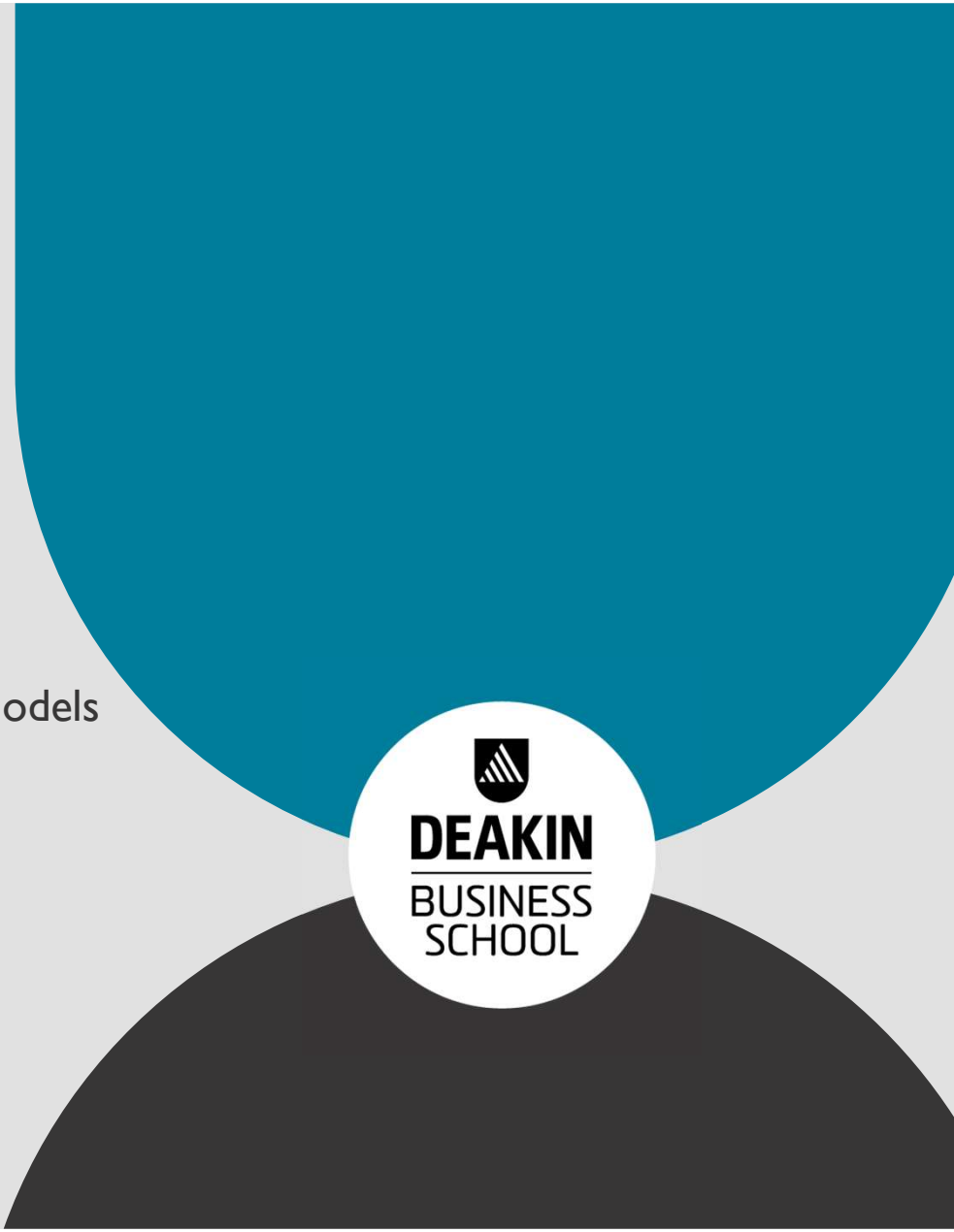
MIS775 Decision Modelling for Business Analytics

TOPIC 10: Applications of Simulation Modelling – Queueing Models



AACSB ACCREDITED

Recap

- Stochastic modelling produces more realistic models that capture and highlight risk
- We discussed:
  - **Using empirical distributions to generate stochastic input data**
  - **Capturing probabilistic outcomes using simulation**
  - **Analysing simulation output**
  - **Model accuracy and limitations of simulation**

Application: Queueing Models

- This topic considers the application of decision models to queueing
- Capacity problems are very common in industry and are one of the main drivers of process redesign
  - **Need to trade-off the costs of increasing capacity against the costs from customers becoming more dissatisfied as they wait for service**
- The objective is to find the "best" system (minimise total cost)

Learning Objectives

- This topic considers the application of decision models to queueing
  - **The elements of queueing models**
  - **The Exponential distribution**
  - **Important queueing relationships**
  - **Analytic steady-state queueing models**
    - **The basic single-server model**
    - **The basic multiple-server model**

Textbook reading: Chapter 12 (12.1-12.5)

Analytical vs Simulation Modelling

- Two approaches to modelling queueing systems
  - **Analytical (mathematical)**
  - **Simulation**
- Analytical approach relies on simplifying assumptions which may not be realistic
- Simulation doesn't require unrealistic assumptions, so more realistic systems can be analysed, but it requires specialised software

Elements of Queueing Models

- Queueing models involve customers arriving and entering a system, probably waiting in one or more queues, being served, then departing
- Inputs are average arrival rate and average service time
- Outputs typically include average waiting time in queues, average queue length, and the proportion of time that servers are busy

Characteristics of Arrivals

Arrival process includes the timing of arrivals and types of arrivals

- Inter-arrival times (i.e. time between successive customer arrivals) are usually random with some probability distribution
- Customers can arrive one at a time or in groups
- How long customers are prepared to wait
  - **E.g. 1. A customer might decide not to join a line on seeing too many customers waiting**
  - **E.g. 2. A waiting customer might become impatient and leave before being served**

Service Discipline

- The service discipline is the rule that states which of the customers waiting in the queueing system is served next
- Most common is first-come-first-served
- Another aspect of the waiting process is the number of lines
  - **E.g. 1, banks have a single line; customer are served in order of arrival**
  - **E.g. 2, supermarkets have multiple lines**

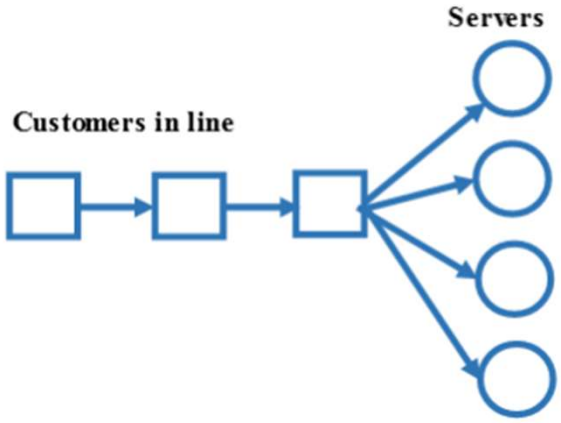
Service Characteristics

- Service times are typically random, and so the probability distribution must be specified
- The probability distribution can be the same for all customers and servers, or it can depend on the server and/or the customer

Service Characteristics

- In a bank, customers join a single line and are served by the first available teller. The servers (tellers) are said to be in parallel

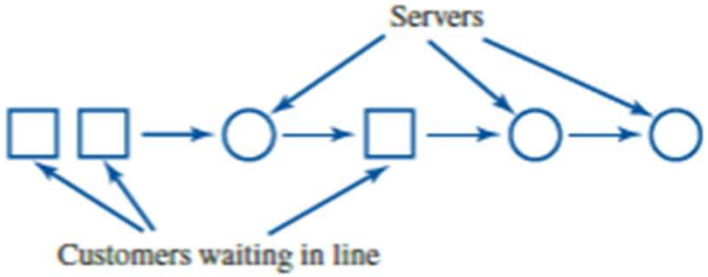
Figure 12.1
Queueing System with Servers in Parallel



Service Characteristics

- A different type of service process found in manufacturing is a **queueing network** where parts ("customers") enter a system with several machines ("servers")
- Simplest type of queueing network is a series system

Figure 12.2
Queueing System with Servers in Series



- Each machine has its own service time distribution, and a part might have to wait in line behind the machines on its routing

Short-run vs Steady-state Behaviour

- If you run a fast-food restaurant, you are particularly interested in the queueing behaviour during the peak lunchtime period
- Arrival rate during this period increases sharply, and you would probably employ more workers to meet the increased customer load
- In this case, your primary interest is in the **short-run behaviour** of the system - the next hour or two
- Unfortunately, short-run behaviour is difficult to analyse with analytical models

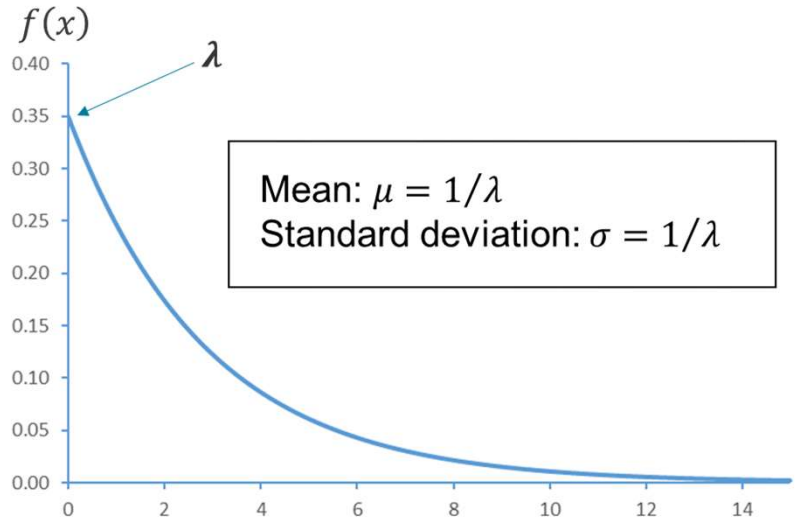
Short-run vs Steady-state Behaviour

- Analytical models are best suited for studying long-run behaviour, when the effects of initial conditions have dissipated
- This type of analysis is called **steady-state analysis** and is our main focus
  - **One requirement for steady-state analysis is that the parameters (e.g. arrival rate) remain constant**
  - **Another requirement is that the system must be stable, i.e. servers must serve fast enough to keep up with arrivals - otherwise, the queue can become unmanageable**

Exponential Distribution

- The most common probability distribution used to model inter-arrival times and service times is the **Exponential distribution** (see Topic 8)
- It has a single parameter, λ, the expected number of arrivals or expected number served per time interval
- E.g. If *X* is the time to serve a customer, in minutes, and the mean service time is three mins, then $1/\lambda = 3$, so that $\lambda = 1/3$

  λ can be interpreted as the average rate at which customers are being served - in this case, an average of one customer every 3 mins



Memoryless Property

- The property that makes the Exponential distribution so useful in queueing models is its **memoryless property**:

$$Pr(X > x + h | X > x) = Pr(X > h)$$

- Interpretation: if a server has already been serving a customer for x minutes, then the chance that the service continues for at least another h minutes is the same no matter what the current length of service (x) is. In a sense, the probability "starts over"
- Exponential is the only continuous distribution with this property

Poisson Process

- When inter-arrival times are Exponentially distributed, we often state that "arrivals occur according to a Poisson process"
- There is a close relationship between the Exponential distribution, which measures times between events such as arrivals, and the Poisson distribution, which counts the number of events in a certain length of time (See Topic 8)
- If customers arrive at a bank according to a Poisson process with rate one every three minutes, this implies that the inter-arrival times are Exponentially distributed with parameter $\lambda = 1/3$

Testing the Exponential Assumption

- The Exponential distribution is a simplifying assumption, but its properties don't necessarily align with the real world
  - **E.g. most services require some minimum time in which to be completed, whereas the Exponential has no such requirement**
- We therefore apply a "goodness of fit" test* to see if the Exponential provides an adequate fit to the data, before making this assumption

*Tests such as the Chi-square goodness of fit test are not covered in this unit

Important Queueing Relationships

- There are some general relationships that hold for a wide variety of queueing models
- We typically calculate two general types of outputs in a queueing model - **averages over time** and **averages across customers**

Averages Over Time

- L_Q is the expected no. of customers in the queue
  - **Estimate L_Q by counting the no. of customers in the queue at randomly selected times, and finding the average**
- L_S is the expected no. of customers being served
- $L = L_Q + L_S$ is the expected no. in system (i.e. in queue or in service)
- $\boldsymbol{Pr(All\ idle)}$ is the probability that all servers are idle
- $\boldsymbol{Pr(All\ busy)}$ is the probability that all servers are busy
  - **To estimate $Pr(All\ busy)$ observe the servers at random times and calculate the proportion of times that all servers are busy**

Averages Across Customers

- W_Q is the expected time a customer will spend in the queue
  - **Estimate W_Q by averaging waiting times over a random sample of customers**
- W_S is the expected time it takes to serve a customer
- $W = W_Q + W_S$ is expected time a customer will spend in the system

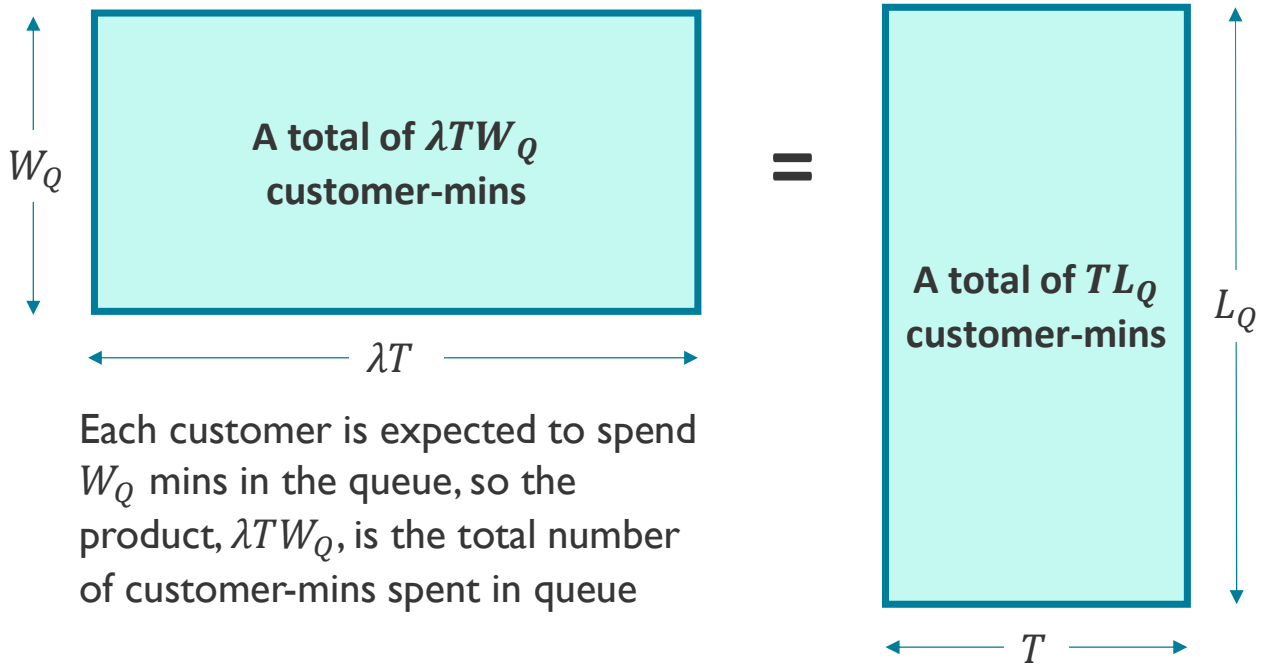
Little's Formula

- Consider any queueing system **in steady state**
  - λ is the average rate at which customers enter the system
  - L is the expected number of customers in the system
  - W is the expected time a customer spends in the system
- Little's formula, $L = \lambda W$, relates customer averages and time averages
- It also applies to the queue and service components:

$$L_Q = \lambda W_Q$$
$$L_S = \lambda W_S$$

Little's Formula Explained

- λ is the arrival rate (say, in mins). Then over a long time period, T mins, λT customers will enter the system



Each customer is expected to spend W_Q mins in the queue, so the product, $\lambda T W_Q$, is the total number of customer-mins spent in queue

But over that same period, T, the expected number of customers in the queue is L_Q, giving another calculation of the total number of customer-mins

Equating the two expressions, and cancelling out T, gives

$$L_Q = \lambda W_Q$$

Server Utilisation

- An important queueing measure is server utilisation, U
- It is defined as the long-run proportion of time a server is busy
- In a multiple-server system, where there are s identical servers in parallel, server utilisation is defined as $U = L_S/s$, i.e. the expected number of busy servers / no. of servers
  - **E.g. if $s = 3$ and $L_S = 2.55$, then $U = 0.85$ so the three servers are busy about 85% of the time**

Analytical Steady-state Queueing Models

Basic Single-server Model

- Most basic single-server model is the M/M/1 model
  - **First M means inter-arrival times are Exponentially distributed**
  - **Second M means the service times are Exponentially distributed**
  - **The "1" implies that there is a single server**
- Denote the arrival rate by λ and service rate by μ, and define the **traffic intensity** by (rho) $\rho = \lambda/\mu$ and assume $\rho < 1$ (necessary for a steady-state solution). Then

$$\boldsymbol{Pr(n\ customers\ in\ system) = (1-\rho) \times \rho^n}$$

- This gives the expected no. of customers in the system, L

Basic Single-server Model Formulas

- The expected no. of customers in the system is $L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$
- From Little's formula, the expected time in the system is $W = L/\lambda$
- Expected time being served is $W_S = 1/\mu$
- Expected waiting time in the queue is $W_Q = W - W_S = W - 1/\mu$
- From Little's formula, expected queue length is $L_Q = \lambda W_Q$
- Server utilisation, $U = Pr(server\ is\ busy)$

$$= 1 - Pr(0\ customers\ in\ system)$$

$$= 1 - (1 - \rho) = \rho$$

Example

Suppose an M/M/1 queue has an arrival rate of 0.75 customers per min. and a service rate of 1 per min. Then:

- Expected no. of customers in the system is $L = \frac{\lambda}{\mu - \lambda} = \frac{0.75}{1 - 0.75} = 3$ customers
- Expected time in the system is $W = L/\lambda = \frac{3}{0.75} = 4$ mins
- Expected time being served is $W_S = 1/\mu = 1$ min
- Expected waiting time in the queue is $W_Q = W - W_S = 4 - 1 = 3$ min
- Expected queue length, $L_Q = \lambda W_Q = 0.75 \times 3 = 2.25$ customers
- Probability server is busy, $U = \rho = \lambda/\mu = 0.75$
- Probability that an arriving customer has to wait for service $= U = 0.75$

Basic Multiple-server Model

- Many service facilities (e.g. banks) employ multiple servers that work in parallel, so each customer is served by one server and then departs
- The simplest version of this **multiple server parallel system**, the M/M/s model, where s denotes the number of servers
- There are three inputs to this system:
  - **Arrival rate λ**
  - **Service rate (per server) μ**
  - **Number of servers s**
- To ensure the system is stable we require the traffic intensity, now defined by $\rho = \lambda/(s\mu)$, to be less than 1

Comparison of Models

- Would you rather go to a system with one fast server or a system with several slow servers?
- In the latter case, we assume that only one waiting line forms, so that you can't get unlucky by joining the "wrong" line.
- The solution to the question is fairly straightforward, once we know how to obtain outputs for *M*/*M*/1 and *M*/*M*/s models

The Effect of Traffic Intensity

- We have mentioned that for an M/M/1 or M/M/s system to be stable, the traffic intensity must be less than 1
- In words, the system must be able to service the customers faster than they arrive; otherwise, the queue length eventually grows without limit
- It is interesting to see what happens to a system when the traffic intensity gets closer and closer to 1 but stays less than 1 (seminar task)

Other Exponential Models

- The basic M/M/s model and its special case, the M/M/1 model, represent only two of numerous analytical queueing models that researchers have studied
- Some of these are relatively simple extensions of the models such as the limited waiting room model
- It starts with the basic M/M/s (or M/M/1) model but assumes that any new arrivals are turned away when the number already in the queue is at some maximum level (e.g. 10)
- Stability is not a concern with this system. That is, there is no need to require that a traffic intensity be less than 1 to ensure stability

Queueing Simulation Models

- A popular alternative to using the analytical models is to develop queueing simulations
- There are several advantages to using simulation
- The most important advantage is that you are not restricted to the assumptions required by the standard analytical queueing models
- A second advantage of queueing simulation is that you can see how the system behaves through time
- The downside of queueing simulation is that it requires a programmer, or a specialised software package

Conclusion

- This topic has introduced two basic approaches for analysing queueing systems.
- The first is the analytical approach, where the goal is to find formulas to calculate steady-state performance of the system
- The second is the simulation approach, where the random elements of the system are generated and then the events are played out as they occur through time

Module 2 Learning Outcomes

1. Building effective decision models in spreadsheets
2. Treatment of stochastic variables and modelling uncertainty
3. Stochastic modelling in spreadsheets
4. Risk analysis in spreadsheets
5. Steps in decision analysis
6. Sensitivity analysis and risk profile in decision analysis
7. Investigated queueing models as an advanced application

Next Class

- **Topic 11: Review the sample exam paper and solutions**