IEOR 4615: February 22, 2012, Professors Whitt and Yom-Tov

Lecture 6. Modeling Arrivals in a Service System

1 Looking At Call Center Arrivals

We will be learning how to look at arrival process data. This means (i) how to create appropriate figures using the call center data and (ii) how to see things in the plots.

1.1 Different Time Scales

It is useful to look at arrival process data in different time scales. Depending on the application, this might range from decades down to a minute or less. Figure **??** shows plots of call center data over four time scales.

Years over a decade. For a service system, the longest time scale might be a decade. We look at yearly totals over the multi-year period to see the long-term trends. Such a plot gives an estimate of the arrival rate in calls per year. Is the arrival rate changing systematically?

Months over a year. The next time scale is a single year. We now look at monthly totals over each year. This should be done for both individual years and taking averages over multiple years. Are their seasonal effects?

Days over a month. The next time scale is a single month. We now look at daily totals over each month. This should be done for both individual months and taking averages over multiple months. Are their daily effects? Given that service systems tend to differently on weekends than on weekdays, it is actually usually better to focus on days within a week.

Special days in a year. It is also appropriate to consider if there are systematic differences in arrivals over holidays and other special days.

Hours over a day. The next time scale is a single day. We now look at hourly totals over each day. This should be done for both individual days and taking averages over multiple days, where we use the previous larger-scale analysis to identify what days should be used in the averages. Here we expect to see systematic **predictable variability** (that will be seen repeatedly on each sample).

Minutes over an hour. The next time scale is a single hour. We now look at totals over $0.5 - 5.0$-minute intervals during each hour. This should be done for both individual hours and taking averages over multiple hours, where we use the previous larger-scale analysis to identify what hours should be used in the averages. It is usually better to use the same hour on multiple days than different hours on the same day. Here we expect to be seeing stochastic fluctuations, for which we use stochastic models.

1.2 Focusing on and Removing Stochastic Variability

It can be useful to show more arrival counts per time period, especially over a day. Figure **??** shows 408 totals in 2.5-minute intervals over an entire day. When we plot fewer totals (e.g., 10-20), we see clear trends. When we plot more totals, we see the stochastic variability plus the trend. We then can fit an arrival rate function by using a statistical smoothing algorithm, but note that the smoothling algorithm does not capture the evident spike around 9 am, evidently a startup effect. From such a plot showing the additional fluctuations, we can confirm that a Poisson process model (but a nonhomogeneous one!!) is reasonable. Figures **??**-**??** shows creative plots, where we divide the interval counts by the daily total.

2 Stochastic Arrival Process Models

2.1 An Arrival Process is a Point Process.

An arrival process is a special kind of stochastic process, often called a point process. See the Appendix for additional details and discussion.

2.1.1 Three Equivalent Alternative Representations of a Point Process

In an arrival process model, the arrival times are random times or random points in the positive half line (assuming that we start at time 0).

Point processes on the positive half line $[0, \infty)$ can be represented in three alternative ways: (i) in terms of A_n, the location of point n, as a function of $n \geq 1$, with $A_0 \equiv 0$, (ii) in terms of $A(t)$, the number of points in the interval $[0, t]$, as a function of $t \geq 0$, and (iii) in terms of X_k, the interval between point $k-1$ and point k, for $k \geq 1$. In terms of the intervals X_k, we write

$$A_k \equiv X_1 + \cdots + X_k, \quad k \geq 1, \quad \text{and} \quad A(t) \equiv \max\{k \geq 0 : A_k \leq t\}, \quad t \geq 0. \tag{1}$$

1. **arrival times** A_k, $k \geq 1$, (A_k is the **sum** of the first n interarrival times)
2. **arrival counting process** $A(t)$, $t \geq 0$,
3. **interarrival times** $X_k \equiv A_k - A_{k-1}$

2.1.2 The Basic Inverse Relation

In this setting, the two stochastic processes $\{A(t) : t \geq 0\}$ and $\{A_n : n \geq 0\}$ are inverse processes. This is easily seen by looking at a sample path of a counting process (such as a Poisson process). In such a plot, we show, $A(t)$, the number of points in the interval $[0, t]$ as a function of t. The sample path is a nondecreasing nonnegative integer-valued function of time t. We represent time t on the horizontal "x" axis, and the values of $A(t)$ on the vertical "y" axis. If instead, we look at the plot regarding the y axis as the domain and the x axis as the range, we see a plot of the locations of the successive points, A_n as a function of the discrete variable n. (We allow multiple points, but a common case is unit jumps in $N(t)$.)

2.2 The Arrival Rate

Let $\{A(t) : t \geq 0\}$ be a stochastic counting process. An important partial characterization is its mean, which is the cumulative arrival rate function

$$\Lambda(t) \equiv E[A(t)], \quad t \geq 0. \tag{2}$$

We assume that $\Lambda(t)$ is differentiable and that

$$\Lambda(t) = \int_0^t \lambda(s)\, ds, \quad t \geq 0. \tag{3}$$

The function $\lambda(t)$ is called the **arrival-rate function** of the point process. It is well defined in considerable generality, without requiring Poisson assumptions.

2.3 A Nonhomogeneous Poisson Process (NHPP)

The standard model for the arrival process over a day of a service system is a nonhomogeneous Poisson process (NHPP); see §5.4.1 of Ross [11], the standard textbook for IEOR 3106/4106.

An arrival (counting) process $\{A(t) : t \geq 0\}$ is an NHPP with arrival-rate function (intensity function) $\lambda(t)$, $t \geq 0$, if

1. **no batches:** Arrivals occur one at a time,
2. **Poisson distribution:** $P(A(t) = k) = \frac{e^{-m(t)}m(t)^k}{k!}$, ($A(t)$ has a Poisson distribution for all $t \geq 0$),
3. **the mean is the integral of the intensity:** $m(t) = \int_0^t \lambda(x)\,dx$ (The mean is a 1-dimensional integral over $[0, t]$.)
4. **independent increments:** $A(t_2) - A(t_1)$, $A(t_4) - A(t_3), \ldots, A(t_{2k}) - A(t_{2k-1})$ are mutually independent random variables for all k if the intervals $(t_1, t_2], (t_3, t_4], \ldots, (t_{2k-1}, t_{2k}]$ are disjoint subintervals of $[0, \infty)$, i.e., if $t_1 < t_2 \leq t_3 < t_4 \leq \cdots \leq t_{2k-1} < t_{2k}$

An NHPP becomes an ordinary Poisson process if $\lambda(t) = \lambda$, $t \geq 0$. The first condition above is actually implied by the others, but it is easier to check directly.

2.4 The (Stationary) Poisson Process

From plots of the arrival-rate function over a day, one should be convinced that the arrival rate is not constant. Nevertheless, the ordinary stationary Poisson process and associated classical queueing models that assume such a stationary arrival process are often applicable.

2.4.1 The Relevant Time Scale of an Arrival Process in a Queueing System

We have emphasized that it is useful to look at arrival processes in many different time scales. Each time scale has its own story. However, when we look at an arrival process as a component of a queueing system, then a particular time scale becomes relevant. To a large extent, **the service-time distribution in a queueing system determines the relevant time scale for the arrival process.** In particular, the relevant time scale for an arrival process in a queueing system tends to be from $1 - 10$ mean service times. Having identified that time scale, we want to look at the arrival process closely in that time scale.

If the mean service time is about 3 minutes, then it may be appropriate to look at the arrival process over intervals of 30 minutes or 1/2 hour. Over such shorter intervals, the arrival process might reasonably be regarded as approximately stationary. This insight is the basis for the standard **pointwise-stationary approximation (PSA)** approaches to staffing in call centers, using steady-state performance measures associated with stationary queueing models over short staffing intervals of length (1/4)-2 hours; e.g. see [5]. Over these shorter time intervals, the steady-state performance of the stationary model is used to approximate the performance of the queueing system (distribution of queue lengths and waiting times).

This perspective makes it important to consider the standard stationary Poisson process as well as the NHPP. However, given the strong time-varying arrival rate, we cannot expect stationary Poisson processes to fit perfectly. Careful statistical fitting with ample data will usually lead to rejecting the Poisson hypothesis, even though the Poisson arrival process model might be very useful.

2.5 Poisson Process, Chapter 5 of Ross [11]

The following are alternative ways to define a (stationary) Poisson process:

1. A Poisson process is a renewal process where the time between renewals has an exponential cdf, i.e., $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$, for some parameter $\lambda > 0$.

   IMPLICATION: As a renewal process, the interarrival times of a Poisson process are independent and identically distributed (i.i.d.), but unlike a renewal process, the interarrival times of a Poisson process are exponentially distributed as well. Non-Poisson renewal processes have different interarrival-time distributions.

2. A Poisson process is a counting process with unit jumps that has stationary and independent increments.

   IMPLICATION: Surprisingly, it turns out that, as a consequence of the stated properties, the distribution of each increment is necessarily Poisson, even though it is not assumed directly.

3. A Poisson process is a pure-birth stochastic process, i.e., a birth-and-death (BD) stochastic process (a special kind of continuous time Markov chain (CTMC), see Chapter 6 of Ross [11]) with all death rates equal to 0 and constant positive birth rates $\lambda_k = \lambda > 0$.

   IMPLICATION: As a CTMC and BD process, the Poisson process can be analyzed by that theory, but since the sample paths are nondecreasing, the process does not have a proper steady-state distribution. Hence, there is no CTMC steady-state theory to apply.

4. A Poisson process is an NHPP with the additional property that $\lambda(t) = \lambda$, $t \geq 0$.

   IMPLICATION: This is just emphasizing that a Poisson process is a special case of a NHPP. All properties of a NHPP necessarily hold for a Poisson process. A Poisson process is an NHPP with the extra property of stationarity, which manifests itself simply in the form of a constant arrival-rate function.

A key property of a Poisson process (not appearing in the first three definitions above) is: A Poisson counting process $A(t)$ has a **Poisson distribution** with mean λt for each $t \geq 0$, i.e.,

$$P(A(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \quad \text{for all} \quad k \geq 0. \tag{4}$$

Given (4) and the renewal-process representation listed first above, two probability distributions play a key role in Poisson processes:

2.5.1 Properties of a Poisson Distribution

1. A Poisson distribution has a single parameter, its mean, typically denoted by λ.

2. The variance of a Poisson distribution is equal to its mean.

3. As the mean of a Poisson distribution increases, the Poisson distribution is better and better approximated by a normal distribution with the same mean and variance. (There is a central limit theorem (CLT) for the Poisson distribution.)

2.5.2 Properties of the Exponential Distribution

1. An exponential distribution has a single parameter, its mean, typically denoted by λ^{-1}.
2. The variance of an exponential distribution is equal to **the square** of its mean.
3. The exponential distribution has squared coefficient of variation (SCV, variance divided by the square of the mean) of $c^2 = 1$.
4. The exponential distribution is the unique probability distribution on the nonnegative real line that has the lack of memory property:

$$P(X > s + t | X > s) = P(X > t) \quad \text{for all} \quad s \geq 0 \quad \text{and} \quad t \geq 0.$$

2.5.3 Properties of an NHPP (To be contrasted with a Poisson Process)

1. The interarrival times are *not* independent (but may be nearly so in short time intervals).
2. The interarrival times are *not* exponentially distributed (but may be nearly so in short time intervals).
3. The counts over intervals are still Poisson.
4. The counts over disjoint intervals are still independent.

3 Operations on Point Processes

3.1 Superpositions of Point Processes

A superposition of n point processes can be represented as the sum of the corresponding counting processes, i.e.,

$$A(t) \equiv A_1(t) + \cdots + A_n(t), \quad t \geq 0, \tag{5}$$

where the component stochastic processes $\{A_i(t) : t \geq 0\}$ are counting processes.

3.1.1 Good News: Superpositions of Independent Poisson Processes

Theorem 3.1 (*superpositions of independent Poisson processes*) *The superposition of n independent stationary (nonhomogeneous) Poisson processes is itself a stationary (nonhomogeneous) Poisson process process with an intensity equal to the sum of the component intensities.*

3.1.2 [Advanced topic] Bad News: Superpositions of Non-Poisson Processes

Even if the intervals between points are mutually independent in each component process in a superposition process, the intervals between points are dependent in the superposition process, unless all processes are Poisson process.

From renewal process theory, we know that a stationary renewal process is a delayed renewal process (the first interval has a different distribution from the others) in which the distribution until the first point has the equilibrium excess distribution. Give that the other intervals are i.i.d. with cdf F with mean $m < \infty$, the equilibrium excess cdf is

$$F_e(t) \equiv \frac{1}{m} \int_0^t (1 - F(s))\, ds, \quad t \geq 0. \tag{6}$$

A stationary renewal process is a special case of a stationary point process. The superposition of independent stationary point processes is always itself a stationary point process.

In general, the rate of a superposition process is the sum of the component rates. The following theorem summarizes both positive and negative properties of superposition processes.

Theorem 3.2 (*superpositions of independent renewal processes*) *The superposition of n independent stationary (ordinary) renewal processes is itself either a stationary or ordinary renewal process if and only if all the component processes and the superposition process are homogeneous Poisson processes.*

On the positive side, Theorem 3.2 states that the superposition of independent Poisson processes is itself a Poisson process (repeating the prior Theorem 3.1). On the other hand, if any of the component processes are non-Poisson renewal processes, either stationary or ordinary, then Theorem 3.2 states that the superposition process is neither a stationary renewal process nor an ordinary renewal process (and thus also not a Poisson process). That means that the successive intervals between points in the superposition process necessarily have dependence. Thus, dependence often plays a role in superposition processes.

3.1.3 More Good News: The Law of Rare Events

However, in applications we often have an arrival process be the superposition of a very large number of independent processes, each with relatively low rate, See §4.

3.2 Independent Splitting or Thinning of a Point Process

A single point process can be split into n other point processes by assigning each of the successive points to one of the designated n point process. Hence, each new split process contains a subset of all the points in the original point process. Every point is assigned to one and only one split process. The original point process is the disjoint union of the points in the split processes.

The points are said to be split independently with splitting functions $p_i(t)$, with $p_1(t) + p_n(t) = 1$ for all t, if any point at time t is assigned to split process i with probability $p_i(t)$, independent of the assignment of any other points.

Theorem 3.3 (*independent splitting of a Poisson process*) *The independent splitting of a stationary Poisson process with intensity λ into n split processes with splitting functions $p_i(t) = p_i$, $t \geq 0$, produces n independent Poisson processes with intensities $\lambda_i \equiv \lambda p_i$.*

Theorem 3.4 (*independent splitting of a nonhomogeneous Poisson process*) *The independent splitting of a nonhomogeneous Poisson process with intensity function $\lambda(t)$ into n split processes with splitting functions $p_i(t) = p_i$, $t \geq 0$, produces n independent nonhomogeneous Poisson processes with intensity functions $\lambda_i(t) \equiv \lambda(t)p_i(t)$.*

4 Why Poisson? The Law of Rare Events

We now provide theoretical explanation for why Poisson processes, stationary and nonstationary, have proven to work well as models of arrival processes for many service systems.

In applications, an arrival process is often a superposition of a very large number of independent point processes, each with relatively low rate. The component processes are the arrival

processes of individuals acting for the most part independently. Even though the superposition of a fixed number of component non-Poisson arrival processes is not Poisson, it approaches Poisson as the number of component processes increases. This explains why a Poisson process (possibly a nonstationary one) often provides a good fit to arrival processes.

Theorem 4.1 (*superpositions of many sparse point processes*) *The superposition of n independent i.i.d. stationary* (*nonhomogenous*) *point processes with intensities* λ/n ($\lambda(t)/n$) *converges in distribution to a stationary* (*nonhomogenous*) *Poisson process with intensity* λ ($\lambda(t)$) *as* $n \to \infty$.

In Theorem 4.1, the component processes need not actually be i.i.d., but each must be asymptotically negligible and the total rate must go to a fixed limit. This theorem can be understood by recalling the **Poisson approximation for the binomial distribution**, see §2.2.4 of Ross [11]. Here is a specific model: Let each of n customers choose to come to our service system during some specified time interval $[0, t]$ with probability p/n, making this choice independently of one another Let their arrival times during the time interval if they do elect to come, be uniformly distributed throughout the interval $[0, t]$. Then as $n \to \infty$, the arrival counting process $\{A(s) : s \geq 0\}$ over the interval $[0, t]$ converges to a Poisson process with arrival rate $\lambda \equiv p$. The expected number of arrivals in the interval $[a, b]$, where $0 \leq a < b \leq t$, is $p(b - a)$.

4.1 Practical Examples: Judging if Poisson Models should be Good or Not

The cases when a Poisson process or a NHPP is a good fit usually are due to Theorem 4.1. However, there are many examples where we can anticipate that a Poisson process or a NHPP might *not* provide a good fit. However, even in these cases, the Poisson model might prove to be very useful.

4.1.1 Examples where the Poisson and NHPP Models are Suspect

The following examples illustrate situations in which we can see that the axioms of the Poisson models are not satisfied. That does not entirely rule out the Poisson models, however.

1. scheduled arrivals

   Explanation: In many service systems, such as a doctor's office, the arrivals appear at scheduled times. Even though the arrival times fluctuate randomly around the scheduled times, usually a Poisson process or NHPP model does not fit well.

2. enforced separation

   Explanation: When considering landing times at airports, there are enforced intervals of separation between successive landings on each runway, such as 90 seconds. That is on top of scheduled arrivals. As a result the landing arrival process is typically not well modeled as a Poisson process or as a NHPP.

3. overflow processes

   Explanation: In finite capacity systems, when there is insufficient capacity for arrivals to be admitted immediately or soon enough, the arrivals may be allowed to overflow to other alternative service facilities. This was a classic phenomenon in telecommunication

networks. Since overflows tend to occur in bunches when the system is heavily loaded and then not at all in lightly loaded periods, overflow processes tend to be more bursty or highly variable than Poisson processes.

4. batch arrival processes

   Explanation: In many systems arrivals tend to occur in batches. Arrivals to restaurants, amusement parks and zoos typically come in batches. If the batches tend to stay together, then maybe the entire batch can be regarded as a single customer. However, in many cases, as for limited-capacity rides in an amusement park, a batch will use more resources than a single customer. Then it may be necessary to consider batch arrival processes.

4.1.2 Examples where the Poisson and NHPP Models Should Be Good

Theorem 4.1 indicates that Poisson models should be good for all of the following situations.

1. telephone calls at a call center

   Explanation: Usually these arrivals involve single customers operating independently of others.

2. arrival times at a supermarket or a bank

   Explanation: Usually these arrivals involve single customers operating independently of others.

3. walk-in arrivals in an emergency depertment

   Explanation: These often involve single patients without scheduling.

5 Relating NHPP's to Associated Poisson Processes

5.1 How to Simulate Poisson Processes and NHPP's

1. We can simulate a Poisson process with rate λ by generating the sequence of i.i.d. inter-arrival times with an exponential distribution having mean λ^{-1}. Then we can construct the associated arrival times and arrival counting process exploiting §2.1.1 above.

2. We can generate a NHPP with arrival-rate function $\lambda(t)$ by performing time-dependent independent thinning of a stationary Poisson process with rate $\bar{\lambda}$, where

$$\bar{\lambda} \geq \lambda(t) \quad \text{for all} \quad t.$$

   As the time-dependent thinning probability, we use

$$p(t) \equiv \frac{\lambda(t)}{\bar{\lambda}}.$$

5.2 How to Construct a NHPP from a Poisson Process

Let $\{N(t) : t \geq 0\}$ be a rate-1 Poisson process and let $\lambda(t)$ be an arrival-rate function of a desired NHPP $\{A(t) : t \geq 0\}$. We can easily construct the stochastic process $\{A(t) : t \geq 0\}$ given the two components $\{N(t) : t \geq 0\}$ and $\lambda(t)$. To do so, let $\Lambda(t)$ be the associated cumulative arrival rate function, defined by

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0. \tag{7}$$

The we define $A(t)$ for all $t \geq 0$ by setting

$$A(t) \equiv N(\Lambda(t)), \quad t \geq 0. \tag{8}$$

The following result is elementary:

Theorem 5.1 *The stochastic process* $\{A(t) : t \geq 0\}$ *constructed in* (8) *above is a NHPP with arrival-rate function* $\lambda(t)$*.*

Proof Independent increments extends. Look at the intensity:

$$E[A(t)] = E[N(\Lambda(t))] = \Lambda(t) = \int_0^t \lambda(s)\, ds, \quad t \geq 0. \quad \blacksquare$$

5.3 How to Transform a NHPP Into a Poisson Process

It might be convenient to transform a given NHPP into an associated Poisson process. To do so, we simply invert the construction carried out in the previous subsection. In particular, we define an inverse function $\Lambda^{-1}(t)$ of the cumulative arrival rate function $\Lambda(t)$ in (7) above.

Consider a NHPP $\{A(t) : t \geq 0\}$ with arrival-rate function $\lambda(t)$. Let $\Lambda(t)$ be the associated cumulative arrival rate function, defined by

$$\Lambda^{-1}(t) \equiv \inf\{s > 0 : \Lambda(s) = t\}, \quad t \geq 0. \tag{9}$$

We can now convert the NHPP $\{A(t) : t \geq 0\}$ into a rate-1 Poisson process by letting

$$N(t) \equiv A(\Lambda^{-1}(t)), \quad t \geq 0. \tag{10}$$

The following result is elementary:

Theorem 5.2 *The stochastic process* $\{N(t) : t \geq 0\}$ *constructed in* (10) *above is a rate-*1 *Poisson process.*

Proof Independent increments extends. Look at the intensity:

$$E[N(t)] = E[A(\Lambda^{-1}(t))] = \Lambda(\Lambda^{-1}(t)) = t, \quad t \geq 0. \quad \blacksquare$$

6 Poisson Arrivals See Time Averages (PASTA)

6.1 The Inspection Paradox

See §7.7 in Ross [11].

6.2 PASTA

With a Poisson arrival process, in great generality, the state of the system seen by an arrival is the same as seen at an arbitrary time.

A full understanding depends on the notions of stationary processes in continuous and discrete time, discussed briefly in Appendix §B.1.

7 Fitting a NHPP to Arrival Data

There are two issues:

(i) How can we fit an NHPP to data?

(ii) How can we test to see if an NHPP model is appropriate?

7.1 Estimating the Arrival-Rate Function

The parameters of the NHPP model are the arrival rates, To estimate the arrival rates, we partition the overall time interval into disjoint subintervals of fixed length, called bins. We then count the number of arrivals in each bin and plot it. The total number of arrivals in a bin is a direct estimate of the arrival rate per bin length for each time in the bin.

We often include data for multiple similar days. The arrival rate over that subinterval is then estimated by the sample average of the bin counts over the similar days. By this procedure, we produce an estimate of the arrival rate function $\lambda(t)$. Specifically, if we use bin sizes Δ, then we estimate $\lambda(t)$ by the piecewise-constant function

$$\bar{\lambda}(t) \equiv \frac{\bar{A}(a, a+\Delta)}{\Delta}, \quad a \le t \le a+\Delta, \tag{11}$$

where $\bar{A}(a, a+\Delta)$ is the average number of arrivals found in the interval $[a, a+\Delta]$.

However, caution is needed in looking at similar days, because days may appear to be more similar than they are. It is good to form confidence intervals for our estimates of arrival rates. There may be more day-to-day variability than we realize. That is illustrated by Figure **??**.

7.2 A Test for a Poisson Process

7.2.1 The Test

(a) Break up the day into shorter subintervals, over each of which the arrival rate will be considered constant.

(b) Consider one interval of length L; refer to it as $[0, L]$. But a virtue of this test is that the data over all the intervals can be combined (concatenated) after the transformation.

(c) Let n be the observed number of arrivals in the one (any one) interval.

(d) Let A_j be the j^{th} ordered arrival time within the interval $[0, L]$; i.e.,

$$0 < A_1 < A_2 < \cdots < A_n < L.$$

(e) Define new random variables

$$Y_j \equiv (n + 1 - j) \left(- \log_e \left(\frac{L - A_j}{L - A_{j-1}} \right) \right), \quad 1 \leq j \leq n. \tag{12}$$

(f) Under the null hypothesis that the arrival process is a Poisson process (with constant arrival rate) over the interval $[0, L]$ (with any fixed arrival rate), the random variables Y_j are independent and identically distributed (i.i.d.) random variables with a standard (mean-1) exponential distribution. [We explain in Section III below.]

(g) As a consequence, the data from all the intervals over the day can be combined. By combining the data from all intervals, starting at the beginning of the day, we obtain a single sequence of i.i.d. standard exponential random variables. Denote it by $\{Y_k : 1 \leq k \geq n\}$, where now n is the total number of arrivals.

(h) Since the arrival rate over any fixed interval we consider can be any fixed value, it could also be a random variable. We could include in our overall sequence data from different days with different arrival rates. In doing so, we should treat intervals on different days separately, making one long sequence.

7.2.2 Explanation

(a) First, we use the well-known **conditioning property of a Poisson process**; see §5.3.5 of Ross. If we condition upon the number of arrivals in a Poisson process over any subinterval, then those arrival times are distributed as independent and identically distributed random variables, each uniformly distributed over the interval. Thus the successive points we see are distributed as the uniform order statistics of i.i.d. uniform random variables; i.e., the first is the minimum, the second is the second smallest, and the last is the maximum.

(b) Second, we use a property of uniform order statistics, giving the **conditional distribution of the remaining variables, given the minimum**. Given n i.i.d. random variables uniformly distributed over an interval $[0, L]$, conditional on the minimum, say M_n, the remaining $n - 1$ random variables are i.i.d. random variables uniformly distributed on the interval $[M_n, L]$. We can repeat that construction to generate independent random variables.

(c) Given the recursive procedure in part (b) above, it suffices to convert the distribution of the minimum of i.i.d. uniform random variables into a standard exponential. For simplicity,

suppose that $L = 1$. Let us first observe that

$$P(M_n > 1 - t) = P(U > 1 - t) = t^n, \quad 0 \le t \le 1. \tag{13}$$

Lemma 7.1 *If M_n is the minimum of n i.i.d. uniform random variables on the interval $[0, 1]$, then*

$$P(-n \log_e (1 - M_n) > t) = e^{-t}, \quad t \ge 0.$$

Proof Using (13) in the last step,

$$\begin{aligned} P(-n \log_e (1 - M_n) > t) &= P(\log_e (1 - M_n) < -t/n) \\ &= P((1 - M_n) < e^{-t/n}) \\ &= P(M_n > 1 - e^{-t/n}) \\ &= (e^{-t/n})^n = e^{-t}. \quad \blacksquare \end{aligned}$$

References

[1] S. Asmussen. *Applied Probability and Queues*, second edition, Springer, 2003. [Advanced reference, concise treatment of the main probability topics related to queues. Good after having a couple of graduate probability courses.]

[2] R. B. Cooper *Introduction to Queueing Theory*, second edition, North Holland, Amsterdam, 1981. [A good introductory queueing book, with telecommunications flavor, available online. Has basic theory plus useful engineering methods, e.g., the equivalent random method.]

[3] Eick, S. G., W. A. Massey, W. Whitt. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** (1993a) 731–742. [This week will cover §1 (3 pages). Other highlights: Theorem 2, Corollary 4 and formulas (14) and (20).]

[4] Eick, S. G., W. A. Massey, W. Whitt. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* **39** (1993b) 241–252. [Goes into greater detail than reference above in the special idealized case of sinusoidal arrival rate functions. Explicit formulas in the sinusoidal case provide general understanding.]

[5] L. V. Green, P. J. Kolesar, W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16** (2007) 13–39. [Relatively readable introductory survey paper, approximately at the level of this course.]

[6] R. W. Hall. *Queueing Models for Services and Manufacturing*, Prentice-Hall, Englewood Cliffs, NJ, 1991. [The course textbook. This lecture relates to Chapters 3 and 6. This book tends to be different from most queueing textbooks, emphasizing important practical matters, but not carefully covering the basic queueing models. It is good to read this book AND an introductory queueing book, at least the relevant sections of [11], such as Chapters 5, 6 and 8.]

[7] Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. Server staffing to meet time-varying demand. *Management Sci.* **42** (1996) 1383–1394. [Shows how to apply [3] to provide practical solution to the staffing problem. This lecture will go through §2 and Figures 1-4 (3 pages).]

[8] L. Kleinrock. *Queueing Systems*, Wiley, 1975. [Excellent introductory queueing textbook with computer science and communication network perspective; first of two volumes.]

[9] W. A. Massey, W. Whitt. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* **13** (1993) 183–250. [More advanced reference, showing results for $M_t/GI/\infty$ models in [3, 4] extend to networks of these models.]

[10] B. Melamed, W. Whitt. On Arrivals That See Time Averages. *Operations Research*, vol. 38, No. 1, 1990, pp. 156–172. [Focuses on Poisson Arrivals See Time Averages (PASTA). See seminal paper by Wolff (1982) referenced here.]

[11] S. M. Ross. *Introduction to Probability Models*, 10^{th} edition, Elsevier, Amsterdam, 2010. [Used in IEOR 3106 and 4106. Has much useful material; e.g., Chapters 5, 6 and 8.]

[12] K. Sigman. *Stationary Marked Point Processes, An Intuitive Approach*, Chapman and Hall, New York, 1995. [More advanced reference, focusing on stationary point processes and the two representations, in continuous time and in discrete time. The generalization including marks, allows us to consider the service times together with the arrival process, which represents the full work brought to the system.]

[13] W. Whitt. *Stochastic-Process Limits*, Springer, New York, 2002. [More advanced reference, focusing on asymptotics, e.g., scaling and heavy-traffic limits for queueing systems.]

[14] R. W. Wolff. *Stochastic Models and the Theory of Queues*, Prentice-Hall, Englewood Cliffs, NJ, 1989. [A good introductory book on queueing theory.]

Appendix

This appendix contains additional material supplementing the main lecture notes.

A More on Arrival Process Models

A.1 What Makes a Good Model?

1. **purpose?** The model choice depends on the purpose. What do we want to do with the model?

   ANSWER: We are thinking of the arrival process as a component of a queueing model. In turn, we are thinking of that queueing model being used to predict staffing needs in the future, i.e., to decide how many resources we need (e.g., agents in a call center or beds in a hospital).

2. **analyzable?** To be useful, we must be able to work with the model. We must be able to perform useful analysis.

   ANSWER: We will highlight stationary and nonstationary (nonhomogeneous) Poisson arrival process models. See Chapter 5 of Ross [11] and Chapters 3 and 6 of Hall [6]. We will be showing how they can be analyzed. There are other arrival process models and associated ways to analyze them, but we will not get into that much.

3. **realistic?** The model should be suitably realistic. It should be consistent with the system and fit the data.

   ANSWER: We will see how to use the Poisson arrival process models with the call center data in the homework. One important issue is whether the arrival rate should be considered stationary or nonstationary. This actually a complicated issue. The same arrival process might be usefully modelled by a stationary Poisson process and by a nonhomogeneous Poisson process in different time scales. We will explain why the Poisson models make sense and we will consider tests to apply to data.

A.2 The Arrival Process as a Component of a Queueing Model

1. Often our goal is to use the arrival process as a **component** of a queueing model.

   IMPLICATION: The three questions above then apply to the potential queueing model. We now need to be able to analyze the queueing model with the arrival process model. We will highlight queueing models with stationary and nonstationary (nonhomogeneous) Poisson arrival processes. We will be showing how they can be analyzed.

2. Many service systems can be modeled, at least in part (i.e., one resource queue within the larger service system), by a multi-server queueing model. The classical multi-server queueing models are the Markovian **Erlang models**: $M/M/s/0$ (Erlang B or loss model), $M/M/s/\infty$ (Erlang C or delay model), or $M/M/s/\infty + M$ (Erlang A or delay model with customer abandonment). [These models are all birth and death processes; see §6.3 of the IEOR 3106/4106 textbook Ross [11].]

   NOTATION: The M in the notation stands for "Markov." In the $M/M/s/r$ model, the first M refers to the arrival process, meaning that the arrival process is a Poisson process

with rate λ; the second M refers to the service times, meaning that the service times are assumed to be independent of the arrival process and themselves independent and identically distributed (i.i.d.) with an exponential distribution having mean μ^{-1} (and thus service rate μ); the s means that there are s servers; and the r means that there are r extra waiting places in the queue. When the system has $s + r$ customers, s in service and r waiting, then there is no more space; then new arrivals are assumed to be blocked (leave without affecting future arrivals). Customers are assigned to servers as soon as the servers become available in order of the customer arrival times. The $+M$ in the $M/M/s\infty + M$ model refers to customer abandonment. The patience times (times to abandon) are assumed to be independent of the system history and are assumed to be exponentially distributed with mean θ^{-1}.

IMPLICATION: These models are discussed in introductory queueing textbooks such as [1, 2, 8, 14]. A quick introduction can be obtained from §§6.3, 8.3 and 8.9 of Ross [11]. This important background is mostly missing from our textbook Hall [6].

3. For time-varying arrival rates, only one model can be analyzed well exactly: the $M_t/GI/\infty$ infinite-server model (**IS model**). [See the paper distributed in class, [3], and Examples 5.18 and 5.25 of Ross [11].]

   NOTATION: The M_t means that the arrival process is a nonhomogeneous Poisson process (NHPP) with arrival-rate function $\lambda(t)$. The GI means that the service times are independent of the arrival process and themselves are i.i.d. random variables distributed as a random variable S with a general cumulative distribution function (cdf) $G(x) \equiv P(S \le x)$ having mean $E[S] = \mu^{-1}$. (We use $\equiv$ to denote equality by definition.) the ∞ means that there are infinitely many servers.

   IMPLICATION: With time-varying arrival rates, it is useful to exploit the IS model. The IS model serves as an offered-load (OL) model, showing how many resources would be used (and thus needed) if there were no limit on their availability. The number used is of course, random so we want to focus on its distribution, but the expected number used, denoted by $m(t)$, called **the offered load** is a vital partial characterization. That is a major theme of the course.

4. All the queueing models above have Poisson arrival processes, either stationary or nonstationary.

   IMPLICATION: We will focus on Poisson arrival processes, both stationary and nonstationary.

B More on Stochastic Point Processes

B.1 Stationary Point Process

There are two forms of stationary point processes, depending on whether we define stationarity in continuous time or discrete time. For more on this, see Sigman [12].

B.1.1 Counting Processes with Stationary Increments

A point process is said to be **stationary** in **continuous time** if the counting process $\{A(t) : t \ge 0\}$ has **stationary increments**. For the most part, a point process is stationary in contin-

uous time if the arrival rate function is a constant function. For example, a nonhomogeneous Poisson process (NHPP) is a Poisson process that is also a stationary point process.

MORE CAREFUL DEFINITIONS: An increment in continuous time is the count over an interval, such as $A(s+t) - A(s)$, which counts the number of points in the interval $[0, s+t] - [0, s] = (s, s+t]$, open on the left and closed on the right. We use the convention of open on the left and closed on the right to ensure that each time point is in only one interval. We look at the counting process over increment when we divide time into subintervals, called bins, and count the number of arrivals in each bin. In that measurement process, we want to put each arrival in one and only one bin. A single increment of a counting process (which is a random variable) is stationary if its probability distribution is independent of time shifts; i.e., if the distribution of $A(s+t) - A(s)$ is independent of $s \geq 0$.

More generally, we say that a point process is stationary if the joint distribution of any k increments is independent of a time shift; i.e., the joint distribution of $(A(s+t_2) - A(s+t_1), A(s+t_4) - A(s+t_3), \ldots, A(s+t_{2k}) - A(s+t_{2k-1})$ in $\mathbb{R}^k$ is independent of $s \geq 0$ for $0 < t_1 < t_2 \leq t_3 < t_4 \leq \ldots \leq t_{2k-1} < t_{2k}$. [That is a technical refinement. The main point concerns one increment.]

A point process that is stationary in continuous time necessarily has a constant arrival rate function.

B.1.2 Point Processes with Stationary Increments in Discrete Time

A point process is said to be **stationary** in **discrete time** if the discrete-time sequence of arrival times $\{A_k : k \geq 0\}$ is a **stationary sequence**, i.e., if the joint distribution of the random vector $(A_{j+n_1}, A_{j+n_2}, \ldots, A_{j+n_k})$ in $\mathbb{R}^k$ is independent of $j \geq 0$, where j is an integer and $n_1 < n_2 < \cdots < n_k$.

B.2 Counting Processes with Independent Increments

A point process is said to have independent increments (understood to be in continuous time) if any k disjoint increments, each of the form $A(s+t) - A(s)$, are mutually independent.

B.3 Renewal Processes, see Chapter 7 of Ross [11]

An arrival (point) process is a renewal process if the sequence of interarrival times $\{X_k : k \geq 1\}$ are independent and identically distributed (i.i.d.) with some interarrival-time cumulative distribution function (cdf) $F(x) \equiv P(X_k \leq x)$, $x \geq 0$. (The renewal process has stationary increments and independent increments in discrete time.) Note that the interarrival times are independent in a renewal process, but the renewal counting process typically (in general) does not have independent increments. A renewal counting process turns out to have independent increments in continuous time if and only if it is a Poisson process.

B.3.1 [Advanced Topic] An NHPP as a Poisson Random Measure on $\mathbb{R}^k$

A NHPP can also be defined as the special case of a Poisson random measure on Euclidean space $\mathbb{R}^k$ when $k = 1$, i.e., when the space is the positive half line $[0, \infty)$.

A random counting measure $A(B)$ in $\mathbb{R}^k$ counts the random number of points in each subset B in $\mathbb{R}^k$, with the usual properties of a measure as a function of sets B. A random counting measure is a Poisson random measure on $\mathbb{R}^k$ (or a Poisson process on the Euclidean space $\mathbb{R}^k$) determined by an intensity function λ if

1. $P(A(B) = k) = \frac{e^{-m(B)}m(B)^k}{k!}$, i.e., if $A(B)$ has a Poisson distribution with mean $m(B)$ for all B,
2. $A(B_1), A(B_2), \ldots A(B_k)$ are mutually independent random variables for all k if the sets $B_1, B_2, \ldots, B_k$ are disjoint subsets of $\mathbb{R}^k$
3. $m(B) = \int_B \lambda(x)\, dx$ (the mean number of points in set B is a k-dimensional integral)

For an example, see Exercise 5.94 of [11], where the intensity is constant.

A Poisson random measure on the positive half line, $[0, \infty)$ is a nonhomogeneous Poisson process. Now the possible subsets are subsets B of $[0, \infty$. It suffices to focus on intervals $0, t]$. We can define a counting process $A(t)$ as a function of $t \geq 0$ in terms of a random measure if we identify $A(t)$ with $A([0, t])$. With this new notation, the properties in this subsection reduce to those in the previous direct definition.

C Application (later in course): Staffing with a NHPP Arrival Process

We now show how the NHPP can be used to allocate resources, i.e., determine a candidate staffing function $s(t)$.

C.1 Performance of the $M_t/GI/\infty$ model, §1 of [3]

C.1.1 The Model Can Be Analyzed!

For an IS model with i.i.d. service times having a general service-time cdf $G(x) \equiv P(S \leq x)$ that are independent of an NHPP arrival process, the number in system at any time t has a Poisson distribution with the single parameter, its mean $m(t)$, which has a convenient explicit expression; see §1 of [3]. (Surprisingly, this nonstationary model can be analyzed!!)

C.1.2 The Offered Load

The IS model is useful because it shows the number of servers that would be used if there were unlimited resources. Of course that number is actually random, and so is not fixed, but we know its full distribution and its mean. Thus, this explicit performance of the IS model helps us understand the performance of associated models with finitely many servers. The mean shows the expected number of servers that would be used if there were unlimited resources. Thus, the mean $m(t)$ is often called **the offered load**. It is often denoted by $R(t)$ by Mandelbaum.

C.2 Staffing in the $M_t/GI/s_t/\infty$ and $M_t/GI/s_t/\infty + GI$ models, §2 of [7]

We can use the IS model as a basis for staffing. Since the number in system in the IS model has a Poisson distribution with mean $m(t)$, we can approximate by a normal distribution with mean $m(t)$ and variance $m(t)$. (The variance of a Poisson distribution equals its mean.) As a rough engineering approximation we can use the **square root formula**, we can staff at

$$s(t) \equiv m(t) + \lceil \beta \sqrt{m(t)} \rceil, \tag{14}$$

where $\lceil x \rceil$ is the least integer greater than x and β is a Quality-of-Service (QoS) parameter. A simple conservative choice is $\beta = 2$. That amounts to staffing at two standard deviations above the mean.

There are various refinements, such as the so-called modified offered-load (MOL) approximations, but in view of the typically rough model fit, they tend to be second-order.

D [Advanced topic] Forecasting

There are whole courses devoted to this topic, under time series analysis. Generally, the goal is to predict the arrival rate function at some future time, e.g., over the day on some day in the near future, tomorrow or next week, say, so that we can then make plans for staffing. For a NHPP, the goal is to predict the arrival-rate function to use with the NHPP. Given the arrival-rate function, we have a completely defined NHPP model.

In predicting the arrival-rate function, we want to account for seasonal effects (month of the year), daily effects (day of the week), special days (e.g., holidays and major sale events), and time-of-day effect. For a NHPP, the goal is to predict the arrival-rate function to use with the NHPP.

With call center data, we can choose to look at it over different resolutions (time scales) in order to see these different effects.

D.1 Time Series Methods

1. Moving average
2. Weighted moving average
3. Exponential smoothing
4. Autoregressive moving average (ARMA)
5. Autoregressive integrated moving average (ARIMA), e.g. Box-Jenkins
6. Extrapolation
7. Linear prediction
8. Trend estimation
9. Growth curve