

Analysis of Queueing Systems with Sample Paths and Simulation

Nicky D. van Foreest

October 6, 2017

Contents

1	Introduction	5
2	Single-Station Queueing Systems	9
2.1	Exponential Distribution	9
2.2	Poisson Distribution	20
2.3	Kendall's Notation to Characterize Queueing Processes	29
2.4	Construction of Discrete-Time Queueing Processes	31

1 Introduction

Motivation and Examples Queueing systems abound, and the analysis and control of queueing systems are major topics in the control, performance evaluation and optimization of production and service systems.

At my local supermarket, for instance, any customer that joins a queue of 4 or more customers get his/her shopping for free. Of course, there are some constraints: at least one of the cashier facilities has to be unoccupied by a server and the customers in queue should be equally divided over the cashiers that are open. (And perhaps there are some further rules, of which I am unaware.) The manager that controls the occupation of the cashier positions is focused on keeping $\pi(4) + \pi(5) + \dots$, i.e., the fraction of customers that see upon arrival a queue length longer or equal than 4, very small. In a sense, this is easy enough: just hire many cashiers. However, the cost of personnel may then outweigh the yearly average cost of paying the customer penalties. Thus, the manager's problem becomes to plan and control the service capacity in such a way that both the penalties and the personnel cost are small.

Fast food restaurants also deal with many interesting queueing situations. Consider, for instance, the making of hamburgers. Typically, hamburgers are made-to-stock, in other words, they are prepared before the actual demand has arrived. Thus, hamburgers in stock can be interpreted as customers in queue waiting for service, where the service time is the time between the arrival of two customers that buy hamburgers. The hamburgers have a typical lifetime, and they have to be scrapped if they remain on the shelf longer than some amount of time. Thus, the waiting time of hamburgers has to be closely monitored. Of course, it is easy to achieve zero scrap cost, simply by keeping no stock at all. However, to prevent lost-sales it is very important to maintain a certain amount of hamburgers on stock. Thus, the manager has to balance the scrap cost against the cost of lost sales. In more formal terms, the problem is to choose a policy to prepare hamburgers such that the cost of excess waiting time (scrap) is balanced against the cost of an empty queue (lost sales).

Service systems, such as hospitals, call centers, courts, and so on, have a certain capacity available to serve customers. The performance of such systems is, in part, measured by the total number of jobs processed per year and the fraction of jobs processed within a certain time between receiving and closing the job. Here the problem is to organize the capacity such that the sojourn time, i.e., the typical time a job spends in the system, does not exceed some threshold, and such that the system achieves a certain throughput, i.e., jobs served per year.

Clearly, all the above systems can be seen as queueing systems that have to be monitored and controlled to achieve a certain performance. The performance analysis of such systems can, typically, be characterized by the following performance measures:

1. The fraction of time $p(n)$ that the system contains n customers. In particular, $1 - p(0)$, i.e., the fraction of time the system contains jobs, is important, as this is a measure of the time-average occupancy of the servers, hence related to personnel cost.
2. The fraction of customers $\pi(n)$ that 'see upon arrival' the system with n customers. This measure relates to customer perception and lost sales, i.e., fractions of arriving customers

1 Introduction

that do not enter the system.

3. The average, variance, and/or distribution of the waiting time.
4. The average, variance, and/or distribution of the number of customers in the system.

Here the system can be anything that is capable of holding jobs, such as a queue, the server(s), an entire court house, patients waiting for an MRI scan in a hospital, and so on.

It is important to realize that a queueing system can, typically, be decomposed into *two subsystems*, the queue itself and the service system. Thus, we are concerned with three types of waiting: waiting in queue, i.e., *queueing time*, waiting while being in service, i.e., the *service time*, and the total waiting time in the system, i.e., the *sojourn time*.

Organization In these notes we will be primarily concerned with making models of queueing systems such that we can compute or estimate the above performance measures. Part of our work is to derive analytic models. The benefit of such models is that they offer structural insights into the behavior of the system and scaling laws, such as that the average waiting time scales (more or less) linearly in the variance of the service times of individual customers. However, these models have severe shortcomings when it comes to analyzing real queueing systems, in particular when particular control rules have to be assessed. Consider, for example, the service process at a check-in desk of KLM. Business customers and economy customers are served by two separate queueing systems. The business customers are served by one server, server A say, while the economy class customers by three servers, say. What would happen to the sojourn time of the business customers if server A would be allowed to serve economy class customers when the business queue is empty? For the analysis of such cases simulation is a very useful and natural approach.

In the first part of these notes we concentrate on the analysis of *sample paths of a queueing process*. We assume that a typical sample path captures the ‘normal’ stochastic behavior of the system. This sample-path approach has two advantages. In the first place, most of the theoretical results follow from very concrete aspects of these sample paths. Second, the analysis of sample-paths carries over right away to simulation. In fact, simulation of a queueing system offers us one (or more) sample paths, and based on such sample paths we derive behavioral and statistical properties of the system. Thus, the performance measures defined for sample paths are precisely those used for simulation. Our aim is not to provide rigorous proofs for all results derived below; for the proofs and further background discussion we refer to ?. As a consequence we tacitly assume in the remainder that results derived from the (long-run) analysis of a particular sample path are equal to their ‘probabilistic counterpart’.

In the second part we construct algorithms to analyze open and closed queueing networks. Many of the sample path results developed for the single-station case can be applied to these networks. As such, theory, simulation and algorithms form a nicely round out part of work. For this part we refer to book of Prof. Zijm; the present set of notes augment the discussion there.

Exercises I urge you to make *all* exercises in this set of notes. Many exercises require many of the tools you learned previously in courses on calculus, probability, and linear algebra. Here you can see them applied. Moreover, many of these tools will be useful for other, future, courses. Thus, the investments made here will pay off for the rest of your (student) career. Moreover, the exercises are meant to *illustrate* the material and to force you to *think* about it. Thus, the main text does not contain many examples; the exercises form the examples.

You'll notice that many of these problems are quite difficult, often not because the problem itself is difficult, but because you need to combine a substantial amount of knowledge all at the same time. All this takes time and effort. Next to this, I did not include the exercises with the intention that you would find them easy. The problems should be doable, but hard.

The solution manual is meant to prevent you from getting stuck and to help you increase your knowledge of probability, linear algebra, programming (analysis with computer support), and queueing in particular. Thus, read the solutions very carefully.

As a guideline to making the exercises I recommend the following approach. First read the notes. Then attempt to make a exercises for 10 minutes or so by yourself. If by that time you have not obtained a good idea on how to approach the problem, check the solution manual. Once you have understood the solution, try to repeat the arguments *with the solution manual closed*.

Symbols The meaning of the symbols in the margin of pages are as follows:

- The symbol in the margin means that you have to memorize the *emphasized concepts*.
- The symbol in the margin means that this question has a *hint*.
- The symbol in the margin means that this question or its solution requires still some *work on my part*; you can skip it.



Acknowledgments I would like to acknowledge dr. J.W. Nieuwenhuis for our many discussions on the formal aspects of queueing theory and prof. dr. W.H.M. Zijm for allowing me to use the first few chapters of his book. Finally, without ? I could not have written these notes.

2 Single-Station Queueing Systems

In this chapter, we start with a discussion of the exponential distribution and the related Poisson process, as these concepts are perhaps the most important building blocks of queueing theory. With these concepts we can specify the arrival and service processes of customers, so that we can construct queueing processes and define performance measures to provide insight into the (transient and average) behavior of queueing processes. As it turns out, these constructions can be easily implemented as computer programs, thereby allowing to use simulation to analyze queueing systems. We then continue with developing models for various single-station queueing systems in steady-state, which is, in a sense to be discussed later, the long-run behavior of a stochastic system.¹ In the analysis we use sample-path arguments to count how often certain events occur as functions of time. Then we define probabilities in terms of limits of fractions of these counting processes. Another useful aspect of sample-path analysis is that the definitions for the performance measures are entirely constructive, hence by leaving out the limits, they provide expressions that can be right away used in statistical analysis of (simulations of) queueing systems. Level-crossing arguments will be of particular importance as we use these time and again to develop recursions by which we can compute steady-state probabilities of the queue length or waiting time process.

2.1 Exponential Distribution

As we will see in the sections to come, the modeling and analysis of any queueing system involves the specification of the (probability) distribution of the time between consecutive arrival epochs of jobs, or the specification of the distribution of the number of jobs that arrive in a certain interval. For the first case, the most common distribution is the exponential distribution, while for the second it is the Poisson distribution. For these reasons we start our discussion of the analysis of queueing system with these two exceedingly important distributions. In the ensuing sections we will use these distributions time and again.

As mentioned, one of the most useful models for the interarrival times of jobs assumes that the sequence $\{X_i\}$ of interarrival times is a set of *independent and identically distributed (i.i.d.)* random variables. Let us write X for the generic random time between two successive arrivals. For many queueing systems, measurements of the interarrival times between consecutive arrivals show that it is reasonable to model an interarrival X as an *exponentially distributed* random variable, i.e.,

$$P\{X \leq t\} = 1 - e^{-\lambda t} := G(t)$$

The constant λ is often called the *rate*. The reason behind this will be clarified once we relate the exponential distribution to the Poisson process in Section 2.2. In the sequel we often write $X \sim \exp(\lambda)$ to mean that X is exponentially distributed with rate λ .

Let us show with simulation how the exponential distribution originates. Consider N people that regularly visit a shop. We assume that we can characterize the interarrival times $\{X_k^i, k =$

¹This statement is, admittedly, vague, to say the least.

2 Single-Station Queueing Systems

$1, 2, \dots\}$ of customer i by some distribution function, for instance the uniform distribution. Then, with $A_0^i = 0$ for all i ,

$$A_k^i = A_{k-1}^i + X_k^i = \sum_{j=1}^n X_j^i,$$

is the arrival moment of the k th visit of customer i . Now the shop owner ‘sees’ the superposition of the arrivals of all customers. One way to compute the arrival moments of all customers together as seen by the shop is to put all the numbers $\{A_k^i, k = 1, \dots, n, i = 1, \dots, N\}$ into one set, and sort these numbers in increasing order. This results in the (sorted) set of arrival times $\{A_k, k = 1, 2, \dots\}$ at the shop, and then

$$X_k = A_k - A_{k-1},$$

with $A_0 = 0$, must be the interarrival time between the $k - 1$ th and k th visit to the shop. Thus, starting from interarrival times of individual customers we can construct interarrival times as seen by the shop.

To plot the *empirical distribution function*, or the histogram, of the interarrival times at the shop, we need to count the number of interarrival times smaller than time t for any t . For this, we introduce the *indicator function*:

$$\mathbb{1}_A = \begin{cases} 1, & \text{if the event } A \text{ is true,} \\ 0, & \text{if the event } A \text{ is false.} \end{cases}$$

With the indicator function we define the empirical distribution of the interarrival times for a simulation with a total of $n \cdot N$ as

$$P_{nN}\{X \leq t\} = \frac{1}{nN} \sum_{k=1}^{nN} \mathbb{1}_{X_k \leq t},$$

where $\mathbb{1}_{X_k \leq t} = 1$ if $X_k \leq t$ and $\mathbb{1}_{X_k \leq t} = 0$ if $X_k > t$.

Let us now compare the probability density as obtained for several simulation scenarios to the density of the exponential distribution, i.e., to $\lambda e^{-\lambda t}$. As a first example, take $N = 1$ customer and let the computer generate $n = 100$ uniformly distributed numbers on the set $[4, 6]$. Thus, the time between two visits of this customer is somewhere between 4 and 6 hours, and the average interarrival times $E\{X\} = 5$. In a second simulation we take $N = 3$ customers, and in the third, $N = 10$ customers. The empirical distributions are shown, from left to right, in the three panels in Figure 2.1. The continuous curve is the graph of $\lambda e^{-\lambda x}$ where $\lambda = N/E\{X\} = N/5$. (In Eq. (??) we show that when 1 person visits the shop with an average interarrival time of 5 hours, it must be that the arrival rate is $1/E\{X\} = 1/5$. Hence, when N customers visit the shop, each with an average interarrival time of 5 hours, the total arrival rate as seen by the shop must be $N/5$.) As a second example, we extend the simulation to $n = 1000$ visits to the shop, see Figure 2.2. In the third example we take the interarrival times to be normally distributed times with mean 5 and $\sigma = 1$, see Figure 2.3.

As the graphs show, even when the customer population consists of 10 members, each visiting the shop with an interarrival time that is quite ‘far’ from exponential, the distribution of the interarrival times as observed by the shop is very well approximated by an exponential distribution. Thus, for a real shop, with many thousands of customers, or a hospital, call center, in fact for nearly every system that deals with random demand, it seems reasonable to use the exponential distribution to model interarrival times. In conclusion, the main conditions to use

2.1 Exponential Distribution

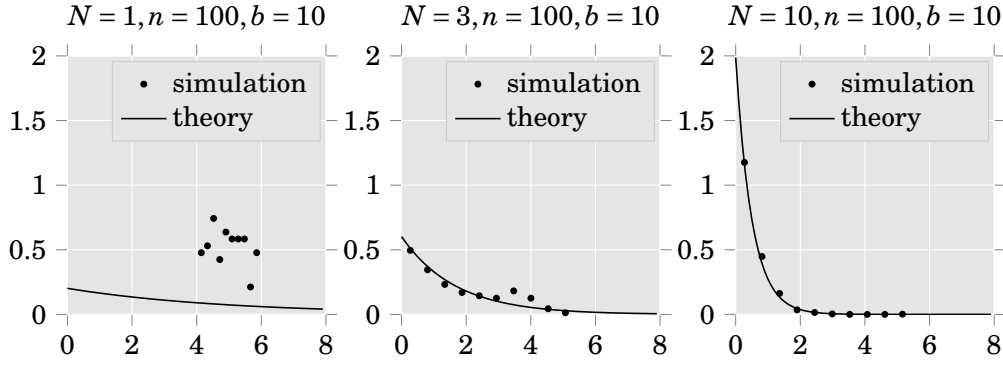


Figure 2.1: The interarrival process as seen by the shop owner. Observe that the density $\lambda e^{-\lambda x}$ intersects the y -axis at level $N/5$, which is equal to the arrival rate when N persons visit the shop. The parameter $n = 100$ is the simulation length, i.e., the number of visits per customer, and $b = 10$ is number of bins to collect the data.

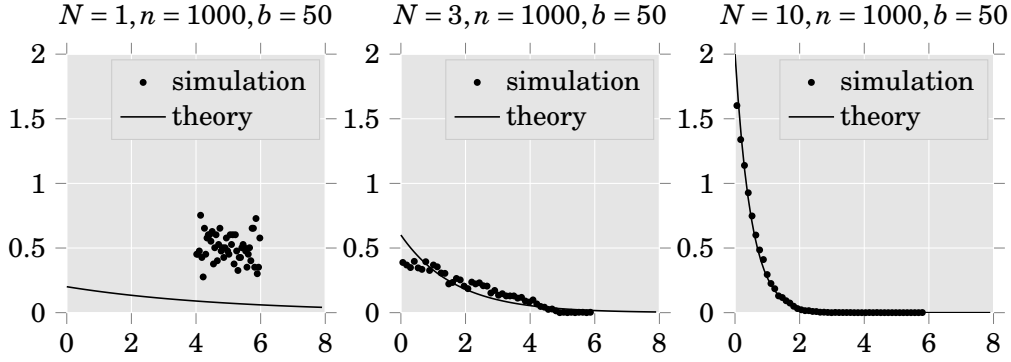


Figure 2.2: Each of the N customer visits the shop with exponentially distributed interarrival times, but now the number of visits is $n = 1000$.

an exponential distribution are: 1) arrivals have to be drawn from a large population, and 2) each of the arriving customers decides, independent of the others, to visit the system.

Another, but theoretical, reason to use the exponential distribution is that an exponentially distributed random variable is *memoryless*, that is, X is memoryless if it satisfies the property that

$$P\{X > t + h | X > t\} = P\{X > h\}.$$

In words, the probability that X is larger than some time $t + h$, conditional on it being larger than a time t , is equal to the probability that X is larger than h . Thus, no matter how long we have been waiting for the next arrival to occur, the probability that it will occur in the next h seconds remains the same. This property seems to be vindicated also in practice: suppose that a patient with a broken arm just arrived at the emergency room of a hospital, what does that tell us about the time the next patient will be brought in? Not much, as most of us will agree.

It can be shown that only exponential random variables have the memoryless property. The proof of this fact requires quite some work; we refer the reader to the literature if s/he want to check this, see e.g. [?], Appendix 3.

2 Single-Station Queueing Systems

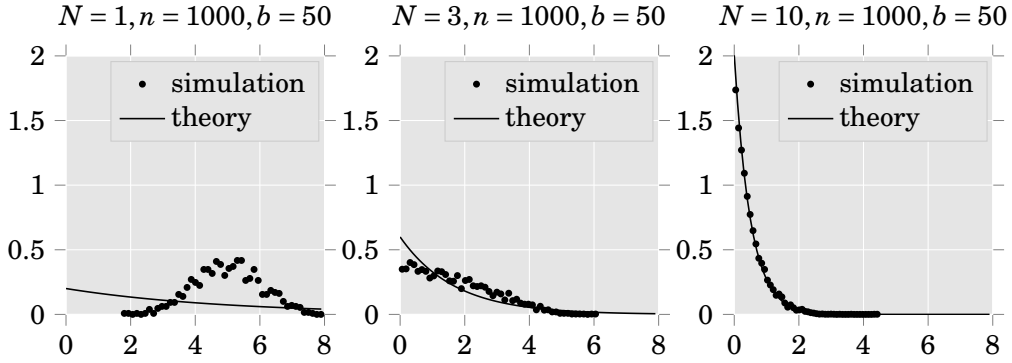


Figure 2.3: Each of the N customer visits the shop with normally distributed interarrival times with $\mu = 5$ and $\sigma = 1$.

Finally, the reader should realize that it is simple, by means of computers, to generate exponentially distributed interarrival times. Thus, it is easy to use such exponentially distributed random variables to simulate queueing systems.

Exercises

Exercise 2.1.1. (Using conditional probability) We have to give one present to one of three children. As we cannot divide the present into parts, we decide to let ‘fate decide’. That is, we choose a random number in the set $\{1, 2, 3\}$. The first child that guesses the number wins the present. Show that the probability of winning the present is the same for each child.

Exercise 2.1.2. Assume that the time X to fail of a machine is uniformly distributed on the interval $[0, 10]$. If the machine fails at time t , the cost to repair it is $h(t)$. What is the expected repair cost?

Exercise 2.1.3. Show that an exponentially distributed random variable is memoryless.

Exercise 2.1.4. If the random variable $X \sim \exp(\lambda)$, show that

$$\mathbb{E}\{X\} = \int_0^\infty t \, dF(t) = \int_0^\infty t f(t) \, dt = \int_0^\infty t \lambda e^{-\lambda t} \, dt = \frac{1}{\lambda},$$

where f is the density function of the distribution function F of X .

Exercise 2.1.5. If the random variable $X \sim \exp(\lambda)$, show that

$$\mathbb{E}\{X^2\} = \int_0^\infty t^2 \lambda e^{-\lambda t} \, dt = \frac{2}{\lambda^2}.$$

Exercise 2.1.6. If the random variable $X \sim \exp(\lambda)$, show that the *variance*

$$\mathbb{V}\{X\} = \mathbb{E}\{X\}^2 - (\mathbb{E}\{X\})^2 = \frac{1}{\lambda^2}.$$

Recall in particular this middle term to compute $\mathbb{V}\{X\}$; it is very practical.

Exercise 2.1.7. Define the *square coefficient of variation (SCV)* as

$$C_a^2 = \frac{V\{X\}}{(E\{X\})^2}. \quad (2.1)$$

Prove that when X is exponentially distributed, $C_a^2 = 1$. As will become clear later, the SCV is a very important concept in queueing theory. Memorize it as a measure of *relative variability*.

Exercise 2.1.8. If X is an exponentially distributed random variable with parameter λ , show that its moment generating function

$$M_X(t) = E\{e^{tX}\} = \frac{\lambda}{\lambda - t}.$$

Exercise 2.1.9. Let A_i be the arrival time of customer i and set $A_0 = 0$. Assume that the interarrival times $\{X_i\}$ are i.i.d. with exponential distribution with mean $1/\lambda$ for some $\lambda > 0$. Prove that the density of $A_i = X_1 + X_2 + \cdots + X_i = \sum_{k=1}^i X_k$ with $i \geq 1$ is

$$f_{A_i}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{i-1}}{(i-1)!}.$$

Exercise 2.1.10. Assume that the interarrival times $\{X_i\}$ are i.i.d. and $X_i \sim \exp(\lambda)$. Let $A_i = X_1 + X_2 + \cdots + X_i = \sum_{k=1}^i X_k$ with $i \geq 1$. Use the density of A_i or the moment generating function of A_i to show that

$$E\{A_i\} = \frac{i}{\lambda},$$

that is, the expected time to see i jobs is i/λ .

Exercise 2.1.11. If $X \sim \exp(\lambda)$ and $S \sim \exp(\mu)$ and X and S are independent, show that

$$Z = \min\{X, S\} \sim \exp(\lambda + \mu),$$

hence $E\{Z\} = (\lambda + \mu)^{-1}$.

Exercise 2.1.12. If $X \sim \exp(\lambda)$, $S \sim \exp(\mu)$ and independent, show that

$$P\{X \leq S\} = \frac{\lambda}{\lambda + \mu}.$$

Exercise 2.1.13. A machine serves two types of jobs. The processing time of jobs of type i , $i = 1, 2$, is exponentially distributed with parameter μ_i . The type T of job is random and independent of anything else, and such that $P\{T = 1\} = p = 1 - q = 1 - P\{T = 2\}$. (An example is a desk serving men and women, both requiring different average service times, and p is the probability that the customer in service is a man.) What is the expected processing time and what is the variance?

Exercise 2.1.14. Try to make Figure 2.2 with simulation.

Hints

Hint 2.1.1: For the second child, condition on the event that the first does not chose the right number.

Hint 2.1.3: Condition on the event $X > t$.

Hint 2.1.9: Check the result for $i = 1$ by filling in $i = 1$ (just to be sure that you have read the formula right), and compare the result to the exponential density. Then write $A_i = \sum_{k=1}^i X_k$, and compute the moment generating function for A_i and use that the interarrival times X_i are independent. Use the moment generating function of X_i .

Hint 2.1.10: Use that $\int_0^\infty (\lambda t)^i e^{-\lambda t} dt = \frac{i!}{\lambda}$. Another way would be to use that, once you have the moment generating function of some random variable X , $E\{X\} = \frac{d}{dt}M(t)|_{t=0}$.

Hint 2.1.11: Use that if $Z = \min\{X, S\} > x$ that then it must be that $X > x$ and $S > x$. Then use independence of X and S .

Hint 2.1.12: Define the joint distribution of X and S and carry out the computations, or use conditioning, or use the result of the previous exercise.

Solutions

Solution 2.1.1: Use the definition of conditional probability ($P\{A|B\} = P\{AB\}/P\{B\}$, provided $P\{B\} > 0$)

The probability that the first child to guess also wins is $1/3$. What is the probability for child number two? Well, for him/her to win, it is necessary that child one does not win and that child two guesses the right number of the remaining numbers. Assume, without loss of generality that child 1 chooses 3 and that this is not the right number. Then

$$\begin{aligned} &P\{\text{Child 2 wins}\} \\ &= P\{\text{Child 2 guesses the right number and child 1 does not win}\} \\ &= P\{\text{Child 2 guesses the right number} | \text{child 1 does not win}\} \cdot P\{\text{Child 1 does not win}\} \\ &= P\{\text{Child 2 makes the right guess in the set } \{1, 2\}\} \cdot \frac{2}{3} \\ &= \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}. \end{aligned}$$

Similar conditional reasoning gives that child 3 wins with probability $1/3$.

Solution 2.1.2: Write for $F(x) = P\{X \leq x\}$ and $f(x) = dF(x)/dx$ for the density of F .

$$\begin{aligned} E\{h(X)\} &= \int_0^{10} E\{h(X) | X = x\} P\{X \in dx\} \\ &= \int_0^{10} E\{h(x) | X = x\} dF(x) \\ &= \int_0^{10} E\{h(x) | X = x\} F(dx) \\ &= \int_0^{10} E\{h(x) | X = x\} f(x) dx \\ &= \int_0^{10} h(x) \frac{dx}{10}. \end{aligned}$$

2.1 Exponential Distribution

Here we introduce some notation that is commonly used in the probability literature to indicate the same conceptual idea, i.e, $P\{X \in dx\} = dF(x) = F(dx) = f(x)dx$, where the last equality follows from the fact that F has a density f everywhere on $[0, 10]$.

The concept of conditional expectation is of fundamental importance in probability theory. Any *good* probability book defines this concept as a random variable measurable with respect to some σ -algebra. In this course we will not deal with this elegant idea, due to lack of time.

Solution 2.1.3: This is easy, but be sure you can do it.

To see that an exponentially distributed is memoryless, use the definition of conditional probability ($P\{A|B\} = P\{AB\}/P\{B\}$, provided $P\{B\} > 0$):

$$P\{X > t+h|X > t\} = \frac{P\{X > t+h, X > t\}}{P\{X > t\}} = \frac{P\{X > t+h\}}{P\{X > t\}} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = P\{X > h\}.$$

Solution 2.1.4:

$$\begin{aligned} E\{X\} &= \int_0^\infty E\{X|X=t\}f(t)dt \\ &= \int_0^\infty t\lambda e^{-\lambda t} dt, \quad \text{density is } \lambda e^{-\lambda t} \\ &= \lambda^{-1} \int_0^\infty u e^{-u} du, \quad \text{by change of variable } u = \lambda t, \\ &= -\lambda^{-1} t e^{-t} \Big|_0^\infty + \lambda^{-1} \int_0^\infty e^{-t} dt \\ &= -\lambda^{-1} e^{-t} \Big|_0^\infty = \frac{1}{\lambda}. \end{aligned}$$

Solution 2.1.5:

$$\begin{aligned} E\{X^2\} &= \int_0^\infty E\{X^2|X=t\}f(t)dt \\ &= \int_0^\infty t^2 \lambda e^{-\lambda t} dt \\ &= \lambda^{-2} \int_0^\infty u^2 e^{-u} du, \quad \text{by change of variable } u = \lambda t, \\ &= -\lambda^{-2} t^2 e^{-t} \Big|_0^\infty + 2\lambda^{-2} \int_0^\infty t e^{-t} dt \\ &= -2\lambda^{-2} t e^{-t} \Big|_0^\infty + 2\lambda^{-2} \int_0^\infty e^{-t} dt \\ &= -2\lambda^{-2} e^{-t} \Big|_0^\infty \\ &= 2/\lambda^2. \end{aligned}$$

Solution 2.1.6: By the previous problems, $E\{X^2\} = 2/\lambda^2$ and $E\{X\} = 1/\lambda$.

Solution 2.1.7: By the previous problems, $V\{X\} = 1/\lambda^2$ and $E\{X\} = 1/\lambda$.

Solution 2.1.8:

$$\begin{aligned} M_X(t) &= E\{\exp(tX)\} = \int_0^\infty e^{tx} dF(x) = \int_0^\infty e^{tx} f(x) dx = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty e^{(t-\lambda)x} dx = \frac{\lambda}{\lambda-t}. \end{aligned}$$

2 Single-Station Queueing Systems

Solution 2.1.9: One way to find the distribution of A_i is by using the moment generating function $M_{A_i}(t) = E\{e^{tA_i}\}$ of A_i . Let X_i be the interarrival time between customers i and $i-1$, and $M_X(t)$ the associated moment generating function. Using the i.i.d. property of the $\{X_i\}$,

$$\begin{aligned} M_{A_i}(t) &= E\{e^{tA_i}\} = E\left\{\exp\left(t \sum_{j=1}^i X_j\right)\right\} \\ &= \prod_{j=1}^i E\{e^{tX_j}\} = \prod_{j=1}^i M_{X_j}(t) = \prod_{j=1}^i \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^i. \end{aligned}$$

From a table of moment generation functions it follows immediately that $A_i \sim \Gamma(n, \lambda)$, i.e., A_i is Gamma distributed.

Solution 2.1.10:

$$E\{A_i\} = \int_0^\infty t P\{A_i \in dt\} = \int_0^\infty t f_{A_i}(t) dt = \int_0^\infty t \lambda e^{-\lambda t} \frac{(\lambda t)^{i-1}}{(i-1)!} dt.$$

Thus,

$$E\{A_i\} = \frac{1}{(i-1)!} \int_0^\infty e^{-\lambda t} (\lambda t)^i dt = \frac{i!}{(i-1)! \lambda} = \frac{i}{\lambda},$$

where we used the hint.

What if we would use the moment generating function?

$$\begin{aligned} E\{A_i\} &= \left. \frac{d}{dt} M_{A_i}(t) \right|_{t=0} \\ &= \left. \frac{d}{dt} \left(\frac{\lambda}{\lambda - t}\right)^i \right|_{t=0} \\ &= i \left(\frac{\lambda}{\lambda - t}\right)^{i-1} \frac{\lambda}{(\lambda - t)^2} \Big|_{t=0} \\ &= \frac{i}{\lambda} \left(\frac{\lambda}{\lambda - t}\right)^{i-1} \Big|_{t=0} \\ &= \frac{i}{\lambda}. \end{aligned}$$

And indeed, by the fact that $E\{X + Y\} = E\{X\} + E\{Y\}$ for any r.v. X and Y ,

$$E\{A_i\} = E\left\{\sum_{k=1}^i X_k\right\} = i E\{X\} = \frac{i}{\lambda}.$$

We get the same answer.

Solution 2.1.11: Use that X and S are independent to get

$$\begin{aligned} P\{Z > x\} &= P\{\min X, S > x\} = P\{X > x \text{ and } S > x\} = P\{X > x\} P\{S > x\} \\ &= e^{-\lambda x} e^{-\mu x} = e^{-(\lambda + \mu)x}. \end{aligned}$$

Solution 2.1.12: There is more than one way to show that $P\{X \leq S\} = \lambda/(\lambda + \mu)$.

Method 1. (I admit that, although the simplest, least technical, method, I did not think of this right away. I am ‘conditioned’ to use conditioning...) Observe first that X and S , being exponentially distributed, both have a density. Moreover, as they are independent, we can sensibly speak of the joint density $f_{X,S}(x,y) = f_X(x)f_S(y) = \lambda\mu e^{-\lambda x}e^{-\mu y}$. With this,

$$\begin{aligned}
 P\{X \leq S\} &= E\{\mathbb{1}_{X \leq S}\} \\
 &= \int_0^\infty \int_0^\infty \mathbb{1}_{x \leq y} f_{X,S}(x,y) \, dy \, dx \\
 &= \lambda\mu \int_0^\infty \int_0^\infty \mathbb{1}_{x \leq y} e^{-\lambda x} e^{-\mu y} \, dy \, dx \\
 &= \lambda\mu \int_0^\infty e^{-\mu y} \int_0^y e^{-\lambda x} \, dx \, dy \\
 &= \mu \int_0^\infty e^{-\mu y} (1 - e^{-\lambda y}) \, dy \\
 &= \mu \int_0^\infty (e^{-\mu y} - e^{-(\lambda+\mu)y}) \, dy \\
 &= \mu \int_0^\infty (e^{-\mu y} - e^{-(\lambda+\mu)y}) \, dy \\
 &= 1 - \frac{\mu}{\lambda + \mu}
 \end{aligned}$$

This argument is provided in the probability book you use in the first year.

Method 2. Applying a standard conditioning argument

$$P\{X \leq S\} = \int_0^\infty P\{X \leq S | S = s\} \mu e^{-\mu s} \, ds.$$

Now, $P\{X \leq S | S = s\}$ is a conditional probability distribution. This is a bit of tricky object, but very useful once you get used to it. The tricky part is that $P\{S = s\} = 0$. Therefore $P\{X \leq S | S = s\}$ cannot be defined as $\frac{P\{X \leq s, S = s\}}{P\{S = s\}}$. However, if we proceed nonetheless and use the independence of S and X , we get

$$P\{X \leq S | S = s\} = \frac{P\{X \leq s, S = s\}}{P\{S = s\}} = \frac{P\{X \leq s\} P\{S = s\}}{P\{S = s\}} = P\{X \leq s\}$$

and thus, indeed, $P\{X \leq S | S = s\} = P\{X \leq s\}$. Then,

$$\begin{aligned}
 P\{X \leq S\} &= \int_0^\infty P\{X \leq S | S = s\} \mu e^{-\mu s} \, ds \\
 &= \int_0^\infty P\{X \leq s\} \mu e^{-\mu s} \, ds \\
 &= \int_0^\infty (1 - e^{-\lambda s}) \mu e^{-\mu s} \, ds
 \end{aligned}$$

and we arrive at the integral we have seen above.

So we get the correct answer, but by the wrong method. How can we repair this? As a first step, let's not fix S to a set of measure zero, but let's assume that $S \in [s, t]$ for $s < t$. Then it follows that

$$\mathbb{1}_{X \leq s} \mathbb{1}_{S \in [s, t]} \leq \mathbb{1}_{X \leq S} \mathbb{1}_{S \in [s, t]} \leq \mathbb{1}_{X \leq t} \mathbb{1}_{S \in [s, t]}$$

2 Single-Station Queueing Systems

As a second step, using that $P\{S \in [s, t]\} > 0$ if $s < t$ and the independence of X and S ,

$$\begin{aligned} P\{X \leq s\} &= \frac{P\{X \leq s\} P\{S \in [s, t]\}}{P\{S \in [s, t]\}} = \frac{P\{X \leq s, S \in [s, t]\}}{P\{S \in [s, t]\}} \\ P\{X \leq t\} &= \frac{P\{X \leq t\} P\{S \in [s, t]\}}{P\{S \in [s, t]\}} = \frac{P\{X \leq t, S \in [s, t]\}}{P\{S \in [s, t]\}} \end{aligned}$$

Now with the result of the first step

$$\begin{aligned} P\{X \leq s\} &= \frac{P\{X \leq s, S \in [s, t]\}}{P\{S \in [s, t]\}} \\ &\leq \frac{P\{X \leq S, S \in [s, t]\}}{P\{S \in [s, t]\}} \\ &= P\{X \leq S | S \in [s, t]\} \\ &\leq \frac{P\{X \leq t, S \in [s, t]\}}{P\{S \in [s, t]\}} \\ &= P\{X \leq t\}. \end{aligned}$$

Hence,

$$P\{X \leq s\} \leq P\{X \leq S | S \in [s, t]\} \leq P\{X \leq t\}.$$

Finally, taking the limit $t \downarrow s$, and defining $P\{X \leq S | S = s\} = \lim_{t \downarrow s} P\{X \leq S | S \in [s, t]\}$, it follows that

$$P\{X \leq s\} = P\{X \leq S | S = s\}$$

A more direct way to properly define $P\{X \leq S | S = s\}$ is as follows. For any y such that $f_S(y) > 0$, we can define the conditional probability density function of X , given that $S = s$, as

$$f_{X|S}(x|s) = \frac{f_{X,S}(x, s)}{f_S(s)},$$

where, as before, $f_{X,S}(x, s)$ is the joint density of X and S . Now that the conditional probability density is defined, we can properly define

$$E\{X | S = s\} = \int_0^\infty x f_{X|S}(x|s) dx$$

and also

$$P\{X \leq S | S = s\} = E\{\mathbb{1}_{X \leq S} | S = s\} = \int_0^\infty \mathbb{1}_{x \leq s} f_{X|S}(x|s) dx.$$

Using the definition of $f_{X|S}(x|s)$ and the independence of X and S it follows that

$$f_{X|S}(x|s) = \frac{f_{X,S}(x, s)}{f_S(s)} = \frac{\lambda e^{-\lambda x} \mu e^{-\mu s}}{\mu e^{-\mu s}} = \lambda e^{-\lambda x}$$

from which we get that

$$\begin{aligned} E\{\mathbb{1}_{X \leq S} | S = s\} &= \int_0^\infty \mathbb{1}_{x \leq s} f_{X|S}(x|s) dx \\ &= \int_0^\infty \mathbb{1}_{x \leq s} \lambda e^{-\lambda x} dx \\ &= \int_0^s \lambda e^{-\lambda x} dx \\ &= 1 - e^{-\lambda s}, \end{aligned}$$

that is,

$$P\{X \leq S | S = s\} = E\{\mathbb{1}_{X \leq S} | S = s\} = 1 - e^{-\lambda s} = P\{X \leq s\}.$$

All of these problems can be put on solid ground by using measure theory. We do not pursue these matters any further, but trust on our intuition that all is well.

Solution 2.1.13: Let X be the processing (or service) time at the server, and X_i the service time of a type i job. Then,

$$X = \mathbb{1}_{T=1}X_1 + \mathbb{1}_{T=2}X_2,$$

where $\mathbb{1}$ is the indicator function, that is, $\mathbb{1}_A = 1$ if the event A is true, and $\mathbb{1}_A = 0$ if A is not true. With this,

$$\begin{aligned} E\{X\} &= E\{\mathbb{1}_{T=1}X_1\} + E\{\mathbb{1}_{T=2}X_2\} \\ &= E\{\mathbb{1}_{T=1}\}E\{X_1\} + E\{\mathbb{1}_{T=2}\}E\{X_2\}, \text{ by the independence of } T, \\ &= P\{T=1\}/\mu_1 + P\{T=2\}/\mu_2 \\ &= p/\mu_1 + q/\mu_2 \\ &= pE\{X_1\} + qE\{X_2\}. \end{aligned}$$

(The next derivation may seem a bit long, but the algebra is standard. I include all steps so that you don't have to use pen and paper yourself if you want to check the result.) Next, using that

$$\mathbb{1}_{T=1}\mathbb{1}_{T=2} = 0 \text{ and } \mathbb{1}_{T=1}^2 = \mathbb{1}_{T=1},$$

we get

$$\begin{aligned} V\{X\} &= E\{X^2\} - (E\{X\})^2 \\ &= E\{(\mathbb{1}_{T=1}X_1 + \mathbb{1}_{T=2}X_2)^2\} - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\ &= E\{\mathbb{1}_{T=1}X_1^2 + \mathbb{1}_{T=2}X_2^2\} - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\ &= pE\{X_1^2\} + qE\{X_2^2\} - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\ &= pV\{X_1\} + p(E\{X_1\})^2 + qV\{X_2\} + q(E\{X_2\})^2 - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\ &= pV\{X_1\} + \frac{p}{\mu_1^2} + qV\{X_2\} + \frac{q}{\mu_2^2} - \left(\frac{p}{\mu_1} + \frac{q}{\mu_2}\right)^2 \\ &= pV\{X_1\} + qV\{X_2\} + \frac{p}{\mu_1^2} + \frac{q}{\mu_2^2} - \frac{p^2}{\mu_1^2} - \frac{q^2}{\mu_2^2} - \frac{2pq}{\mu_1\mu_2} \\ &= pV\{X_1\} + qV\{X_2\} + \frac{p(1-p)}{\mu_1^2} + \frac{q(1-q)}{\mu_2^2} - \frac{2pq}{\mu_1\mu_2} \\ &= pV\{X_1\} + qV\{X_2\} + \frac{pq}{\mu_1^2} + \frac{qp}{\mu_2^2} - \frac{2pq}{\mu_1\mu_2} \\ &= pV\{X_1\} + qV\{X_2\} + pq(E\{X_1\} - E\{X_2\})^2. \end{aligned}$$

2 Single-Station Queueing Systems

Interestingly, we see that even if $V\{X_1\} = V\{X_2\} = 0$, $V\{X\} > 0$ if $E\{X_1\} \neq E\{X_2\}$. Bear this in mind; we will use these ideas later when we discuss the effects of failures on the variance of service times of jobs.

Solution 2.1.14: The source code can be found in `progs/converge_to_exp.py`.

2.2 Poisson Distribution

In this section we provide a derivation of, and motivation for, the Poisson process, and clarify its relation with the exponential distribution at the end.

Consider a machine that fails occasionally. Let us write $N(s, t)$ for the number of failures occurring during a time interval of $(s, t]$. We assume, without loss of generality, that repairs are instantaneous. Clearly, as we do not know in advance how often the machine will fail, we model $N(s, t)$ as a random variable for all times s and t .

Our first assumption is that the failure behavior of the machine does not significantly change over time. Then it is reasonable to assume that the expected number of failure is proportional to the length of the interval T . Thus, it is reasonable to assume that there exists some constant λ such that

$$E\{N(s, t)\} = \lambda(t - s) \quad (2.2)$$

The constant λ is often called the *arrival rate*, or failure rate in this case.

The second assumption is that $\{N(s, t), s \leq t\}$ has *stationary* and *independent increments*. Stationarity means that the distribution of the number of arrivals are the same for all intervals of equal length. Formally, $N(s_1, t_1)$ has the same distribution as $N(s_2, t_2)$ if $t_2 - s_2 = t_1 - s_1$. Independence means, roughly speaking, that knowing that $N(s_1, t_1) = n$, does not help to make any predictions about $N(s_2, t_2)$ if the intervals (s_1, t_1) and (s_2, t_2) have no overlap.

To find the distribution of $N(0, t)$, let us split the interval $[0, t]$ into n sub-intervals, all of equal length, and ask: ‘What is the probability that the machine will fail in some given sub-interval.’ By our second assumption, the failure behavior is constant over time. Therefore, the probability p to fail in each interval should be equal. Moreover, if n is large, p must be small, for otherwise (2.2) could not be true. As a consequence, if the time intervals are very small, we can safely neglect the probability that two or more failures occur in one such tiny interval.

As a consequence, then, we can model the occurrence of a failure in some period i as a Bernoulli distributed random variable B_i such that $P\{B_i = 1\}$ and $P\{B_i = 0\} = 1 - P\{B_i = 1\}$, and we assume that $\{B_i\}$ are independent. The total number of failures $N_n(t)$ that occur in n intervals is then binomially distributed

$$P\{N_n(t) = k\} = \binom{n}{k} p^k (1 - p)^{n-k}. \quad (2.3)$$

In the exercises we ask you to use this to motivate that, in some appropriate sense, $N_n(t)$ converges to $N(t)$ for $n \rightarrow \infty$ such that

$$P\{N(t) = k\} = e^{-\lambda t} \frac{(\lambda t)^k}{k!}. \quad (2.4)$$

We say that $N(t)$ is *Poisson distributed* with rate λ , and write $N(t) \sim P(\lambda t)$.

Moreover, if there are no failures in some interval $[0, t]$, then it must be that $N(t) = 0$ and the interarrival time $X_1 = A_1 - 0$ (as $A_0 = 0$) must be larger than t . Therefore,

$$P\{X_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t} \frac{(\lambda t)^0}{0!} = e^{-\lambda t}.$$

The relations we discussed above are of paramount importance in the analysis of queueing process. We summarize this by a theorem.

Theorem 2.2.1. A counting process $\{N(t)\}$ is a Poisson process with rate λ if and only if the inter-arrival times $\{X_i\}$, i.e., the times between consecutive arrivals, are i.i.d. and $P\{X_1 \leq t\} = 1 - e^{-\lambda t}$. In other words, $X_i \sim \exp(\lambda) \Leftrightarrow N(t) \sim P(\lambda t)$

Exercises

Exercise 2.2.1. Show that $E\{N_n(t)\} = \sum_{i=1}^n E\{B_i\} = np$.

Exercise 2.2.2. What is the difference between $N_n(t)$ and $N(t)$?


Exercise 2.2.3. Show how the binomial distribution (2.3) converges to the Poisson distribution (2.4) if $n \rightarrow \infty$, $p \rightarrow 0$ such that $np = \lambda$.


Exercise 2.2.4. If the inter-arrival times $\{X_i\}$ are i.i.d. and exponentially distributed with mean $1/\lambda$, prove that the number $N(t)$ of arrivals during interval $[0, t]$ is Poisson distributed.

Exercise 2.2.5. Show that if $N(t) \sim P(\lambda t)$, we have for small h ,

1. $P\{N(h) = n \mid N(0) = n\} = 1 - \lambda h + o(h)$
2. $P\{N(h) = n + 1 \mid N(0) = n\} = \lambda h + o(h)$
3. $P\{N(h) \geq n + 2 \mid N(0) = n\} = o(h)$,

where $o(h)$ is a function $f(h)$ such $f(h)/h \rightarrow 0$ as $h \rightarrow 0$.

Exercise 2.2.6. Assume a timer fires at times $0 = T_0 < T_1 < T_2 < \dots$, such that $T_k - T_{k-1} \sim \exp(\lambda)$. Define $N(t) = \sum_{k=0}^{\infty} k \mathbb{1}_{T_k \leq t < T_{k+1}}$, where $\mathbb{1}$ is the *indicator function*, that is, $\mathbb{1}_A = 1$ if the event A is true, and $\mathbb{1}_A = 0$ if A is not true. What is the distribution of $N(t)$? 

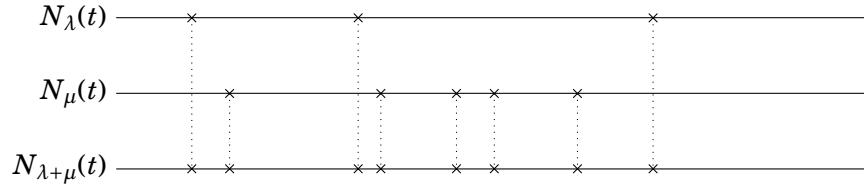
Exercise 2.2.7. (I added this question to clarify the next question, you can skip it for year 16/17.) Show that for a Poisson process N , 

$$P\{N(0, s] = 1 \mid N(0, t] = 1\} = \frac{s}{t}.$$

Thus, if you know that $N(0, t] = 1$, the arrival is distributed uniformly on the interval $[0, t]$.

Exercise 2.2.8. (Merged Poisson streams form a new Poisson process with the sum of the rates.) Assume that $N_a(t) \sim P(\lambda t)$, $N_s(t) \sim P(\mu t)$ and independent. Show that $N_a(t) + N_s(t) \sim P((\lambda + \mu)t)$. The figure below provides a graphical representation of merging (also called superposition) of streams.

2 Single-Station Queueing Systems



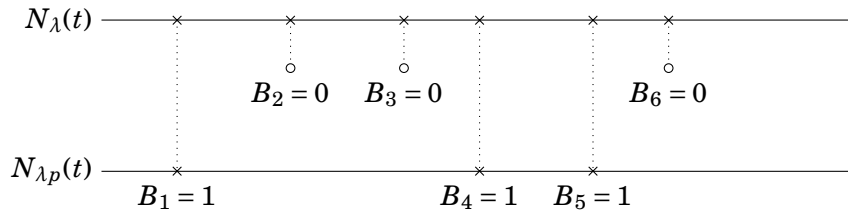
Exercise 2.2.9. Assume that $N_a(t) \sim P(\lambda t)$, $N_s(t) \sim P(\mu t)$ and independent. Use that $N_a(t) + N_s(t) \sim P((\lambda + \mu)t)$ to conclude

$$P\{N_a(h) = 1, N_s(h) = 0 | N_a(h) + N_s(h) = 1\} = \frac{\lambda}{\lambda + \mu}.$$

Note that the right-hand-side does not depend on h , hence it holds for any time h , whether it is small or not.

Exercise 2.2.10. Assume that $N_a(t) \sim P(\lambda t)$, $N_s(t) \sim P(\mu t)$ and independent. What is actually the meaning of the event $\{N_a(h) = 1, N_s(h) = 0\} \cap \{N_a(h) + N_s(h) = 1\}$?

Exercise 2.2.11. (A Bernoulli-thinned Poisson process is still a Poisson process) Consider a Poisson process. Split the process in the following way. When a job arrives, throw a coin that lands heads with probability p and tails with $q = 1 - p$. When the coin lands heads, call the job of type 1, otherwise of type 2. Another way of thinning is by modeling the stream of people passing a shop as a Poisson process with rate λ . With probability p a person decides, independent of anything else, to enter the shop, c.f. the figure below. The crosses at the upper line are passersby in the street. The crosses at the lower line are the customers that enter the shop. The outcome of the k th throw of a coin is indicated by B_k . Since $B_1 = 1$ the first passerby turns into a customer entering the shop; the second passerby does not enter as $B_2 = 0$, and so on. Like this, the stream of passerby's is thinned with Bernoulli distributed random variables.



The concept of thinning is particularly useful to analyze queueing networks. Suppose the departure stream of a machine is split into two substreams, e.g., a fraction p of the jobs move on another machine and the rest $(1 - p)$ of the jobs leave the system. Then we can model the arrival stream at the second machine as a thinned stream (with probability p) of the departures of the first machine.

Show that the Poisson process obtained by thinning the original process is $\sim P(\lambda t p)$ for any t .

Exercise 2.2.12. Show that $E\{N(t)\} = \lambda t$ and $V\{N(t)\} = \lambda t$. Why is the SCV of $N(t)$ equal to $1/\lambda t$? Conclude that the relative variability of $N(t)$ becomes smaller as t becomes larger. Observe that the SCV of a Poisson distributed random variable is not the same as the SCV of an exponentially distributed random variable.

Hints

Hint 2.2.1: Recall that $E\{X + Y\} = E\{X\} + E\{Y\}$. Just as a reminder, $E\{XY\} \neq E\{X\}E\{Y\}$ in general. Only when X and Y are uncorrelated (which is implied by independence), the product of the expectations is the expectation of the products.

Hint 2.2.3: First find p , n , λ and t are such that the rate at which event occur in both processes are the same. Then consider the binomial distribution and use the standard limit $(1 - x/n)^n \rightarrow e^{-x}$ as $n \rightarrow \infty$.

Hint 2.2.4: Realize that $P\{N(t) = k\} = P\{A_k \leq t\} - P\{A_{k+1} \leq t\}$.

Hint 2.2.5: Think about the meaning of the formula $P\{N(h) = n | N(0) = n\}$. It is a conditional probability that should be read like this: given that up to time 0 we have seen n arrivals (i.e., $N(0) = n$), what is the probability that just a little later (h) the number of arrivals is still n , i.e., $N(h) = n$? Then use the definition of the Poisson distribution to compute this probability. Finally, use Taylor's expansion of e^x to see that $e^x = 1 + x + o(x)$ for $|x| \ll 1$. Furthermore, use that $\sum_{i=2}^{\infty} x^i/i! = \sum_{i=0}^{\infty} x^i/i! - x - 1 = e^x - x - 1$.

Hint 2.2.7: Use Bayes' law for conditional probability. Observe that

$$\{N(0, s] + N(s, t] = 1\} \cap \{N(0, s] = 1\} = \{1 + N(s, t] = 1\} \cap \{N(0, s] = 1\} = \{N(s, t] = 0\} \cap \{N(0, s] = 1\}.$$

Hint 2.2.8: Use a conditioning argument or use probability generating functions (i.e. $E\{z^X\} = \sum_k z^k P\{X = k\}$). In particular, for conditioning, use that $P\{AB\} = P\{A|B\}P\{B\}$. More generally, if the set A can be split into disjoint sets B_i , i.e., $A = \bigcup_{i=1}^n B_i$, then

$$P\{A\} = \sum_{i=1}^n P\{AB_i\} = \sum_{i=1}^n P\{A|B_i\}P\{B_i\},$$

where we use the conditioning formula to see that $P\{AB_i\} = P\{A|B_i\}P\{B_i\}$. Now choose practical sets B_i .

Hint 2.2.9: Suppose we write $N(t) = N_a(t) + N_s(t)$. Then

$$P\{N_a(h) = 1, N_s(h) = 0 | N(h) = 1\}$$

is the probability that $N_a(h) = 1$ and $N_s(h) = 0$ given that $N(h) = 1$. In other words, the question is find out that, given one of the two processes 'fired', what is the probability that N_a was the one that 'fired'.

Hint 2.2.11: Condition on the total number of arrivals $N(t) = m$ up to time t . Realize that the probability that a job is of type 1 is Bernoulli distributed, hence when you consider m jobs in total, the number of type 1 jobs is binomially distributed.

Again use that if the set A can be split into disjoint sets B_i , i.e., $A = \bigcup_{i=1}^n B_i$, then

$$P\{A\} = \sum_{i=1}^n P\{A|B_i\}P\{B_i\}.$$

Now choose practical sets B_i .

2 Single-Station Queueing Systems

You might also consider the random variable

$$Y = \sum_{i=1}^N Z_i,$$

with $N \sim P(\lambda)$ and $Z_i \sim B(p)$. Show that the moment generating function of Y is equal to the moment generating function of a Poisson random variable with parameter λp .

Hint 2.2.12: You might use generating functions here.

Solutions

Solution 2.2.1:

$$\mathbb{E}\{N_n(t)\} = \mathbb{E}\left\{\sum_{i=1}^n B_i\right\} = \sum_{i=1}^n \mathbb{E}\{B_i\} = n \mathbb{E}\{B_i\} = np.$$

Solution 2.2.2: $N_n(t)$ is a binomially distributed random variable with parameters n and p . The maximum value of $N_n(t)$ is n . The random variable $N(t)$ models the number of failures that can occur during $[0, t]$. As such it is not necessarily bounded by n . Thus, $N_n(t)$ and $N(t)$ cannot represent the same random variable.

Solution 2.2.3: Now we like to relate $N_n(t)$ and $N(t)$. It is clear that we at least want the expectations to be the same, that is, $np = \lambda t$. This implies that

$$p = \frac{\lambda t}{n},$$

so that

$$\mathbb{P}\{N_n(t) = k\} = \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k}.$$

To see that

$$\lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} = e^{-\lambda t} \frac{(\lambda t)^k}{k!}.$$

use that

$$\begin{aligned} \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^n \\ &= \frac{(\lambda t)^k}{k!} \left(\frac{n}{n-\lambda t}\right)^k \frac{n!}{n^k(n-k)!} \left(1 - \frac{\lambda t}{n}\right)^n \\ &= \frac{(\lambda t)^k}{k!} \left(\frac{n}{n-\lambda t}\right)^k \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \left(1 - \frac{\lambda t}{n}\right)^n. \end{aligned}$$

Observe now that, as λt is finite, $n/(n-\lambda t) \rightarrow 1$ as $n \rightarrow \infty$. Also for any finite k , $(n-k)/n \rightarrow 1$. Finally, we use that $(1+x/n)^n \rightarrow e^x$ so that $\left(1 - \frac{\lambda t}{n}\right)^n \rightarrow e^{-\lambda t}$. The rest is easy, so that, as $n \rightarrow \infty$, the above converges to

$$\frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

Solution 2.2.4: We want to show that

$$P\{N(t) = k\} = e^{-\lambda t} \frac{(\lambda t)^k}{k!}.$$

Now observe that $P\{N(t) = k\} = P\{A_k \leq t\} - P\{A_{k+1} \leq t\}$. Using the density of A_{k+1} as obtained previously and applying partial integration leads to

$$\begin{aligned} P\{A_{k+1} \leq t\} &= \lambda \int_0^t \frac{(\lambda s)^k}{k!} e^{-\lambda s} ds \\ &= \lambda \frac{(\lambda s)^k}{k!} \frac{e^{-\lambda s}}{-\lambda} \Big|_0^t + \lambda \int_0^t \frac{(\lambda s)^{k-1}}{(k-1)!} e^{-\lambda s} ds \\ &= -\frac{(\lambda t)^k}{k!} e^{-\lambda t} + P\{A_k \leq t\} \end{aligned}$$

We are done.

Solution 2.2.5: 1. $P\{N(h) = n | N(0) = n\} = P\{N(h) = 0\} = e^{-\lambda h} (\lambda h)^0 / 0! = e^{-\lambda h} = 1 - \lambda h + o(h)$, as follows from a standard argument in analysis that $e^{-\lambda h} = 1 - \lambda h + o(h)$ for h small.

2. $P\{N(h) = n + 1 | N(0) = n\} = P\{N(h) = 1\} = e^{-\lambda h} (\lambda h)^1 / 1! = (1 - \lambda h + o(h)) \lambda h = \lambda h - \lambda^2 h^2 + o(h) = \lambda h + o(h)$.

3.

$$\begin{aligned} P\{N(h) \geq n + 2 | N(0) = n\} &= P\{N(h) \geq 2\} \\ &= e^{-\lambda h} \sum_{i=2}^{\infty} \frac{(\lambda h)^i}{i!} = e^{-\lambda h} \left(\sum_{i=0}^{\infty} \frac{(\lambda h)^i}{i!} - \lambda h - 1 \right) \\ &= e^{-\lambda h} (e^{\lambda h} - 1 - \lambda h) = 1 - e^{-\lambda h} (1 + \lambda h) \\ &= 1 - (1 - \lambda h + o(h))(1 + \lambda h) = 1 - (1 - \lambda^2 h^2 + o(h)) \\ &= \lambda^2 h^2 + o(h) = o(h). \end{aligned}$$

We can also use the results of the previous parts to see that

$$\begin{aligned} P\{N(h) \geq n + 2 | N(0) = n\} &= P\{N(h) \geq 2\} = 1 - P\{N(h) < 2\} \\ &= 1 - P\{N(h) = 0\} - P\{N(h) = 1\} \\ &= 1 - (1 - \lambda h + o(h)) - (\lambda h + o(h)) \\ &= o(h). \end{aligned}$$

Solution 2.2.6: $N(t) \sim P(\lambda t)$.

Solution 2.2.7:

$$\begin{aligned} P\{N(0, s] = 1 | N(0, t] = 1\} &= \frac{P\{N(0, s] = 1, N(0, t] = 1\}}{P\{N(0, t] = 1\}} \\ &= P\{N(0, t] = 1 | N(0, s] = 1\} \frac{P\{N(0, s] = 1\}}{P\{N(0, t] = 1\}} \\ &= P\{N(0, s] + N(s, t] = 1 | N(0, s] = 1\} \frac{e^{-\lambda s} \lambda s}{e^{-\lambda t} \lambda t} \\ &= P\{1 + N(s, t] = 1 | N(0, s] = 1\} e^{-\lambda(s-t)} \frac{s}{t} \\ &= P\{N(s, t] = 0\} e^{-\lambda(s-t)} \frac{s}{t} = e^{-\lambda(t-s)} e^{-\lambda(s-t)} \frac{s}{t} = \frac{s}{t}. \end{aligned}$$

2 Single-Station Queueing Systems

Solution 2.2.8: First we show how to use conditioning.

$$P\{N_a(t) + N_s(t) = n\} = \sum_{i=0}^n P\{N_a(t) + N_s(t) = n | N_a(t) = i\} P\{N_a(t) = i\}$$

Now, if $N_a(t) = i$, then

$$N_a(t) + N_s(t) = n \iff i + N_s(t) = n \iff N_s(t) = n - i.$$

Thus,

$$\begin{aligned} P\{N_a(t) + N_s(t) = n\} &= \sum_{i=0}^n P\{N_s(t) = n - i\} P\{N_a(t) = i\} \\ &= \sum_{i=0}^n \frac{(\mu t)^{n-i}}{(n-i)!} \frac{(\lambda t)^i}{i!} e^{-(\mu+\lambda)t} \\ &= e^{-(\mu+\lambda)t} \sum_{i=0}^n \frac{(\mu t)^{n-i}}{(n-i)!} \frac{(\lambda t)^i}{i!} \\ &= e^{-(\mu+\lambda)t} \frac{1}{n!} \sum_{i=0}^n \binom{n}{i} (\mu t)^{n-i} (\lambda t)^i \\ &= \frac{((\mu + \lambda)t)^n}{n!} e^{-(\mu+\lambda)t}. \end{aligned} \tag{2.5}$$

Now with the probability generating functions.

$$\begin{aligned} M_a(z) &= E\{z^{N_a(t)}\} = \sum_{k=0}^{\infty} z^k P\{N_a(t) = k\} \\ &= \sum_{k=0}^{\infty} z^k e^{-\lambda t} \frac{(\lambda t)^k}{k!} \\ &= e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(z\lambda t)^k}{k!} \\ &= \exp(\lambda t(z - 1)). \end{aligned}$$

Use this, and the similar expression for $N_s(t)$ and independence to see that

$$M(z) = E\{z^{N(t)}\} = E\{z^{N_a(t)}\} E\{z^{N_s(t)}\} = \exp((\lambda + \mu)t(z - 1)).$$

Finally, since the generating function uniquely characterizes the distribution, and since the above expression has the same form as $M_1(z)$ but with $\lambda + \mu$ replacing λ , we can conclude that $N(t) \sim P((\lambda + \mu)t)$.

Solution 2.2.9: With the above:

$$\begin{aligned}
& P\{N_a(h) = 1, N_s(h) = 0 | N_a(h) + N_s(h) = 1\} \\
&= \frac{P\{N_a(h) = 1, N_s(h) = 0, N_a(h) + N_s(h) = 1\}}{P\{N_a(h) + N_s(h) = 1\}} \\
&= \frac{P\{N_a(h) = 1, N_s(h) = 0\}}{P\{N_a(h) + N_s(h) = 1\}} \\
&= \frac{P\{N_a(h) = 1\}P\{N_s(h) = 0\}}{P\{N_a(h) + N_s(h) = 1\}} \\
&= \frac{\lambda h \exp(-\lambda h) \exp(-\mu h)}{((\lambda + \mu)h) \exp(-(\lambda + \mu)h)} \\
&= \frac{\lambda h \exp(-(\lambda + \mu)h)}{((\lambda + \mu)h) \exp(-(\lambda + \mu)h)} \\
&= \frac{\lambda}{\lambda + \mu}
\end{aligned}$$

Solution 2.2.10: This means that, given that an event occurred, the event was an arrival, i.e., N_a was the first.

Solution 2.2.11: Suppose that N_1 is the thinned stream, and N the total stream. Then

$$\begin{aligned}
P\{N_1 = k\} &= \sum_{n=k}^{\infty} P\{N_1 = k, N = n\} = \sum_{n=k}^{\infty} P\{N_1 = k | N = n\} P\{N = n\} \\
&= \sum_{n=k}^{\infty} P\{N_1 = k | N = n\} e^{-\lambda} \frac{\lambda^n}{n!} = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda} \frac{\lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=k}^{\infty} \frac{p^k (1-p)^{n-k}}{k!(n-k)!} \lambda^n = e^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{(\lambda(1-p))^{n-k}}{(n-k)!} \\
&= e^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{n=0}^{\infty} \frac{(\lambda(1-p))^n}{n!} = e^{-\lambda} \frac{(\lambda p)^k}{k!} e^{\lambda(1-p)} \\
&= e^{-\lambda p} \frac{(\lambda p)^k}{k!}.
\end{aligned}$$

We see that the thinned stream is Poisson with parameter λp . (For notational ease, we left out the t , otherwise it is $P(\lambda t p)$).

Now consider $Y = \sum_{i=1}^N Z_i$. Suppose that $N = n$, so that n arrivals occurred. Then we throw n coins with success probability p . It follows that Y is indeed a thinned Poisson random variable. Model the coins as a generic Bernoulli distributed random variable Z . We first need

$$E\{e^{sZ}\} = e^0 P\{Z = 0\} + e^s P\{Z = 1\} = (1-p) + e^s p.$$

Suppose that $N = n$, then since the Z_i are i.i.d.,

$$E\{e^{s \sum_{i=1}^n Z_i}\} = \left(E\{e^{sZ}\}\right)^n = (1 + p(e^s - 1))^n$$

Then, using conditioning on N ,

$$\begin{aligned}
E\{e^{sY}\} &= E\left\{E\left\{e^{s \sum_{i=1}^N Z_i} \mid N = n\right\}\right\} = E\{E\{(1 + p(e^s - 1))^n \mid N = n\}\} \\
&= \sum_{n=0}^{\infty} (1 + p(e^s - 1))^n e^{-\lambda} \frac{\lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(1 + p(e^s - 1))^n \lambda^n}{n!} \\
&= e^{-\lambda} e^{\lambda(1 + p(e^s - 1))} = e^{\lambda p(e^s - 1)}.
\end{aligned}$$

2 Single-Station Queueing Systems

Thus, Y has the same moment generating function as a Poisson distributed random variable with parameter λp . Since moment-generating functions specify the distribution uniquely, $Y \sim P(\lambda p)$.

Solution 2.2.12: I expect that this is easy for you by now. The exercise is just meant to help you recall this.

When a random variable N is Poisson distributed with parameter λ ,

$$\begin{aligned}
 E\{N\} &= \sum_{n=0}^{\infty} n e^{-\lambda} \frac{\lambda^n}{n!} \\
 &= \sum_{n=1}^{\infty} n e^{-\lambda} \frac{\lambda^n}{n!}, \text{ since the term with } n=0 \text{ cannot contribute} \\
 &= e^{-\lambda} \lambda \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} \\
 &= e^{-\lambda} \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}, \text{ by a change of variable} \\
 &= e^{-\lambda} \lambda e^{\lambda} \\
 &= \lambda.
 \end{aligned}$$

Similarly, using that $V\{N\} = E\{N^2\} - (E\{N\})^2$,

$$\begin{aligned}
 E\{N^2\} &= \sum_{n=0}^{\infty} n^2 e^{-\lambda} \frac{\lambda^n}{n!} \\
 &= e^{-\lambda} \sum_{n=1}^{\infty} n \frac{\lambda^n}{(n-1)!} \\
 &= e^{-\lambda} \sum_{n=0}^{\infty} (n+1) \frac{\lambda^{n+1}}{n!} \\
 &= e^{-\lambda} \lambda \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} + e^{-\lambda} \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \\
 &= \lambda^2 + \lambda.
 \end{aligned}$$

The Poisson *process* $\{N(t)\}$ is a much more complicated object than the Poisson distributed random variable $N(t)$. The process contains it is an *uncountable set* of random variables, whereas $N(t)$ is just *one* random variable. However, for a fixed t the element $N(t)$ is a random variable that is Poisson distributed with parameter λt . Using the above, the answer of the question follows immediately.

With generating functions we get the result without too much effort. Suppose N is a Poisson

2.3 Kendall's Notation to Characterize Queueing Processes

distributed random variable. Then

$$\begin{aligned}
 \phi(z) &= E\{z^N\} \\
 &= \sum_{k=0}^{\infty} z^k P\{N = k\} \\
 &= \sum_{k=0}^{\infty} z^k \frac{\lambda^k}{k!} e^{-\lambda} \\
 &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(z\lambda)^k}{k!} \\
 &= e^{-\lambda} e^{z\lambda} \\
 &= e^{(z-1)\lambda}.
 \end{aligned}$$

Observe that $\phi'(z) = E\{Nz^N\}$, so that $\phi'(1) = E\{N\}$; also $\phi''(z) = E\{N(N-1)z^{N-2}\}$, hence $\phi''(1) = E\{N(N-1)\}$. With this,

$$\begin{aligned}
 E\{N\} &= \phi'(1) = \lambda, \\
 E\{N^2\} &= \phi''(1) + E\{N\} = \lambda^2 + \lambda, \\
 V\{N\} &= E\{N^2\} - (E\{N\})^2 = \lambda, \\
 SCV &= \frac{V\{N(t)\}}{(E\{N(t)\})^2} = \frac{\lambda t}{(\lambda t)^2} = \frac{1}{(\lambda t)^2}.
 \end{aligned}$$

Clearly, as t increases, $1/\lambda t$ decreases.

Here is a point of confusion for some students. The SCV of an exponentially distributed random variable is 1; the SCV of the related Poisson process $N_\lambda(t)$ is *not* identically 1 for all t .

2.3 Kendall's Notation to Characterize Queueing Processes

As will become apparent in Sections 2.4 and ??, the construction of any queueing process involves three main elements: the distribution of the interarrival times between consecutive jobs, the distribution of the service times of the individual jobs, and the number of servers present to process jobs. In this characterization it is implicit that the interarrival times form a set of i.i.d. (independent and identically distributed) random variables, the service times are also i.i.d., and finally, the interarrival times and service times are mutually independent.

To characterize the type of queueing process it is common to use the *abbreviation* $A/B/c/K$ where A is the distribution of the interarrival times, B the distribution of the services, c the number of servers, and K the size of the system. In this notation it is assumed that jobs are served in first-in-first-out (FIFO) order; FIFO scheduling is also often called first-come-first-serve (FCFS).

Let us illustrate the shorthand $A/B/c/K$ with some examples:

- $M/M/1$: the distribution of the interarrival times is *Memory-less*, hence exponential, the service times are also *Memoryless*, and there is 1 server. As K is unspecified, it is assumed to be infinite.
- $M/M/c$: A *multi-server* queue with c servers in which all servers have the same capacity. Jobs arrive according to a Poisson process and have exponentially distributed processing times.