following theorem.

Theorem 2.2.4 *For the regenerative process* $\{X(t), t \in T\}$,

$$\lim_{t\to\infty} P\{X(t) \in B\} = \frac{E(T_B)}{E(C_1)}$$

provided that the probability distribution of the cycle length has a continuous part in the continuous-time case and is aperiodic in the discrete-time case.

A distribution function is said to have a continuous part if it has a positive density on some interval. A discrete distribution $\{a_j, j = 0, 1, \ldots\}$ is said to be aperiodic if the greatest common divisor of the indices $j \geq 1$ for which $a_j > 0$ is equal to 1. The proof of Theorem 2.2.4 requires deep mathematics and is beyond the scope of this book. The interested reader is referred to Miller (1972). It is remarkable that the proof of Theorem 2.2.3 for the time-average limit $\lim_{t\to\infty} (1/t) \int_0^t I_B(u)\, du$ is much simpler than the proof of Theorem 2.2.4 for the ordinary limit $\lim_{t\to\infty} P\{X(t) \in B\}$. This is all the more striking when we take into account that the time-average limit is in general much more useful for practical purposes than the ordinary limit. Another advantage of the time-average limit is that it is easier to understand than the ordinary limit. In interpreting the ordinary limit one should be quite careful. The ordinary limit represents the probability that an outside person will find the process in some state of the set B when inspecting the process at an arbitrary point in time after the process has been in operation for a very long time. It is essential for this interpretation that the outside person has *no information* about the past of the process when inspecting the process. How much more concrete is the interpretation of the time-average limit as the long-run fraction of time the process will spend in the set B of states!

To illustrate Theorem 2.2.4, consider again Example 2.2.1. In this example we analysed the long-run average behaviour of the regenerative process $\{X(t)\}$, where $X(t) = 1$ if the machine is up at time t and $X(t) = 0$ otherwise. It was shown that the long-run fraction of time the machine is down equals $E(D)/[E(U) + E(D)]$, where the random variables U and D denote the lengths of an up-period and a down-period. This result does not require any assumption about the shapes of the probability distributions of U and D. However, some assumption is needed in order to conclude that

$$\lim_{t\to\infty} P\{\text{the system is down at time } t\} = \frac{E(D)}{E(U) + E(D)}. \tag{2.2.2}$$

It is sufficient to assume that the distribution function of the length of an up-period has a positive density on some interval.

We state without proof a central limit theorem for the renewal-reward process.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Theorem 2.2.5 *Assume that* $R(t) \geq 0$ *with* $E(C_1^2) < \infty$ *and* $E(R_1^2) < \infty$. *Then*

$$\lim_{t\to\infty} P\left\{\frac{R(t) - gt}{\nu\sqrt{t/\mu_1}} \leq x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\, dy, \quad x \geq 0,$$

where $\mu_1 = E(C_1)$, $\mu_2 = E(C_1^2)$, $g = E(R_1)/E(C_1)$ *and* $\nu^2 = E(R_1 - gC_1)^2$.

A proof of this theorem can be found in Wolff (1989). In applying this theorem, the difficulty is usually to find the constant ν. In specific applications one might use simulation to find ν. As a special case, Theorem 2.2.5 includes a central limit theorem for the renewal process $\{N(t)\}$ studied in Section 2.1. Taking the rewards R_n equal to 1 it follows that the renewal process $\{N(t)\}$ is asymptotically $N(t/\mu_1, \sigma^2 t/\mu_1^3)$ distributed with $\sigma^2 = \mu_2 - \mu_1^2$.

Next we give two illustrative examples of the renewal-reward model.

Example 2.2.2 A stochastic clearing system

In a communication system messages requiring transmission arrive according to a Poisson process with rate λ. The messages are temporarily stored in a buffer having ample capacity. Every T time units, the buffer is cleared from all messages present. The buffer is empty at time $t = 0$. A fixed cost of $K > 0$ is incurred for each clearing of the buffer. Also, for each message there is a holding cost of $h > 0$ for each time unit the message has to wait in the buffer. What is the value of T for which the long-run average cost per time unit is minimal?

We first derive an expression for the average cost per time unit for a given value of the control parameter T. To do so, observe that the stochastic process describing the number of messages in the system regenerates itself each time the buffer is cleared from all messages present. This fact uses the lack of memory of the Poisson arrival process so that at any clearing epoch it is not relevant how long ago the last message arrived. Taking a cycle as the time interval between two successive clearings of the buffer, we have

$$\text{the expected length of one cycle } = T.$$

To specify the expected cost incurred during one cycle, we need an expression for the total waiting time of all messages arriving during one cycle. It was shown in Example 1.1.4 that

$$E[\text{total waiting time in } (0, T)] = \frac{1}{2}\lambda T^2.$$

This gives

$$E[\text{cost incurred during one cycle}] = K + \frac{1}{2}h\lambda T^2.$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Hence, by the renewal-reward theorem,

$$\text{the long-run average cost per time unit } = \frac{1}{T}\left(K + \frac{1}{2}h\lambda T^2\right)$$

with probability 1. When $K = 0$ and $h = 1$, the system incurs a cost at rate j whenever there are j messages in the buffer, in which case the average cost per time unit gives the average number of messages in the buffer. Hence

$$\text{the long-run average number of messages in the buffer} = \frac{1}{2}\lambda T.$$

Putting the derivative of the cost function equal to 0, it follows that the long-run average cost is minimal for

$$T^* = \sqrt{\frac{2K}{h\lambda}}.$$

Example 2.2.3 A reliability system with redundancies

An electronic system consists of a number of independent and identical components hooked up in parallel. The lifetime of each component has an exponential distribution with mean $1/\mu$. The system is operative only if m or more components are operating. The non-failed units remain in operation when the system as a whole is in a non-operative state. The system availability is increased by periodic maintenance and by putting r redundant components into operation in addition to the minimum number m of components required. Under the periodic maintenance the system is inspected every T time units, where at inspection the failed components are repaired. The repair time is negligible and each repaired component is again as good as new. The periodic inspections provide the only repair opportunities. The following costs are involved. For each component there is a depreciation cost of $I > 0$ per time unit. A fixed cost of $K > 0$ is made for each inspection and there is a repair cost of $R > 0$ for each failed component. How can we choose the number r of redundant components and the time T between two consecutive inspections such that the long-run average cost per time unit is minimal subject to the requirement that the probability of system failure between two inspections is no more than a prespecified value α?

We first derive the performance measures for given values of the parameters r and T. The stochastic process describing the number of operating components is regenerative. Using the lack of memory of the exponential lifetimes of the components, it follows that the process regenerates itself after each inspection. Taking a cycle as the time interval between two inspections, we have

$$E(\text{length of one cycle}) = T.$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Further, using the fact that a given component fails within a time T with probability $1 - e^{-\mu T}$, it follows that

$$P\{\text{the system as a whole fails between two inspections}\}$$
$$= \sum_{k=r+1}^{m+r} \binom{m+r}{k} (1 - e^{-\mu T})^k e^{-\mu T(m+r-k)}$$

and

$$E(\text{number of components that fail between two inspections})$$
$$= (m + r)(1 - e^{-\mu T}).$$

Hence

$$E(\text{total costs in one cycle}) = (m + r)I \times T + K + (m + r)(1 - e^{-\mu T})R.$$

This gives

$$\text{the long-run average cost per time unit}$$
$$= \frac{1}{T}[(m + r)I \times T + K + (m + r)(1 - e^{-\mu T})R]$$

with probability 1. The optimal values of the parameters r and T are found from the following minimization problem:

$$\text{Minimize } \frac{1}{T}[(m + r)I \times T + K + (m + r)(1 - e^{-\mu T})R]$$
$$\text{subject to } \sum_{k=r+1}^{m+r} \binom{m+r}{k} (1 - e^{-\mu T})^k e^{-\mu T(m+r-k)} \leq \alpha.$$

Using the Lagrange method this problem can be numerically solved.

*Rare events**

In many applied probability problems one has to study rare events. For example, a rare event could be a system failure in reliability applications or buffer overflow in finite-buffer telecommunication problems. Under general conditions it holds that the time until the first occurrence of a rare event is approximately *exponentially* distributed. Loosely formulated, the following result holds. Let $\{X(t)\}$ be a regenerative process having a set B of (bad) states such that the probability q that the process visits the set B during a given cycle is very small. Denote by the random variable U the time until the process visits the set B for the first time. Assuming that the cycle length has a finite and positive mean $E(T)$, it holds that $P\{U > t\} \approx e^{-tq/E(T)}$ for $t \geq 0$; see Keilson (1979) or Solovyez (1971) for

*This section may be skipped at first reading.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

a proof. The result that the time until the first occurrence of a rare event in a regenerative process is approximately exponentially distributed is very useful. It gives not only quantitative insight, but it also implies that the computation of the mean of the first-passage time suffices to get the whole distribution.

In the next example we obtain the above result by elementary arguments.

Example 2.2.4 A reliability problem with periodic inspections

High reliability of an electronic system is often achieved by employing redundant components and having periodic inspections. Let us consider a reliability system with two identical units, where one unit is in full operation and the other unit is in warm standby. The operating unit has a constant failure rate of λ_0 and the unit in standby has a constant failure rate of λ_1, where $0 \leq \lambda_1 < \lambda_0$. Upon failure of the operating unit, the standby unit is put into full operation provided the standby is not in the failure state. Failed units are replaced only at the scheduled times $T, 2T, \ldots$ when the system is inspected. The time to replace any failed unit is negligible. A system failure occurs if both units are down. It is assumed that $(\lambda_0 + \lambda_1)T$ is sufficiently small so that a system failure is a rare event. In designing highly reliable systems a key measure of system performance is the probability distribution of the time until the first system failure.

To find the distribution of the time until the first system failure, we first compute the probability q defined by

$$q = P\{\text{system failure occurs between two inspections}\}.$$

To do so, observe that a constant failure rate λ for the lifetime of a unit implies that the lifetime has an exponential distribution with mean $1/\lambda$. Using the fact that the minimum of two independent exponentials with respective means $1/\lambda_0$ and $1/\lambda_1$ is exponentially distributed with mean $1/(\lambda_0 + \lambda_1)$, we find by conditioning on the epoch of the first failure of a unit that

$$\begin{aligned} q &= \int_0^T \left\{1 - e^{-\lambda_0(T-x)}\right\} (\lambda_0 + \lambda_1)\, e^{-(\lambda_0+\lambda_1)x}\, dx \\ &= 1 - \frac{(\lambda_0 + \lambda_1)}{\lambda_1} e^{-\lambda_0 T} + \frac{\lambda_0}{\lambda_1} e^{-(\lambda_0+\lambda_1)T}. \end{aligned}$$

Assuming that both units are in good condition at epoch 0, let

$$U = \text{the time until the first system failure.}$$

Since the process describing the state of the two units regenerates itself at each inspection, it follows that

$$P\{U > nT\} = (1 - q)^n, \quad n = 0, 1, \ldots .$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Assuming that the failure probability q is close to 0, the approximations $(1-q)^n \approx 1-nq$ and $e^{-nq} \approx 1-nq$ apply. Thus we find that

$$P\{U > t\} \approx e^{-tq/T}, \quad t \geq 0.$$

In other words, the time until the first system failure is approximately exponentially distributed.

2.3 THE FORMULA OF LITTLE

To introduce the formula of Little, we consider first two illustrative examples. In the first example a hospital admits on average 25 new patients per day. A patient stays on average 3 days in the hospital. What is the average number of occupied beds? Let $\lambda = 25$ denote the average number of new patients who are admitted per day, $W = 3$ the average number of days a patient stays in the hospital and L the average number of occupied beds. Then $L = \lambda W = 25 \times 3 = 75$ beds. In the second example a specialist shop sells on average 100 bottles of a famous Mexican premium beer per week. The shop has on average 250 bottles in inventory. What is the average number of weeks that a bottle is kept in inventory? Let $\lambda = 100$ denote the average demand per week, $L = 250$ the average number of bottles kept in stock and W the average number of weeks that a bottle is kept in stock. Then the answer is $W = L/\lambda = 250/100 = 2.5$ weeks. These examples illustrate Little's formula $L = \lambda W$. The formula of Little is a 'law of nature' that applies to almost any type of queueing system. It relates long-run averages such as the long-run average number of customers in a queue (system) and the long-run average amount of time spent per customer in the queue (system). A queueing system is described by the arrival process of customers, the service facility and the service discipline, to name the most important elements. In formulating the law of Little, there is no need to specify those basic elements. For didactical reasons, however, it is convenient to distinguish between queueing systems with infinite queue capacity and queueing systems with finite queue capacity.

Infinite-capacity queues

Consider a queueing system with infinite queue capacity, that is, every arriving customer is allowed to wait until service can be provided. Define the following random variables:

$L_q(t)$ = the number of customers in the queue at time t (excluding those in service),

$L(t)$ = the number of customers in the system at time t (including those in service),

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

D_n = the amount of time spent by the nth customer in the queue (excluding service time),

U_n = the amount of time spent by the nth customer in the system (including service time).

Let us assume that each of the stochastic processes $\{L_q(t)\}$, $\{L(t)\}$, $\{D_n\}$ and $\{U_n\}$ is regenerative and has a cycle length with a finite expectation. Then there are constants L_q, L, W_q and W such that the following limits exist and are equal to the respective constants with probability 1:

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t L_q(u)\,du = L_q \quad \text{(the long-run average number in queue)},$$

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t L(u)\,du = L \quad \text{(the long-run average number in system)},$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} D_k = W_q \quad \text{(the long-run average delay in queue per customer)},$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} U_k = W \quad \text{(the long-run average sojourn time per customer)}.$$

Now define the random variable

$$A(t) = \text{the number of customers arrived by time } t,$$

It is also assumed that, for some constant λ,

$$\lim_{t\to\infty} \frac{A(t)}{t} = \lambda \quad \text{with probability 1.}$$

The constant λ gives the long-run average arrival rate of customers. The limit λ exists when customers arrive according to a renewal process (or batches of customers arrive according to a renewal process with independent and identically distributed batch sizes).

The existence of the above limits is sufficient to prove the basic relations

$$L_q = \lambda W_q \tag{2.3.1}$$

and

$$L = \lambda W \tag{2.3.2}$$

These basic relations are the most familiar form of the formula of Little. The reader is referred to Stidham (1974) and Wolff (1989) for a rigorous proof of the formula of Little. Here we will be content to demonstrate the plausibility of this result. The

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

formula of Little is easiest understood (and reconstructed) when imagining that each customer pays money to the system manager according to some non-discrimination rule. Then it is intuitively obvious that

$$\begin{aligned}&\text{the long-run average reward per time unit earned by the system}\\ &\quad = (\text{the long-run average arrival rate of paying customers})\\ &\qquad \times (\text{the long-run average amount received per paying customer}).\end{aligned} \tag{2.3.3}$$

In regenerative queueing processes this relation can often be directly proved by using the renewal-reward theorem; see Exercise 2.26. Taking the 'money principle' (2.3.3) as starting point, it is easy to reproduce various representations of Little's law. To obtain (2.3.1), imagine that each customer pays \$1 per time unit while waiting in queue. Then the long-run average amount received per customer equals the long-run average time in queue per customer ($= W_q$). On the other hand, the system manager receives $\$j$ for each time unit that there are j customers waiting in queue. Hence the long-run average reward earned per time unit by the system manager equals the long-run average number of customers waiting in queue ($= L_q$). The average arrival rate of paying customers is obviously given by λ. Applying the relation (2.3.3) gives next the formula (2.3.1). The formula (2.3.2) can be seen by a very similar reasoning: imagine that each customer pays \$1 per time unit while in the system. Another interesting relation arises by imagining that each customer pays \$1 per time unit while in service. Denoting by $E(S)$ the long-run average service time per customer, it follows that

$$\text{the long-run average number of customers in service} = \lambda E(S). \tag{2.3.4}$$

If each customer requires only one server and each server can handle only one customer at a time, this relation leads to

$$\text{the long-run average number of busy servers} = \lambda E(S). \tag{2.3.5}$$

Finite-capacity queues

Assume now there is a maximum on the number of customers allowed in the system. In other words, there are only a finite number of waiting places and each arriving customer finding all waiting places occupied is turned away. It is assumed that a rejected customer has no further influence on the system. Let the rejection probability P_{rej} be defined by

$$P_{rej} = \text{the long-run fraction of customers who are turned away,}$$

assuming that this long-run fraction is well defined. The random variables $L(t)$, $L_q(t)$, D_n and U_n are defined as before, except that D_n and U_n now refer to the queueing time and sojourn time of the nth *accepted* customer. The constants W_q and W now represent the long-run average queueing time per *accepted* customer

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

and the long-run average sojourn time per *accepted* customer. The formulas (2.3.1), (2.3.2) and (2.3.4) need only slight modification:

$$L_q = \lambda(1 - P_{rej})W_q \quad \text{and} \quad L = \lambda(1 - P_{rej})W, \tag{2.3.6}$$

the long-run average number of customers in service

$$= \lambda(1 - P_{rej})E(S). \tag{2.3.7}$$

Heuristically, these formulas follow by applying the money principle (2.3.3) and taking only the accepted customers as paying customers.

2.4 POISSON ARRIVALS SEE TIME AVERAGES

In the analysis of queueing (and other) problems, one sometimes needs the long-run fraction of time the system is in a given state and sometimes needs the long-run fraction of arrivals who find the system in a given state. These averages can often be related to each other, but in general they are not equal to each other. To illustrate that the two averages are in general not equal to each other, suppose that customers arrive at a service facility according to a deterministic process in which the interarrival times are 1 minute. If the service of each customer is uniformly distributed between $\frac{1}{4}$ minute and $\frac{3}{4}$ minute, then the long-run fraction of time the system is empty equals $\frac{1}{2}$, whereas the long-run fraction of arrivals finding the system empty equals 1. However the two averages would have been the same if the arrival process of customers had been a Poisson process. As a prelude to the generally valid property that Poisson arrivals see time averages, we first analyse two specific problems by the renewal-reward theorem.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Example 2.4.1 A manufacturing problem

Suppose that jobs arrive at a workstation according to a Poisson process with rate λ. The workstation has no buffer to store temporarily arriving jobs. An arriving job is accepted only when the workstation is idle, and is lost otherwise. The processing times of the jobs are independent random variables having a common probability distribution with finite mean β. What is the long-run fraction of time the workstation is busy and what is the long-run fraction of jobs that are lost?

These two questions are easily answered by using the renewal-reward theorem. Let us define the following random variables. For any $t \geq 0$, let

$$I(t) = \begin{cases} 1 & \text{if the workstation is busy at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Also, for any $n = 1, 2, \ldots$, let

$$I_n = \begin{cases} 1 & \text{if the workstation is busy just prior to the } n\text{th arrival,} \\ 0 & \text{otherwise.} \end{cases}$$

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

The continuous-time process $\{I(t)\}$ and the discrete-time process $\{I_n\}$ are both regenerative. The arrival epochs occurring when the workstation is idle are regeneration epochs for the two processes. Why? Let us say that a cycle starts each time an arriving job finds the workstation idle. The long-run fraction of time the workstation is busy is equal to the expected amount of time the workstation is busy during one cycle divided by the expected length of one cycle. The expected length of the busy period in one cycle equals β. Since the Poisson arrival process is memoryless, the expected length of the idle period during one cycle equals the mean interarrival time $1/\lambda$. Hence, with probability 1,

$$\text{the long-run fraction of time the workstation is busy} = \frac{\beta}{\beta + 1/\lambda}. \tag{2.4.1}$$

The long-run fraction of jobs that are lost equals the expected number of jobs lost during one cycle divided by the expected number of jobs arriving during one cycle. Since the arrival process is a Poisson process, the expected number of (lost) arrivals during the busy period in one cycle equals $\lambda \times E(\text{processing time of a job}) = \lambda\beta$. Hence, with probability 1,

$$\text{the long-run fraction of jobs that are lost} = \frac{\lambda\beta}{1 + \lambda\beta}. \tag{2.4.2}$$

Thus, we obtain from (2.4.1) and (2.4.2) the remarkable result

$$\text{the long-run fraction of arrivals finding the workstation busy} = \text{the long-run fraction of time the workstation is busy.} \tag{2.4.3}$$

Incidentally, it is interesting to note that in this loss system the long-run fraction of lost jobs is insensitive to the form of the distribution function of the processing time but needs only the first moment of this distribution. This simple loss system is a special case of Erlang's loss model to be discussed in Chapter 5.

Example 2.4.2 An inventory model

Consider a single-product inventory system in which customers asking for the product arrive according to a Poisson process with rate λ. Each customer asks for one unit of the product. Each demand which cannot be satisfied directly from stock on hand is lost. Opportunities to replenish the inventory occur according to a Poisson process with rate μ. This process is assumed to be independent of the demand process. For technical reasons a replenishment can only be made when the inventory is zero. The inventory on hand is raised to the level Q each time a replenishment is done. What is the long-run fraction of time the system is out of stock? What is the long-run fraction of demand that is lost?

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

In the same way as in Example 2.4.1, we define the random variables

$$I(t) = \begin{cases} 1 & \text{if the system is out of stock at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

and

$$I_n = \begin{cases} 1 & \text{if the system is out of stock when the } n\text{th demand occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

The continuous-time process $\{I(t)\}$ and the discrete-time process $\{I_n\}$ are both regenerative. The regeneration epochs are the demand epochs at which the stock on hand drops to zero. Why? Let us say that a cycle starts each time the stock on hand drops to zero. The system is out of stock during the time elapsed from the beginning of a cycle until the next inventory replenishment. This amount of time is exponentially distributed with mean $1/\mu$. The expected amount of time it takes to go from stock level Q to 0 equals Q/λ. Hence, with probability 1,

$$\text{the long-run fraction of time the system is out of stock} = \frac{1/\mu}{1/\mu + Q/\lambda}. \tag{2.4.4}$$

To find the fraction of demand that is lost, note that the expected amount of demand lost in one cycle equals $\lambda \times E(\text{amount of time the system is out of stock during one cycle}) = \lambda/\mu$. Hence, with probability 1,

$$\text{the long-run fraction of demand that is lost} = \frac{\lambda/\mu}{\lambda/\mu + Q}. \tag{2.4.5}$$

Together (2.4.4) and (2.4.5) lead to this remarkable result:

$$\text{the long-run fraction of customers finding the system out of stock} = \text{the long-run fraction of time the system is out of stock.} \tag{2.4.6}$$

The relations (2.4.3) and (2.4.6) are particular instances of the property 'Poisson arrivals see time averages'. Roughly stated, this property expresses that in statistical equilibrium the distribution of the state of the system *just prior* to an arrival epoch is the same as the distribution of the state of the system at an *arbitrary* epoch when arrivals occur according to a Poisson process. An intuitive explanation of the property 'Poisson arrivals see time averages' is that Poisson arrivals occur completely randomly in time; cf. Theorem 1.1.5.

Next we discuss the property of 'Poisson arrivals see time averages' in a broader context. For ease of presentation we use the terminology of Poisson arrivals. However, the results below also apply to Poisson processes in other contexts. For some

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

specific problem let the continuous-time stochastic process $\{X(t), t \geq 0\}$ describe the evolution of the state of a system and let $\{N(t), t \geq 0\}$ be a renewal process describing arrivals to that system. As examples:

(a) $X(t)$ is the number of customers present at time t in a queueing system.

(b) $X(t)$ describes jointly the inventory level and the prevailing production rate at time t in a production/inventory problem with a variable production rate.

It is assumed that the arrival process $\{N(t), t \geq 0\}$ can be seen as an exogenous factor to the system and is not affected by the system itself. More precisely, the following assumption is made.

Lack of anticipation assumption *For each $u \geq 0$ the future arrivals occurring after time u are independent of the history of the process $\{X(t)\}$ up to time u.*

It is not necessary to specify how the arrival process $\{N(t)\}$ precisely interacts with the state process $\{X(t)\}$. Denoting by τ_n the nth arrival epoch, let the random variable X_n be defined by $X(\tau_n^-)$. In other words,

$$X_n = \text{the state of the system just prior to the } n\text{th arrival epoch.}$$

Let B be any set of states for the $\{X(t)\}$ process. For each $t \geq 0$, define the indicator variable

$$I_B(t) = \begin{cases} 1 & \text{if } X(t) \in B, \\ 0 & \text{otherwise.} \end{cases}$$

Also, for each $n = 1, 2, \ldots$, define the indicator variable $I_n(B)$ by

$$I_n(B) = \begin{cases} 1 & \text{if } X_n \in B, \\ 0 & \text{otherwise.} \end{cases}$$

The technical assumption is made that the sample paths of the continuous-time process $\{I_B(t), t \geq 0\}$ are right-continuous and have left-hand limits. In practical situations this assumption is always satisfied.

Theorem 2.4.1 (Poisson arrivals see time averages) *Suppose that the arrival process $\{N(t)\}$ is a Poisson process with rate λ. Then:*

(a) For any $t > 0$,

$$E[\textit{number of arrivals in } (0, t) \textit{ finding the system in the set } B] = \lambda E\left[\int_0^t I_B(u)\, du\right].$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

(b) With probability 1*, the long-run fraction of arrivals who find the system in the set B of states equals the long-run fraction of time the system is in the set B of states. That is, with probability* 1*,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_k(B) = \lim_{t\to\infty} \frac{1}{t} \int_0^t I_B(u)\, du.$$

Proof See Wolff (1982).

It is remarkable in Theorem 2.4.1 that E[number of arrivals in $(0, t)$ finding the system in the set B] is equal to $\lambda \times E$[amount of time in $(0, t)$ that the system is in the set B], although there is dependency between the arrivals in $(0, t)$ and the evolution of the state of the system during $(0, t)$. This result is characteristic for the Poisson process.

The property 'Poisson arrivals see time averages' is usually abbreviated as ***PASTA***. Theorem 2.4.1 has a useful corollary when it is assumed that the continuous-time process $\{X(t)\}$ is a regenerative process whose cycle length has a finite positive mean. Define the random variables T_B and N_B by

T_B = amount of time the process $\{X(t)\}$ is in the set B of states during one cycle,

N_B = number of arrivals during one cycle who find the process $\{X(t)\}$ in the set of B states.

The following corollary will be very useful in the algorithmic analysis of queueing systems in Chapter 9.

Corollary 2.4.2 *If the arrival process* $\{N(t)\}$ *is a Poisson process with rate* λ*, then*

$$E(N_B) = \lambda E(T_B).$$

Proof Denote by the random variables T and N the length of one cycle and the number of arrivals during one cycle. Then, by Theorem 2.2.3,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t I_B(u)\, du = \frac{E(T_B)}{E(T)} \quad \text{with probability 1}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_k(B) = \frac{E(N_B)}{E(N)} \quad \text{with probability 1.}$$

It now follows from part (b) of Theorem 2.4.1 that $E(N_B)/E(N) = E(T_B)/E(T)$. Thus the corollary follows if we can verify that $E(N)/E(T) = \lambda$. To do so, note that the regeneration epochs for the process $\{X(t)\}$ are also regeneration epochs for the Poisson arrival process. Thus, by the renewal-reward theorem, the long-run average number of arrivals per time unit equals $E(N)/E(T)$, showing that $E(N)/E(T) = \lambda$.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

To conclude this section, we use the PASTA property to derive in a heuristic way one of the most famous formulas from queueing theory.

2.5 THE POLLACZEK–KHINTCHINE FORMULA

Suppose customers arrive at a service facility according to a Poisson process with rate λ. The service times of the customers are independent random variables having a common probability distribution with finite first two moments $E(S)$ and $E(S^2)$. There is a single server and ample waiting room for arriving customers finding the server busy. Each customer waits until service is provided. The server can handle only one customer at a time. This particular queueing model is abbreviated as the $M/G/1$ queue; see Kendall's notation in Section 9.1. The offered load ρ is defined by

$$\rho = \lambda E(S)$$

and it is assumed that $\rho < 1$. By Little's formula (2.3.5) the load factor ρ can be interpreted as the long-run fraction of time the server is busy. Important performance measures are

L_q = the long-run average number of customers waiting in queue,

W_q = the long-run average time spent per customer in queue.

The Pollaczek–Khintchine formula states that

$$W_q = \frac{\lambda E(S^2)}{2(1-\rho)}. \tag{2.5.1}$$

This formula also implies an explicit expression for L_q by Little's formula

$$L_q = \lambda W_q; \tag{2.5.2}$$

see Section 2.3. The Pollaczek–Khintchine formula gives not only an explicit expression for W_q, but more importantly it gives useful qualitative insights as well. It shows that the average delay per customer in the $M/G/1$ queue uses the service-time distribution only through its first two moments. Denoting by $c_S^2 = \sigma^2(S)/E^2(S)$ the squared coefficient of variation of the service time and using the relation (2.1.10), we can write the Pollaczek–Khintchine formula in the more insightful form

$$W_q = \frac{1}{2}(1 + c_S^2)\frac{\rho E(S)}{1-\rho}. \tag{2.5.3}$$

Hence the Pollaczek–Khintchine formula shows that the average delay per customer decreases according to the factor $\frac{1}{2}(1 + c_S^2)$ when the variability in the service is reduced while the average arrival rate and the mean service time are kept

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

fixed. Noting that $c_S^2 = 1$ for exponentially distributed service times, the expression (2.5.3) can also be written as

$$W_q = \frac{1}{2}(1 + c_S^2) W_q(\text{exp}), \tag{2.5.4}$$

where $W_q(\text{exp}) = \rho E(S)/(1-\rho)$ denotes the average delay per customer for the case of exponential services. In particular, writing $W_q = W_q(\text{det})$ for deterministic services ($c_S^2 = 0$), we have

$$W_q(\text{det}) = \frac{1}{2} W_q(\text{exp}). \tag{2.5.5}$$

It will be seen in Chapter 9 that the structural form (2.5.4) is very useful to design approximations in more complex queueing models.

Another important feature shown by the Pollaczek–Khintchine formula is that the average delay and average queue size increase in a *non-linear* way when the offered load ρ increases. A twice as large value for the offered load does not imply a twice as large value for the average delay! On the contrary, the average delay and the average queue size explode when the average arrival rate becomes very close to the average service rate. Differentiation of W_q as a function of ρ shows that the slope of increase of W_q as a function of ρ is proportional to $(1-\rho)^{-2}$. As an illustration a small increase in the average arrival rate when the load $\rho = 0.9$ causes an increase in the average delay 25 times greater than it would cause when the load $\rho = 0.5$. This non-intuitive finding demonstrates the danger of designing a stochastic system with too high a utilization level, since then a small increase in the offered load will in general cause a dramatic degradation in system performance.

We have not yet proved the Pollaczek–Khintchine formula. First we give a heuristic derivation and next we give a rigorous proof.

Heuristic derivation

Tag a customer who arrives when the system has reached statistical equilibrium. Denote its waiting time in queue by the random variable D_{tag}. Heuristically, $E(D_{tag}) = W_q$. By the PASTA property, the expected number of customers in queue seen upon arrival by the tagged customer equals L_q. Noting that ρ is the long-run fraction of time the server is busy, it also follows that the tagged customer finds the server busy upon arrival with probability ρ. Using the result (2.1.8) for the excess variable, it is plausible that the expected remaining service time of the customer seen in service by a Poisson arrival equals $\frac{1}{2}E(S^2)/E(S)$. Putting the pieces together, we find the relation

$$E(D_{tag}) = L_q E(S) + \rho \frac{E(S^2)}{2E(S)}.$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Substituting $E(D_{tag}) = W_q$ and $L_q = \lambda W_q$, the relation becomes

$$W_q = \lambda E(S) W_q + \frac{\rho E(S^2)}{2E(S)}$$

yielding the Pollaczek–Khintchine formula for W_q.

Rigorous derivation

A rigorous derivation of the Pollaczek–Khintchine formula can be given by using the powerful generating-function approach. Define first the random variables

$L(t) =$ the number of customers present at time t,

$Q_n =$ the number of customers present just after the nth service completion epoch,

$L_n =$ the number of customers present just before the nth arrival epoch.

The processes $\{L(t)\}$, $\{Q_n\}$ and $\{L_n\}$ are regenerative stochastic processes with finite expected cycle lengths. Denote the corresponding limiting distributions by

$$p_j = \lim_{t\to\infty} P\{L(t) = j\},\ q_j = \lim_{n\to\infty} P\{Q_n = j\} \quad \text{and} \quad \pi_j = \lim_{n\to\infty} P\{L_n = j\}$$

for $j = 0, 1, \ldots$. The existence of the limiting distributions can be deduced from Theorem 2.2.4 (the amount of time elapsed between two arrivals that find the system empty has a probability density and the number of customers served during this time has an aperiodic distribution). We omit the details. The limiting probabilities can also be interpreted as long-run averages. For example, q_j is the long-run fraction of customers leaving j other customers behind upon service completion and π_j is the long-run fraction of customers finding j other customers present upon arrival. The following important identity holds:

$$\pi_j = p_j = q_j, \quad j = 0, 1, \ldots . \tag{2.5.6}$$

Since the arrival process is a Poisson process, the equality $\pi_j = p_j$ is readily verified from Theorem 2.4.1. To verify the equality $\pi_j = q_j$, define the random variable $L_n^{(j)}$ as the number of customers over the first n arrivals who see j other customers present upon arrival and define the random variable $Q_n^{(j)}$ as the number of service completion epochs over the first n service completions at which j customers are left behind. Customers arrive singly and are served singly. Thus between any two arrivals that find j other customers present there must be a service completion at which j customers are left behind and, conversely, between any two service completions at which j customers are left behind there must be an arrival that sees j other customers present. By this up- and downcrossing argument, we have for

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

each j that

$$\left| L_n^{(j)} - Q_n^{(j)} \right| \leq 1, \quad n = 1, 2, \ldots .$$

Consequently, $\pi_j = \lim_{n\to\infty} L_n^{(j)}/n = \lim_{n\to\infty} Q_n^{(j)}/n = q_j$ for all j. We are now ready to prove that

$$\lim_{n\to\infty} E(z^{Q_n}) = \frac{(1-z)q_0 A(z)}{A(z) - z}, \tag{2.5.7}$$

where

$$A(z) = \int_0^\infty e^{-\lambda t(1-z)} b(t)\, dt$$

with $b(t)$ denoting the probability density of the service time of a customer. Before proving this result, we note that the unknown q_0 is determined by the fact that the left-hand side of (2.5.7) equals 1 for $z = 1$. By applying L'Hospital's rule, we find $q_0 = 1 - \rho$, in agreement with Little's formula $1 - p_0 = \rho$. By the bounded convergence theorem in Appendix A,

$$\lim_{n\to\infty} E(z^{Q_n}) = \lim_{n\to\infty} \sum_{j=0}^{\infty} P\{Q_n = j\} z^j = \sum_{j=0}^{\infty} q_j z^j, \quad |z| \leq 1.$$

Hence, by (2.5.6) and (2.5.7),

$$\sum_{j=0}^{\infty} p_j z^j = \frac{(1-\rho)(1-z)A(z)}{A(z) - z}. \tag{2.5.8}$$

Since the long-run average queue size L_q is given by

$$L_q = \sum_{j=1}^{\infty} (j-1) p_j = \sum_{j=0}^{\infty} j p_j - (1 - p_0)$$

(see Exercise 2.28), the Pollaczek–Khintchine formula for L_q follows by differentiating the right-hand side of (2.5.8) and taking $z = 1$ in the derivative. It remains to prove (2.5.7). To do so, note that

$$Q_n = Q_{n-1} - \delta(Q_{n-1}) + A_n, \quad n = 1, 2, \ldots ,$$

where $\delta(x) = 1$ for $x > 0$, $\delta(x) = 0$ for $x = 0$ and A_n is the number of customers arriving during the nth service time. By the law of total probability,

$$P\{A_n = k\} = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} b(t)\, dt, \quad k = 0, 1, \ldots$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

and so

$$\sum_{k=0}^{\infty} P\{A_n = k\} z^k = \int_0^{\infty} e^{-\lambda t(1-z)} b(t)\, dt.$$

Since the random variables $Q_{n-1} - \delta(Q_{n-1})$ and A_n are independent of each other,

$$E(z^{Q_n}) = E(z^{Q_{n-1}-\delta(Q_{n-1})}) E(z^{A_n}). \tag{2.5.9}$$

We have

$$\begin{aligned} E(z^{Q_{n-1}-\delta(Q_{n-1})}) &= P\{Q_{n-1} = 0\} + \sum_{j=1}^{\infty} z^{j-1} P\{Q_{n-1} = j\} \\ &= P\{Q_{n-1} = 0\} + \frac{1}{z}[E(z^{Q_{n-1}}) - P\{Q_{n-1} = 0\}]. \end{aligned}$$

Substituting this in (2.5.9), we find

$$zE(z^{Q_n}) = \Big[E(z^{Q_{n-1}}) - (1-z)P\{Q_{n-1} = 0\}\Big] A(z).$$

Letting $n \to \infty$, we next obtain the desired result (2.5.7). This completes the proof.

Before concluding this section, we give an amusing application of the Pollaczek–Khintchine formula.

Example 2.5.1 Ladies in waiting*

Everybody knows women spend on average more time in the loo than men. As worldwide studies show, women typically take 89 seconds to use the loo—about twice as long as the 39 seconds required by the average man. However, this does not mean that the queue for the women's loo is twice as long as the queue for the men's. The sequence for the women's loo is usually far longer. To explain this using the Pollaczek–Khintchine formula, let us make the following reasonable assumptions:

1. Men and women arrive at the loo according to independent Poisson processes with the same rates.
2. The expected amount of time people spend in the loo is twice as large for women as for men.
3. The coefficient of variation of the time people spend in the loo is larger for women than for men.
4. There is one loo for women only and one loo for men only.

*This application is based on the article 'Ladies Waiting' by Robert Matthews in *New Scientist*, Vol. 167, Issue 2249, 29 July 2000.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Let λ_w and λ_m denote the average arrival rates of women and men. Let μ_w and c_w denote the mean and the coefficient of variation of the amount of time a woman spends in the loo. Similarly, μ_m and c_m are defined for men. It is assumed that $\lambda_w \mu_w < 1$. Using the assumptions $\lambda_w = \lambda_m$, $\mu_w = 2\mu_m$ and $c_w \geq c_m$, it follows from (2.5.2) and the Pollaczek–Khintchine formula (2.5.3) that

$$\begin{aligned}\text{the average queue size for the women's loo} &= \frac{1}{2}(1+c_w^2)\frac{(\lambda_w\mu_w)^2}{1-\lambda_w\mu_w} \geq \frac{1}{2}(1+c_m^2)\frac{(2\lambda_m\mu_m)^2}{1-2\lambda_m\mu_m} \\ &\geq 4 \times \frac{1}{2}(1+c_m^2)\frac{(\lambda_m\mu_m)^2}{1-\lambda_m\mu_m}.\end{aligned}$$

Hence

$$\text{the average queue size for the women's loo} \geq 4 \times (\text{the average queue size for the men's loo}).$$

The above derivation uses the estimate $1-2\lambda_m\mu_m \leq 1-\lambda_m\mu_m$ and thus shows that the relative difference actually increases much faster than a factor 4 when the utilization factor $\lambda_w\mu_w$ becomes closer to 1.

*Laplace transform of the waiting-time probabilities**

The generating-function method enabled us to prove the Pollaczek–Khintchine formula for the average queue size. Using Little's formula we next found the Pollaczek–Khintchine formula for the average delay in queue of a customer. The latter formula can also be directly obtained from the Laplace transform of the waiting-time distribution. This Laplace transform is also of great importance in itself. The waiting-time probabilities can be calculated by numerical inversion of the Laplace transform; see Appendix F. A simple derivation can be given for the Laplace transform of the waiting-time distribution in the $M/G/1$ queue when service is in order of arrival. The derivation parallels the derivation of the generating function of the number of customers in the system.

Denote by D_n the delay in queue of the nth arriving customer and let the random variables S_n and τ_n denote the service time of the nth customer and the time elapsed between the arrivals of the nth customer and the $(n+1)$th customer. Since $D_{n+1} = 0$ if $D_n + S_n < \tau_n$ and $D_{n+1} = D_n + S_n - \tau_n$ otherwise, we have

$$D_{n+1} = (D_n + S_n - \tau_n)^+, \quad n = 1, 2, \ldots, \tag{2.5.10}$$

where x^+ is the usual notation for $x = \max(x, 0)$. From the recurrence formula (2.5.10), we can derive that for all s with $\mathrm{Re}(s) \geq 0$ and $n = 1, 2, \ldots$

$$(\lambda - s)E\left(e^{-sD_{n+1}}\right) = \lambda E\left(e^{-sD_n}\right)b^*(s) - sP\{D_{n+1} = 0\}, \tag{2.5.11}$$

*This section can be skipped at first reading.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

where $b^*(s) = \int_0^\infty e^{-sx} b(x)\,dx$ denotes the Laplace transform of the probability density $b(x)$ of the service time. To prove this, note that D_n, S_n and τ_n are independent of each other. This implies that, for any $x > 0$,

$$E\left[e^{-s(D_n+S_n-\tau_n)^+} \mid D_n + S_n = x\right]$$
$$= \int_0^x e^{-s(x-y)} \lambda e^{-\lambda y}\,dy + \int_x^\infty e^{-s\times 0} \lambda e^{-\lambda y}\,dy$$
$$= \frac{\lambda}{\lambda - s}(e^{-sx} - e^{-\lambda x}) + e^{-\lambda x} = \frac{1}{\lambda - s}(\lambda e^{-sx} - s e^{-\lambda x})$$

for $s \neq \lambda$ (using L'Hospital's rule it can be seen that this relation also holds for $s = \lambda$). Hence, using (2.5.10),

$$(\lambda - s) E\left(e^{-sD_{n+1}}\right) = \lambda E\left[e^{-s(D_n+S_n)}\right] - s E\left[e^{-\lambda(D_n+S_n)}\right].$$

Since $P\{(D_n + S_n - \tau_n)^+ = 0 \mid D_n + S_n = x\} = e^{-\lambda x}$, we also have

$$P\{D_{n+1} = 0\} = E\left[e^{-\lambda(D_n+S_n)}\right].$$

The latter two relations and $E\left[e^{-s(D_n+S_n)}\right] = E\left(e^{-sD_n}\right) E\left(e^{-sS_n}\right)$ lead to (2.5.11). The steady-state waiting-time distribution function $W_q(x)$ is defined by

$$W_q(x) = \lim_{n\to\infty} P\{D_n \leq x\}, \quad x \geq 0.$$

The existence of this limit can be proved from Theorem 2.2.4. Let the random variable D_∞ have $W_q(x)$ as probability distribution function. Then, by the bounded convergence theorem in Appendix A, $E(e^{-sD_\infty}) = \lim_{n\to\infty} E(e^{-sD_n})$. Using (2.5.6), it follows from $\lim_{n\to\infty} P\{D_{n+1} = 0\} = \pi_0$ and $q_0 = 1 - \rho$ that $\lim_{n\to\infty} P\{D_{n+1} = 0\} = 1 - \rho$. Letting $n \to \infty$ in (2.5.11), we find that

$$E\left(e^{-sD_\infty}\right) = \frac{(1-\rho)s}{s - \lambda + \lambda b^*(s)}. \tag{2.5.12}$$

Noting that $P\{D_\infty \leq x\} = W_q(x)$ and using relation (E.7) in Appendix E, we get from (2.5.12) the desired result:

$$\int_0^\infty e^{-sx} \left\{1 - W_q(x)\right\} dx = \frac{\rho s - \lambda + \lambda b^*(s)}{s(s - \lambda + \lambda b^*(s))}. \tag{2.5.13}$$

Taking the derivative of the right-hand side of (2.5.13) and putting $s = 0$, we obtain

$$\int_0^\infty \left\{1 - W_q(x)\right\} dx = \frac{\lambda E(S^2)}{2(1-\rho)},$$

in agreement with the Pollaczek–Khintchine formula (2.5.1).

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Remark 2.5.1 Relation between queue size and waiting time

Let the random variable $L_q^{(\infty)}$ be distributed according to the limiting distribution of the number of customers in queue at an arbitrary point in time. That is, $P\{L_q^{(\infty)} = j\} = p_{j+1}$ for $j \geq 1$ and $P\{L_q^{(\infty)} = 0\} = p_0 + p_1$. Then the generating function of $L_q^{(\infty)}$ and the Laplace transform of the delay distribution are related to each other by

$$E(z^{L_q^{(\infty)}}) = E[e^{-\lambda(1-z)D_\infty}], \quad |z| \leq 1. \tag{2.5.14}$$

A direct probabilistic proof of this important relation can be given. Denote by L_n the number of customers left behind in queue when the nth customer enters service. Since service is in order of arrival, L_n is given by the number of customers arriving during the delay D_n of the nth customer. Since the generating function of a Poisson distributed variable with mean δ is $\exp(-\delta(1-z))$, it follows that for any $x \geq 0$ and $n \geq 1$,

$$E(z^{L_n} | D_n = x) = e^{-\lambda x(1-z)}.$$

Hence

$$E(z^{L_n}) = E[e^{-\lambda(1-z)D_n}], \quad n \geq 1. \tag{2.5.15}$$

The limiting distribution of L_n as $n \to \infty$ is the same as the probability distribution of $L_q^{(\infty)}$. This follows from an up- and downcrossing argument: the long-run fraction of customers leaving j other customers behind in queue when entering service equals the long-run fraction of customers finding j other customers in queue upon arrival. Noting that there is a single server and using the PASTA property, it follows that the latter fraction equals p_{j+1} for $j \geq 1$ and $p_0 + p_1$ for $j = 0$. This proves that the limiting distribution of L_n equals the distribution of $L_q^{(\infty)}$. Note that, by Theorem 2.2.4, L_n has a limiting distribution as $n \to \infty$. Letting $n \to \infty$ in (2.5.15), the result (2.5.14) follows.

Letting $w_q(x)$ denote the derivative of the waiting-time distribution function $W_q(x)$ for $x > 0$, note that for the $M/G/1$ queue the relation (2.5.14) can be restated as

$$p_{j+1} = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^j}{j!} w_q(x)\, dx, \quad j = 1, 2, \ldots .$$

The relation (2.5.14) applies to many other queueing systems with Poisson arrivals. The importance of (2.5.14) is that this relation enables us to directly obtain the Laplace transform of the waiting-time distribution function from the generating function of the queue size. To illustrate this, note that $E(z^{L_q^{(\infty)}}) = p_0 + \frac{1}{z}[P(z) - p_0]$ for the $M/G/1$ queue, where $P(z) = \sum_{j=0}^\infty p_j z^j$ is given by (2.5.8). Using this relation together with (2.5.8) and noting that $A(z) = b^*(\lambda(1-z))$, it follows from the basic relation (2.5.14) that $E(e^{-sD_\infty})$ is indeed given by (2.5.12).

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

2.6 A CONTROLLED QUEUE WITH REMOVABLE SERVER*

Consider a production facility at which production orders arrive according to a Poisson process with rate λ. The production times $\tau_1, \tau_2, \ldots$ of the orders are independent random variables having a common probability distribution function F with finite first two moments. Also, the production process is independent of the arrival process. The facility can only work on one order at a time. It is assumed that $E(\tau_1) < 1/\lambda$; that is, the average production time per order is less than the mean interarrival time between two consecutive orders. The facility operates only intermittently and is shut down when no orders are present any more. A fixed set-up cost of $K > 0$ is incurred each time the facility is reopened. Also a holding cost $h > 0$ per time unit is incurred for each order waiting in queue. The facility is only turned on when enough orders have accumulated. The so-called N-policy reactivates the facility as soon as N orders are present. For ease we assume that it takes a zero set-up time to restart production. How do we choose the value of the control parameter N such that the long-run average cost per time unit is minimal?

To analyse this problem, we first observe that for a given N-policy the stochastic process describing jointly the number of orders present and the status of the facility (on or off) regenerates itself each time the facility is turned on. Define a cycle as the time elapsed between two consecutive reactivations of the facility. Clearly, each cycle consists of a busy period B with production and an idle period I with no production. We deal separately with the idle and the busy periods. Using the memoryless property of the Poisson process, the length of the idle period is the sum of N exponential random variables each having mean $1/\lambda$. Hence

$$E(\text{length of the idle period } I) = \frac{N}{\lambda}.$$

Similarly,

$$E(\text{holding cost incurred during } I) = h\left(\frac{N-1}{\lambda} + \cdots + \frac{1}{\lambda}\right).$$

To deal with the busy period, we define for $n = 1, 2, \ldots$ the quantities

$$t_n = \text{the expected time until the facility becomes empty given that at epoch 0 a production starts with } n \text{ orders present,}$$

and

$$h_n = \text{the expected holding costs incurred until the facility becomes empty given that at epoch 0 a production starts with } n \text{ orders present.}$$

These quantities are independent of the control rule considered. In particular, the expected length of a busy period equals t_N and the expected holding costs incurred

*This section contains specialized material and can be skipped at first reading.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

during a busy period equals h_N. By the renewal-reward theorem,

$$\text{the long-run average cost per time unit} = \frac{(h/2\lambda)N(N-1) + K + h_N}{N/\lambda + t_N}$$

with probability 1. To find the functions t_n and h_n, we need

a_j = the probability that j orders arrive during the production time of a single order.

Assume for ease that the production time has a probability density $f(x)$. By conditioning on the production time and noting that the number of orders arriving in a fixed time y is Poisson distributed with mean λy, it follows that

$$a_j = \int_0^\infty e^{-\lambda y} \frac{(\lambda y)^j}{j!} f(y)\, dy, \quad j = 0, 1, \ldots .$$

It is readily verified that

$$\sum_{j=1}^{\infty} j a_j = \lambda E(\tau_1) \quad \text{and} \quad \sum_{j=1}^{\infty} j^2 a_j = \lambda^2 E(\tau_1^2) + \lambda E(\tau_1). \tag{2.6.1}$$

We now derive recursion relations for the quantities t_n and h_n. Suppose that at epoch 0 a production starts with n orders present. If the number of new orders arriving during the production time of the first order is j, then the time to empty the system equals the first production time plus the time to empty the system starting with $n - 1 + j$ orders present. Thus

$$t_n = E(\tau_1) + \sum_{j=0}^{\infty} t_{n-1+j} a_j, \quad n = 1, 2, \ldots ,$$

where $t_0 = 0$. Similarly, we derive a recursion relation for the h_n. To do so, note that relation (1.1.10) implies that the expected holding cost for new orders arriving during the first production time τ_1 equals $\frac{1}{2}h\lambda E(\tau_1^2)$. Hence

$$h_n = (n-1)hE(\tau_1) + \frac{1}{2}h\lambda E(\tau_1^2) + \sum_{j=0}^{\infty} h_{n-1+j} a_j, \quad n = 1, 2, \ldots ,$$

where $h_0 = 0$. In a moment it will be shown that t_n is linear in n and h_n is quadratic in n. Substituting these functional forms in the above recursion relations and using (2.6.1), we find after some algebra that for $n = 1, 2, \ldots ,$

$$t_n = \frac{nE(\tau_1)}{1 - \lambda E(\tau_1)}, \tag{2.6.2}$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

$$h_n = \frac{h}{1-\lambda E(\tau_1)}\left[\frac{1}{2}n(n-1)E(\tau_1) + \frac{\lambda n E(\tau_1^2)}{2\{1-\lambda E(\tau_1)\}}\right]. \tag{2.6.3}$$

To verify that t_n is linear in n and h_n is quadratic in n, a brilliant idea due to Takács (1962) is used. First observe that t_n and h_n do not depend on the specific order in which the production orders are coped with during the production process. Imagine now the following production discipline. The n initial orders $O_1, \ldots, O_n$ are separated. Order O_1 is produced first, after which all orders (if any) are produced that have arrived during the production time of O_1, and this way of production is continued until the facility is free of all orders but $O_2, \ldots, O_n$. Next this procedure is repeated with order O_2, etc. Thus we find that $t_n = nt_1$, proving that t_n is linear in n. The memoryless property of the Poisson process is crucial in this argument. Why? The same separation argument is used to prove that h_n is quadratic in n. Since $h_1 + (n-k) \times ht_1$ gives the expected holding cost incurred during the time to free the system of order O_k and its direct descendants until only the orders $O_{k+1}, \ldots, O_n$ are left, it follows that

$$h_n = \sum_{k=1}^{n}\{h_1 + (n-k)ht_1\} = nh_1 + \frac{1}{2}hn(n-1)t_1.$$

Combining the above results we find for the N-policy that

$$\begin{aligned}&\text{the long-run average cost per time unit} \\ &= \frac{\lambda(1-\rho)K}{N} + h\left\{\frac{\lambda^2 E(\tau_1^2)}{2(1-\rho)} + \frac{N-1}{2}\right\},\end{aligned} \tag{2.6.4}$$

where $\rho = \lambda E(\tau_1)$. It is worth noting here that this expression needs only the first two moments from the production time. Also note that, by putting $K = 0$ and $h = 1$ in (2.6.4),

$$\begin{aligned}&\text{the long-run average number of orders waiting in queue} \\ &= \frac{\lambda^2 E(\tau_1^2)}{2(1-\rho)} + \frac{N-1}{2}.\end{aligned}$$

For the special case of $N = 1$ this formula reduces to the famous Pollaczek–Khintchine formula for the average queue length in the standard $M/G/1$ queue; see Section 2.5.

The optimal value of N can be obtained by differentiating the right-hand side of (2.6.4), in which we take N as a continuous variable. Since the average cost is convex in N, it follows that the average cost is minimal for one of the two integers nearest to

$$N^* = \sqrt{\frac{2\lambda(1-\rho)K}{h}}.$$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

2.7 AN UP- AND DOWNCROSSING TECHNIQUE

In this section we discuss a generally applicable up- and downcrossing technique that, in conjunction with the PASTA property, can be used to establish relations between customer-average and time-average probabilities in queueing systems. To illustrate this, we consider the so-called $GI/M/1$ queue. In this single-server system, customers arrive according to a renewal process and the service times of the customers have a common exponential distribution. The single server can handle only one customer at a time and there is ample waiting room for customers who find the server busy upon arrival. The service times of the customers are independent of each other and are also independent of the arrival process. Denoting by λ the average arrival rate ($1/\lambda$ = the mean interarrival time) and by β the service rate ($1/\beta$ = the mean service time), it is assumed that $\lambda < \beta$.

The continuous-time stochastic process $\{X(t),\ t \geq 0\}$ and the discrete-time stochastic process $\{X_n,\ n = 1, 2, \ldots\}$ are defined by

$$X(t) = \text{the number of customers present at time } t,$$

and

$$X_n = \text{the number of customers present just prior to the } n\text{th arrival epoch.}$$

The stochastic processes $\{X(t)\}$ and $\{X_n\}$ are both regenerative. The regeneration epochs are the epochs at which an arriving customer finds the system empty. It is stated without proof that the assumption of $\lambda/\beta < 1$ implies that the processes have a finite mean cycle length. Thus we can define the time-average and the customer-average probabilities p_j and π_j by

$$p_j = \text{the long-run fraction of time that } j \text{ customers are present}$$

and

$$\pi_j = \text{the long-run fraction of customers who find } j \text{ other customers present upon arrival}$$

for $j = 0, 1, \ldots$. Time averages are averages over time, and customer averages are averages over customers. To be precise, $p_j = \lim_{t\to\infty}(1/t)\int_0^t I_j(u)\,du$ and $\pi_j = \lim_{n\to\infty}(1/n)\sum_{k=1}^{n} I_k(j)$, where $I_j(t) = 1$ if j customers are present at time t and $I_j(t) = 0$ otherwise, and $I_n(j) = 1$ if j other customers are present just before the nth arrival epoch and $I_n(j) = 0$ otherwise. The probabilities p_j and π_j are related to each other by

$$\lambda\pi_{j-1} = \beta p_j, \quad j = 1, 2, \ldots . \tag{2.7.1}$$

The proof of this result is instructive and is based on three observations. Before giving the three steps, let us say that the continuous-time process $\{X(t)\}$ makes an *upcrossing* from state $j-1$ to state j if a customer arrives and finds $j-1$

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

other customers present. The process $\{X(t)\}$ makes a *downcrossing* from state j to state $j-1$ if the service of a customer is completed and $j-1$ other customers are left behind.

Observation 1 Since customers arrive singly and are served singly, the long-run average number of upcrossings from $j-1$ to j per time unit equals the long-run average number of downcrossings from j to $j-1$ per time unit. This follows by noting that in any finite time interval the number of upcrossings from $j-1$ to j and the number of downcrossings from j to $j-1$ can differ at most by 1.

Observation 2 The long-run fraction of customers seeing $j-1$ other customers upon arrival is equal to

$$\frac{\text{the long-run average number of upcrossings from } j-1 \text{ to } j \text{ per time unit}}{\text{the long-run average number of arrivals per time unit}}$$

for $j = 1, 2, \ldots$. In other words, the long-run average number of upcrossings from $j-1$ to j per time unit equals $\lambda \pi_{j-1}$.

The latter relation for fixed j is in fact a special case of the Little relation (2.4.1) by assuming that each customer finding $j-1$ other customers present upon arrival pays \$1 (using this reward structure observation 2 can also be obtained directly from the renewal-reward theorem). Observations 1 and 2 do not use the assumption of exponential services and apply in fact to any regenerative queueing process in which customers arrive singly and are served singly.

Observation 3 For exponential services, the long-run average number of downcrossings from j to $j-1$ per time unit equals βp_j with probability 1 for each $j \geq 1$.

The proof of this result relies heavily on the PASTA property. To make this clear, fix j and note that service completions occur according to a Poisson process with rate β as long as the server is busy. Equivalently, we can assume that an exogenous Poisson process generates events at a rate of β, where a Poisson event results in a service completion only when there are j customers present. Thus, by part (a) of Theorem 2.4.1,

$$\beta E[I_j(t)] = E[D_j(t)] \quad \text{for } t > 0 \tag{2.7.2}$$

for any $j \geq 1$, where $I_j(t)$ is defined as the amount of time that j customers are present during $(0, t]$ and $D_j(t)$ is defined as the number of downcrossings from j to $j-1$ in $(0, t]$. Letting the constant d_j denote the long-run average number of downcrossings from j to $j-1$ per time unit, we have by the renewal-reward theorem that $\lim_{t\to\infty} D_j(t)/t = d_j$ with probability 1. Similarly, $\lim_{t\to\infty} I_j(t)/t = p_j$ with probability 1. The renewal-reward theorem also holds in the expected-value version. Thus, for any $j \geq 1$,

$$\lim_{t\to\infty} \frac{E[D_j(t)]}{t} = d_j \quad \text{and} \quad \lim_{t\to\infty} \frac{E[I_j(t)]}{t} = p_j.$$

Hence relation (2.7.2) gives that $d_j = \beta p_j$ for all $j \geq 1$. By observations 1 and 2 we have $d_j = \lambda \pi_{j-1}$. This gives $\lambda \pi_{j-1} = \beta p_j$ for all $j \geq 1$, as was to be proved.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

In Chapter 3 the method of embedded Markov chains will be used to derive an explicit expression for the customer-average probabilities π_j.

EXERCISES

2.1 A street lamp is replaced by a new one upon failure and upon scheduled times $T, 2T, \ldots$. There is always a replacement at the scheduled times regardless of the age of the street lamp in use. The lifetimes of the street lamps are independent random variables and have a common Erlang $(2, \mu)$ distribution. What is the expected number of street lamps used in a scheduling interval?

2.2 The municipality of Gotham City has opened a depot for temporarily storing chemical waste. The amount of waste brought in each week has a gamma distribution with given shape parameter α and scale parameter λ. The amounts brought in during the successive weeks are independent of each other.

(a) What is the expected number of weeks until the total amount of waste in the depot exceeds the critical level L?

(b) Give an asymptotic estimate for the expected value from question (a).

2.3 Limousines depart from the railway station to the airport from the early morning till late at night. The limousines leave from the railway station with independent interdeparture times that are uniformly distributed between 10 and 20 minutes. Suppose you plan to arrive at the railway station at 3 o'clock in the afternoon. What are the estimates for the mean and the standard deviation of your waiting time at the railway station until a limousine leaves for the airport?

2.4 Consider the expression (2.1.4) for the renewal function $M(t)$.

(a) Prove that for any $k = 0, 1, \ldots$

$$\sum_{n=k+1}^{\infty} F_n(t) \leq \frac{F_k(t)F(t)}{1 - F(t)}$$

for any t with $F(t) < 1$. (*Hint*: use $P\{X_1 + \cdots + X_n \leq t\} \leq P\{X_1 + \cdots + X_k \leq t\}P\{X_{k+1} \leq t\} \cdots P\{X_n \leq t\}$.)

(b) Conclude that $M(t) < \infty$ for all $t \geq 0$.

2.5 Consider a renewal process with Erlang (r, λ) distributed interoccurrence times. Use the phase method to prove:

(a) For any $t > 0$,

$$P\{N(t) > k\} = \sum_{j=(k+1)r}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!}, \quad k = 0, 1, \ldots .$$

(b) The excess variable γ_t is Erlang (j, λ) distributed with probability

$$p_j(t) = \sum_{k=1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{kr-j}}{(kr - j)!}, \quad j = 1, \ldots, r.$$

2.6 Consider a continuous-time stochastic process $\{X(t), t \geq 0\}$ that can assume only the two states 1 and 2. If the process is currently in state i, it moves to the next state after an exponentially distributed time with mean $1/\lambda_i$ for $i = 1, 2$. The next state is state 1 with probability p_1 and state 2 with probability $p_2 = 1 - p_1$ irrespective of the past of the process.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

(a) Use the renewal-reward model to find the long-run fraction of time the process $\{X(t)\}$ is in state i for $i = 1, 2$. Does $\lim_{t\to\infty} P\{X(t) = i\}$ exist for $i = 1, 2$? If so, what is the limit?

(b) Consider a renewal process in which the interoccurrence times have an H_2 distribution with density $p_1\lambda_1 e^{-\lambda_1 t} + p_2\lambda_2 e^{-\lambda_2 t}$. Argue that

$$\lim_{t\to\infty} P\{\gamma_t > x\} = \frac{p_1\lambda_2}{p_1\lambda_2 + p_2\lambda_1} e^{-\lambda_1 x} + \frac{p_2\lambda_1}{p_1\lambda_2 + p_2\lambda_1} e^{-\lambda_2 x}, \quad x \geq 0.$$

2.7 Consider a renewal process with Erlang (r, λ) distributed interoccurrence times. Let the probability $p_j(t)$ be defined as in part (b) of Exercise 2.5. Use the renewal-reward model to argue that $\lim_{t\to\infty} p_j(t) = 1/r$ for $j = 1, \ldots, r$ and conclude that

$$\lim_{t\to\infty} P\{\gamma_t > x\} = \frac{1}{r} \sum_{j=1}^{r} \sum_{k=0}^{j-1} e^{-\lambda x} \frac{(\lambda x)^k}{k!}, \quad x \geq 0.$$

Generalize these results when the interoccurrence time is distributed as an Erlang (j, λ) random variable with probability β_j for $j = 1, \ldots, r$.

2.8 Consider the $E_r/D/\infty$ queueing system with infinitely many servers. Customers arrive according to a renewal process in which the interoccurence times have an Erlang (r,λ) distribution and the service time of each customer is a constant D. Each newly arriving customer gets immediately assigned a free server. Let $p_n(t)$ denote the probability that n servers will be busy at time t. Use an appropriate conditioning argument to verify that

$$\lim_{t\to\infty} p_0(t) = \frac{1}{r} \sum_{j=1}^{r} \sum_{k=0}^{j-1} e^{-\mu D} \frac{(\mu D)^k}{k!}$$

$$\lim_{t\to\infty} p_n(t) = \frac{1}{r} \sum_{j=1}^{r} \sum_{k=0}^{r-1} e^{-\mu D} \frac{(\mu D)^{r-j+1+(n-1)r+k}}{(r-j+1+(n-1)r+k)!}, \quad n \geq 1.$$

(*Hint*: the only customers present at time t are those customers who have arrived in $(t - D, t]$.)

2.9 The lifetime of a street lamp has a given probability distribution function $F(x)$ with probability density $f(x)$. The street lamp is replaced by a new one upon failure or upon reaching the critical age T, whichever occurs first. A cost of $c_f > 0$ is incurred for each failure replacement and a cost of $c_p > 0$ for each preventive replacement, where $c_p < c_f$. The lifetimes of the street lamps are independent of each other.

(a) Define a regenerative process and specify its regeneration epochs.

(b) Show that the long-run average cost per time unit under the age-replacement rule equals $g(T) = [c_p + (c_f - c_p)F(T)]/\int_0^T \{1 - F(x)\}\,dx$.

(c) Verify that the optimal value of T satisfies $g(T) = (c_f - c_p)r(T)$, where $r(x)$ is the failure rate function of the lifetime.

2.10 Consider the $M/G/\infty$ queue from Section 1.1.3 again. Let the random variable L be the length of a busy period. A busy period begins when an arrival finds the system empty and ends when there are no longer any customers in the system. Use the result (2.2.1) to argue that $E(L) = (e^{\lambda\mu} - 1)/\lambda$.

2.11 Consider an electronic system having n identical components that operate independently of each other. If a component breaks down, it goes immediately into repair. There are ample

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

repair facilities. Both the running times and the repair times are sequences of independent and identically distributed random variables. It is also assumed that these two sequences are independent of each other. The running time has a positive density on some interval. Denote by α the mean running time and by β the mean repair time.

(a) Prove that

$$\lim_{t\to\infty} P\{k \text{ components are in repair at time } t\} = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \ldots, n$, where $p = \beta/(\alpha + \beta)$.

(b) Argue that the limiting distribution in (a) becomes a Poisson distribution with mean $\lambda\beta$ when $n \to \infty$ and $1/\alpha \to 0$ such that n/α remains equal to the constant λ. Can you explain the similarity of this result with the insensitivity result (1.1.6) for the $M/G/\infty$ queue in Section 1.1.3?

2.12 A production process in a factory yields waste that is temporarily stored on the factory site. The amounts of waste that are produced in the successive weeks are independent and identically distributed random variables with finite first two moments μ_1 and μ_2. Opportunities to remove the waste from the factory site occur at the end of each week. The following control rule is used. If at the end of a week the total amount of waste present is larger than D, then all the waste present is removed; otherwise, nothing is removed. There is a fixed cost of $K > 0$ for removing the waste and a variable cost of $v > 0$ for each unit of waste in excess of the amount D.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run average cost per time unit.

(c) Assuming that D is sufficiently large compared to μ_1, give an approximate expression for the average cost.

2.13 At a production facility orders arrive according to a renewal process with a mean interarrival time $1/\lambda$. A production is started only when N orders have accumulated. The production time is negligible. A fixed cost of $K > 0$ is incurred for each production set-up and holding costs are incurred at the rate of hj when j orders are waiting to be processed.

(a) Define a regenerative stochastic process and identify its regeneration epochs.

(b) Determine the long-run average cost per time unit.

(c) What value of N minimizes the long-run average cost per time unit?

2.14 Consider again Exercise 2.13. Assume now that it takes a fixed set-up time T to start a production. Any new order that arrives during the set-up time is included in the production run. Answer parts (a) and (b) from Exercise 2.13 for the particular case that the orders arrive according to a Poisson process with rate λ.

2.15 How do you modify the expression for the long-run average cost per time unit in Exercise 2.14 when it is assumed that the set-up time is a random variable with finite first two moments?

2.16 Consider Example 1.3.1 again. Assume that a fixed cost of $K > 0$ is incurred for each round trip and that a fixed amount $R > 0$ is earned for each passenger.

(a) Define a regenerative stochastic process and identify its regeneration epochs.

(b) Determine the long-run average net reward per time unit.

(c) Verify that the average reward is maximal for the unique value of T satisfying the equation $e^{-\mu T}(R\lambda T + R\lambda/\mu) = R\lambda/\mu - K$ when $R\lambda/\mu > K$.

2.17 Passengers arrive at a bus stop according to a Poisson process with rate λ. Buses depart from the stop according to a renewal process with interdeparture time A. Using renewal-reward processes, prove that the long-run average waiting time per passenger equals $E(A^2)/2E(A)$. Specify the regenerative process you need to prove this result. Can you give

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

a heuristic explanation of why the answer for the average waiting time is the same as the average residual life in a renewal process?

2.18 Consider a renewal process in which the interoccurrence times have a positive density on some interval. For any time t let the age variable δ_t denote the time elapsed since the last occurrence of an event. Use the renewal-reward model to prove that $\lim_{t\to\infty} E(\delta_t) = \mu_2/2\mu_1$, where μ_k is the kth moment of the interoccurrence times. (*Hint*: assume a cost at rate x when a time x has elapsed since the last occurrence of an event.)

2.19 A common car service between cities in Israel is a sheroot. A sheroot is a seven-seat cab that leaves from its stand as soon as it has collected seven passengers. Suppose that potential passengers arrive at the stand according to a Poisson process with rate λ. An arriving person who sees no cab at the stand goes elsewhere and is lost for the particular car service. Empty cabs pass the stand according to a Poisson process with rate μ. An empty cab stops only at the stand when there is no other cab.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run fraction of time there is no cab at the stand and determine the long-run fraction of customers who are lost. Explain why these two fractions are equal to each other.

2.20 Big Jim, a man of few words, runs a one-man business. This business is called upon by loan sharks to collect overdue loans. Big Jim takes his profession seriously and accepts only one assignment at a time. The assignments are classified by Jim into n different categories $j = 1, \ldots, n$. An assignment of type j takes him a random number of τ_j days and gives a random profit of ξ_j dollars for $j = 1, \ldots, n$. Assignments of the types $1, \ldots, n$ arrive according to independent Poisson processes with respective rates $\lambda_1, \ldots, \lambda_n$. Big Jim, once studying at a prestigious business school, is a muscleman with brains. He has decided to accept those type j assignments for which $E(\xi_j)/E(\tau_j)$ is at least g^* dollars per day for a carefully chosen value of g^* (in Exercise 7.4 you are asked to use Markov decision theory to determine g^*). Suppose that Big Jim only accepts type j assignments for $j = 1, \ldots, n_0$. An assignment can only be accepted when Big Jim is not at work on another assignment. Assignments that are refused are handled by a colleague of Big Jim.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run average pay-off per time unit for Big Jim.

(c) Determine the long-run fraction of time Big Jim is at work and the long-run fraction of the assignments of the types $1, \ldots, n_0$ that are not accepted. Explain why these two fractions are equal to each other.

2.21 Consider the $(S-1, S)$ inventory model with back ordering from Section 1.1.3. What is the long-run fraction of customer demand that is back ordered? What is the long-run average amount of time a unit is kept in stock?

2.22 Consider a machine whose state deteriorates through time. The state of the machine is inspected at fixed times $t = 0, 1, \ldots$. In each period between two successive inspections the machine incurs a random amount of damage. The amounts of damage accumulate. The amounts of damage incurred in the successive periods are independent random variables having a common exponential distribution with mean $1/\alpha$. A compulsory repair of the machine is required when an inspection reveals a cumulative amount of damage larger than a critical level L. A compulsory repair involves a fixed cost of $R_c > 0$. A preventive repair at a lower cost of $R_p > 0$ is possible when an inspection reveals a cumulative amount of damage below or at the level L. The following control limit rule is used. A repair is done at each inspection that reveals a cumulative amount of damage larger than some repair limit z with $0 \le z < L$. It is assumed that each repair takes a negligible time and that after each repair the machine is as good as new.

(a) Define a regenerative process and identify its regeneration epochs.

(b) What is the expected number of periods between two successive repairs? What is the

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

probability that a repair involves the high repair cost R_c? Give the long-run average cost per time unit.

(c) Verify that the average cost is minimal for the unique solution z to the equation $\alpha z \exp[-\alpha(L-z)] = R_p/(R_c - R_p)$ when $\alpha L > R_p/(R_c - R_p)$.

2.23 A group of N identical machines is maintained by a single repairman. The machines operate independently of each other and each machine has a constant failure rate μ. Repair is done only if the number of failed machines has reached a given critical level R with $1 \leq R \leq N$. Then all failed machines are repaired simultaneously. Any repair takes a negligible time and a repaired machine is again as good as new. The cost of the simultaneous repair of R machines is $K + cR$, where $K, c > 0$. Also there is an idle-time cost of $\alpha > 0$ per time unit for each failed machine.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run average cost per time unit.

2.24 The following control rule is used for a slow-moving expensive product. No more than one unit of the product is kept in stock. Each time the stock drops to zero a replenishment order for one unit is placed. The replenishment lead time is a positive constant L. Customers asking for the product arrive according to a renewal process in which the interarrival times are Erlang (r, λ) distributed. Each customer asks for one unit of the product. Each demand occurring while the system is out of stock is lost.

(a) Define a regenerative process and identify its regeneration epochs.

(b) Determine the long-run fraction of demand that is lost.

(c) Determine the long-run fraction of time the system is out of stock. (*Hint:* use part (b) of Exercise 2.5.)

2.25 Jobs arrive at a station according to a renewal process. The station can handle only one job at a time, but has no buffer to store other jobs. An arriving job that finds the station busy is lost. The handling time of a job has a given probability density $h(x)$. Use renewal-reward theory to verify for this loss system that the long-run fraction of jobs that are rejected is given by $\int_0^\infty M(x)h(x)\,dx$ divided by $1 + \int_0^\infty M(x)h(x)\,dx$, where $M(x)$ is the renewal function in the renewal process describing the arrival of jobs. What is the long-run fraction of time that the station is busy? Simplify the formulas for the cases of deterministic and Poisson arrivals.

2.26 Use the renewal-reward theorem to prove relation (2.3.3) when customers arrive according to a renewal process and the stochastic processes $\{L(t)\}$ and $\{U_n\}$ regenerate themselves each time an arriving customer finds the system empty, where the cycle lengths have finite expectations. For ease assume the case of an infinite-capacity queue. Use the following relations:

(i) the long-run average reward earned per time unit = (the expected reward earned in one cycle)/(expected length of one cycle),

(ii) the long-run average amount paid per customer = (the expected amount earned in one cycle)/(expected number of arrivals in one cycle),

(iii) the long-run average arrival rate = (expected number of arrivals in one cycle)/(expected length of one cycle).

2.27 Let $\{X(t), t \geq 0\}$ be a continuous-time regenerative stochastic process whose state space is a subset of the non-negative reals. The cycle length is assumed to have a finite expectation. Denote by $\overline{P}(y)$ the long-run fraction of time that the process $\{X(t)\}$ takes on a value larger than y. Use the renewal-reward theorem to prove that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t X(u)\,du = \int_0^\infty \overline{P}(y)\,dy \quad \text{with probability 1.}$$

2.28 Consider a queueing system in which the continuous-time process $\{L(t)\}$ describing the number of customers in the system is regenerative, where the cycle length has a finite

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

expectation. Let p_j denote the long-run fraction of time that j customers are in the system and let L denote the long-run average number of customers in the system. Apply the result of Exercise 2.27 to conclude that $L = \sum_{j=1}^{\infty} j p_j$.

2.29 Verify that the Pollaczek–Khintchine formula for the average waiting time in the $M/G/1$ queue can also be written as

$$W_q = (1 - c_S^2) W_q(\text{det}) + c_S^2 W_q(\text{exp}).$$

This interpolation formula is very useful and goes back to Cox (1955).

2.30 A professional cleaner in the harbour of Rotterdam is faced with the decision to acquire a new clean installation for oil tankers. Oil tankers requiring a clean arrive according to a Poisson process with rate λ. The amount of time needed to clean a tanker has a given probability distribution with mean α and standard deviation β when the standard Fadar installation is used. Cleaning costs at a rate of $c > 0$ are incurred for each time unit this installation is in use. However, it is also possible to buy another installation. An installation that works z times as fast as the standard Fadar installation involves cleaning costs at a rate of cz^2 per time unit. In addition to the cleaning costs, a holding cost at rate of $h > 0$ is incurred for each tanker in the harbour. What is the long-run average cost per time unit as function of z? Assume that the cleaning installation can handle only one tanker at a time and assume that the cleaner has ample berths for tankers.

2.31 Liquid is put into an infinite-capacity buffer at epochs generated by a Poisson process with rate λ. The successive amounts of liquid that are put in the buffer are independent and identically distributed random variables with finite first two moments μ_1 and μ_2. The buffer is emptied at a constant rate of $\sigma > 0$ whenever it is not empty. Use the PASTA property to give an expression for the long-run average buffer content.

2.32 Consider the $M/G/1$ queue with two types of customers. Customers of the types 1 and 2 arrive according to independent Poisson processes with respective rates λ_1 and λ_2. The service times of the customers are independent of each other, where the service times of type i customers are distributed as the random variable S_i having finite first two moments. Customers of type 1 have priority over customers of type 2 when the server is ready to start a new service. It is not allowed to interrupt the service of a type 2 customer when a higher-priority customer arrives. This queueing model is called the *non-pre-emptive priority* $M/G/1$ queue. Letting $\rho_i = \lambda_i E(S_i)$, it is assumed that $\rho_1 + \rho_2 < 1$.

(a) Use Little's formula to argue that the long-run fraction of time the server is servicing type i customers equals ρ_i for $i = 1, 2$. What is the long-run fraction of customers finding the server servicing a type i customer upon arrival?

(b) Extend the heuristic derivation of the Pollaczek–Khintchine formula to show

$$W_{q1} = \frac{\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)}{2(1 - \rho_1)} \quad \text{and} \quad W_{q2} = \frac{\lambda_1 E(S_1^2) + \lambda_2 E(S_2^2)}{2(1 - \rho_1)(1 - \rho_1 - \rho_2)},$$

where W_{qi} is defined as the long-run average waiting time in queue per type i customer for $i = 1, 2$.

(c) Use Little's formula to give a direct argument for the result that the overall average waiting time $W_{q1}\lambda_1/(\lambda_1 + \lambda_2) + W_{q2}\lambda_2/(\lambda_1 + \lambda_2)$ per customer is the same as the average waiting time per customer in the $M/G/1$ queue in which customers are served in order of arrival (view the non-pre-emptive priority rule as a rule that merely changes the order in which the customers are served).

2.33 Customers arrive at a single-server station according to a Poisson process with rate λ. A customer finding the server idle upon arrival gets served immediately, otherwise the customer enters a so-called orbit. A customer in orbit tries whether the server is idle after an

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

exponentially distributed time with mean $1/\nu$. If the server is idle, the customer gets served, otherwise the customer returns to orbit and tries again after an exponentially distributed time until the server is found free. The customers in orbit act independently of each other. The service times of the customers are independent random variables having the same general probability distribution. Letting the random variable S denote the service time of a customer, it is assumed that $\rho = \lambda E(S)$ is less than 1. For this model, known as the $M/G/1$ queue with *retrials*, define $L(t)$ as the number of customers in the system (service station plus orbit) at time t and define Q_n as the number of customers in orbit just after the nth service completion. Let $p_j = \lim_{t\to\infty} P\{L(t) = j\}$ and $q_j = \lim_{n\to\infty} P\{Q_n = j\}$ for $j \geq 0$.

(a) Use an up- and downcrossing argument to argue that $p_j = q_j$ for all $j \geq 0$.

(b) Letting $Q(z) = \sum_{j=0}^{\infty} q_j z^j$, prove that

$$Q(z) = A(z)\{\lambda R(z) + \nu R'(z)\},$$

where $A(z)$ is the generating function of the number of new customers arriving during the service time S and $R(z)$ is defined by $R(z) = \sum_{j=0}^{\infty} z^j q_j/(\lambda + j\nu)$. (*Hint*: under the condition that $Q_{n-1} = i$ it holds that $Q_n = Q_{n-1} + C_n$ with probability $\lambda/(\lambda + i\nu)$ and $Q_n = Q_{n-1} - 1 + C_n$ with probability $i\nu/(\lambda + i\nu)$, where C_n denotes the number of new customers arriving at the nth service time.)

(c) Prove that

$$Q(z) = \frac{(1-\rho)(1-z)A(z)}{A(z)-z} \exp\left[\frac{\lambda}{\nu}\int_1^z \frac{1-A(u)}{A(u)-u}\,du\right].$$

(*Hint*: use that $Q(z) = \lambda R(z) + \nu z R'(z)$, which follows directly from the definition of $R(z)$.)

(d) Show that the long-run average number of customers in the system is given by

$$L = \rho + \frac{\lambda^2 E(S^2)}{2(1-\rho)} + \frac{\lambda^2 E(S)}{\nu(1-\rho)}.$$

Retrial queues are in general much more difficult to analyse than queues without retrials. The Laplace transform for the waiting-time distribution in the $M/G/1$ queue with retrials is very complex; see also Artalejo *et al*. (2002).

2.34 Consider again the production system from Section 2.6 except that the system is now controlled in a different way when it becomes idle. Each time the production facility becomes empty of orders, the facility is used during a period of fixed length T for some other work in order to utilize the idle time. After this vacation period the facility is reactivated for servicing the orders only when at least one order is present; otherwise the facility is used again for some other work during a vacation period of length T. This utilization of idle time is continued until at least one order is present after the end of a vacation period. This control policy is called the T-policy. The cost structure is the same as in Section 2.6. Use renewal-reward theory to show that $K(1-\lambda\mu_1)(1-e^{-\lambda T})/T + \frac{1}{2}h\lambda T + \frac{1}{2}h\lambda^2\mu_2/(1-\lambda\mu_1)$ gives the long-run average cost per time unit under a T-policy.

2.35 Suppose that, at a communication channel, messages of types 1 and 2 arrive according to independent Poisson processes with respective rates λ_1 and λ_2. Messages of type 1 finding the channel occupied upon arrival are lost, whereas messages of type 2 are temporarily stored in a buffer and wait until the channel becomes available. The channel can transmit only one message at a time. The transmission time of a message of type i has a general probability distribution with mean μ_i and the transmission times are independent of each other. It is assumed that $\lambda_2\mu_2 < 1$. Use the renewal-reward theorem to prove that the long-run fraction of time the channel is busy equals $(\rho_1 + \rho_2)/(1 + \rho_1)$, where $\rho_i = \lambda_i\mu_i$ for $i = 1, 2$.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

(*Hint*: use results from Section 2.6 to obtain the expected amount of time elapsed between two arrivals finding the channel free.)

BIBLIOGRAPHIC NOTES

The very readable monograph of Cox (1962) contributed much to the popularization of renewal theory. A good account of renewal theory can also be found in the texts Ross (1996) and Wolff (1989). A basic paper on renewal theory and regenerative processes is that of Smith (1958), a paper which recognized the usefulness of renewal-reward processes in the analysis of applied probability problems. The book of Ross (1970) was influential in promoting the application of renewal-reward processes. The renewal-reward model has many applications in inventory, queueing and reliability. The illustrative queueing example from Section 2.6 is taken from the paper of Yadin and Naor (1963), which initiated the study of control rules for queueing systems. Example 2.2.3 is adapted from the paper of Vered and Yechiali (1979).

The first rigorous proof of $L = \lambda W$ was given by Little (1961) under rather strong conditions; see also Jewell (1967). Under very weak conditions a sample-path proof of $L = \lambda W$ was given by Stidham (1974). The important result that Poisson arrivals see time averages was taken for granted by earlier practitioners. A rigorous proof was given in the paper of Wolff (1982). The derivation of the Laplace transform of the waiting-time distribution in the $M/G/1$ queue is adapted from Cohen (1982) and the relation between this transform and the generating function of the queue size comes from Haji and Newell (1971).

REFERENCES

Artalejo, J.R., Falin, G.I. and Lopez-Herrero, M.J. (2002) A second order analysis of the waiting time in the $M/G/1$ retrial queue. *Asia-Pacific J. Operat. Res.*, **19**, 131–148.

Cohen, J.W. (1982) *The Single Server Queue*, 2nd edn. North-Holland, Amsterdam.

Cox, D.R. (1955) The statistical analysis of congestion. *J. R. Statist. Soc. A.*, **118**, 324–335.

Cox, D.R. (1962) *Renewal Theory*. Methuen, London.

Haji, R. and Newell, G.F. (1971) A relation between stationary queue and waiting-time distribution. *J. Appl. Prob.*, **8**, 617–620.

Jewell, W.S. (1967) A simple proof of $L = \lambda W$. *Operat. Res.*, **15**, 1109–1116.

Keilson, J. (1979) *Markov Chain Models—Rarity and Exponentiality*. Springer-Verlag, Berlin.

Little, J.D.C. (1961) A proof for the queueing formula $L = \lambda W$. *Operat. Res.*, **9**, 383–387.

Miller, D.R. (1972) Existence of limits in regenerative processes. *Ann. Math. Statist.*, **43**, 1275–1282.

Ross, S.M. (1970) *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.

Ross, S.M. (1996) *Stochastic Processes*, 2nd. edn. John Wiley & Sons, Inc., New York.

Smith, W.L. (1958) Renewal theory and its ramifications. *J. R. Statist. Soc. B*, **20**, 243–302.

Solovyez, A.D. (1971) Asymptotic behaviour of the time of first occurrence of a rare event in a regenerating process. *Engineering Cybernetics*, **9**, 1038–1048.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Stidham, S. Jr (1974) A last word on $L = \lambda W$. *Operat. Res.*, **22**, 417–421.

Takács, L. (1962) *Introduction to the Theory of Queues*. Oxford University Press, New York.

Vered, G. and Yechiali, U. (1979) Optimal structures and maintenance policies for PABX power systems. *Operat. Res.*, **27**, 37–47.

Wolff, R.W. (1982) Poisson arrivals see time averages. *Operat. Res.*, **30**, 223–231.

Wolff, R.W. (1989) *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs NJ.

Yadin, M. and Naor, P. (1963) Queueing systems with removable service station. *Operat. Res. Quart.*, **14**, 393–405.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.

Copyright © 2003. John Wiley & Sons, Incorporated. All rights reserved.

Tijms, Henk C.. A First Course in Stochastic Models, John Wiley & Sons, Incorporated, 2003. ProQuest Ebook Central,
http://ebookcentral.proquest.com/lib/sutd/detail.action?docID=158138.
Created from sutd on 2020-03-19 04:41:23.