PROBLEMS OF STOCKS

1 Introduction

1.1 General, 1.2, Characteristics of the problems of stocks, 1.3, Graphical representation, 1.4, Replenishment, 1.5, Replenishment delays.

— Arnold KAUFMANN, 1970, “Méthodes et modèles de la Recherche Opérationnelle”, Vol. I, 2.nd. edition, Dunod, Paris, p 165, Chapter IV, “Les Problèmes de stocks”

1.1 General

The supply of materials and equipment required for a manufacturing process, the customers’ orders, the reasonable availability of reserve parts incur varied problems. It is difficult to make a coherent and logical classification of the problems of stocks. The nature of demand should, however, be considered first:

- Determined (predictable with a certain accuracy);
- Random, but statistically stable;
- Random, but statistically unstable (seasonal);
- Unknown.

In stock problems, there can be constraints:

- Interactions between the various products;
- Limitations of means (volume, weight, financial availability, etc.).

Each time, an economic function will be defined to be optimized, which will often be, when demand is random, in the form of a mathematical expectation of global cost.

1.2 Characteristics of the problems of stocks

Given the variety of recognized stock problems in industrial practice or other areas, just a review of main cases will be done, to identify some simple concepts. The stock problems present themselves in the form of wait phenomena of a particular nature. Rather than assuming (as is done in the theory of queues) that units arrive one by one, it will be assume that arrivals relate to sets of units. The phenomena will be studied with support on probability, but in certain cases, otherwise frequent, in which variances are weak, deterministic models can be associated with them. All the problems of stocks include:

(1) A demand for certain articles, which is generally a random function of time, but may also be known and determined.

(2) The existence of a stock of the items to meet demand, which runs out and has to be replenished. The replenishment can be continuous, periodic or done at any intervals.

(3) Costs associated with these operations, investments, depreciation, insurance, various risks, storage, etc., and also one, more or less arbitrarily, assigned to stockout, which is essential in some problems. These costs allow to establish an economic function that we intend to optimize.

(4) Objectives to achieve or constraints involved as a consequence of the nature of the problem.

1.3 Graphical representation

In order to describe a problem of stocks, it is convenient to use the representation given in Figure 1, in which appear the initial stock, S_i, the final stock, S_f, the interval q separating them. In general, demand quantities are random, represented by steps. Often this path is replaced by a straight line or a curve which will give an easier analytical description of demand.

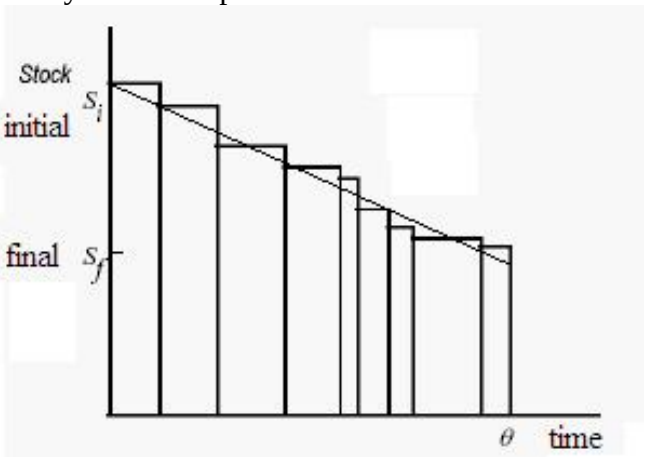

Figure 1

1.4 Replenishment

Suppose that the time interval between the issuance of the order to replenish and the reception is zero (negligible). Two main methods of basic inventory management are used. The first is called ***method by periods.*** A period T is established after which the replenishment is carried out systematically. This method has the drawback of risk of stockout and can lead to a costly management, but has the advantage of being automatic. The second may be termed a ***method of relaxation*** by analogy with physical phenomena of the same nature: the amount provided is constant, but the intervals T_1, T_2, T_3, ..., are no longer equal. There is no risk of stockout, the administration is generally less expensive, but not so easy to become systematic.

1.5 Replenishment delays

Suppose that the replenishment delay (time interval between issuing the order and reception) is independent of the amount ordered, i.e. constant and of duration t. Compare what would occur by either method. In the first (method by periods, T constant), the date of issue of order is known and it is necessary (to determine the quantity to be ordered) to extrapolate what was ordered in the interval $T - t$ preceding t; in certain cases, t can even be greater than T. In the second method (relaxation, several T_i), however, the quantity to order is constant, but the date of issue is unknown and has to be determined through extrapolation, which is sometimes insufficiently precise; in some cases, $t > T_i$. In general, the demand is known in probability. Sometimes, the delay is proportional to or a function of the order, which complicates the situation.

A method widely used for the management of stocks is to issue an order of constant size as soon as the stock reaches a critical value or *replacement level.* This may be called the *two-bin system* (“system of two boxes."). This method offers the advantage of a convenient management, but does not always guarantee against stockouts with sufficient probability.

2. Study of simple cases, proportional costs

2.1, First case: search for the economic (optimal) order quantity; **2.2**, Numerical example, **2.3**, Second case: EOQ with cost of shortage; **2.4**, Numerical example, **2.5**, Third case: random demand with loss on surplus and additional shortage cost (storage cost negligible) **2.6**, Numerical example; **2.7**, Search for the shortage cost; **2.8**, Resolution, **2.9**, Fourth case: random demand with costs of storage and shortage; **2.10**, Numerical example; **2.11**, Resolution by numerical calculation, **2.12**, Fifth case: known demand with storage cost proportional to the price of sale or purchase; **2.13**, Numerical example. (...)

Only the first and fourth cases will be briefly addressed below.

2.1 First case: the economic order quantity

Suppose parts of a certain model that are subject to constant demand, *h* parts per unit time, and stockout is not allowed. The parts are acquired in orders or lots. Suppose that a *fixed* cost of ordering, regardless of the number of parts, is c_L. [1]. The cost of storage of a part per unit time (day, for example) is c_S. The demand for a total *q* time interval, under study (*e.g.*, one year), is *N*. Assuming that all orders contain the same number of parts, *n*, the question is what value to give *n* so that the overall cost of ordering and storage of parts *N* is minimal (excluding the cost of the parts themselves). The number *r* of orders and the period *T* of replacement of the stock will also be determined.

The average level of the stock during a period *T* is $n/2$ (*n* in the beginning, 0 in the end). The storage cost during this period is thus $\frac{1}{2} n c_S T$. The total cost of an order is

$$c_L + \frac{1}{2} n c_S T \qquad \{1\}$$

Moreover, it is

$$n = h T \qquad \{2\}$$

and

$$r = \frac{N}{n} = \frac{q}{T} \qquad \{3\}$$

The total cost for the time interval *q* is:

$$\begin{aligned} z &= \left(c_L + \frac{nT}{2} c_S \right) r = \left(c_L + \frac{nT}{2} c_S \right) \frac{N}{n} = \\ &= \frac{N}{n} c_L + \frac{NT}{2} c_S = \\ &= \frac{N}{n} c_L + \frac{q}{2} c_S n \end{aligned} \qquad \{4\}$$

So, *z* depends on the variable *n*, the other parameters, *N*, *q*, c_L and c_S being known. The minimum *z* (obtained by differentiating or recalling that in the above form the two quantities must be equal)[2] occurs for

---

[1] *L* for "launch".

[2] See Appendix.

$$n_0 = \sqrt{2\frac{N}{q}\frac{c_L}{c_S}} \qquad \{5\}$$

which is the optimum size sought. Substituting $n = n_0$ in $\frac{N}{n} = \frac{q}{T}$, we have

$$T_0 = \frac{n_0}{N}q = \sqrt{2\frac{N}{q}\frac{c_L}{c_S}}\frac{q}{N} = \sqrt{2\frac{q}{N}\frac{c_L}{c_S}} \qquad \{6\}$$

$$\left([T_0] = \sqrt{\frac{\mathrm{T}}{\mathrm{u}}\frac{\$}{\$/(\mathrm{u}.\mathrm{T})}} = \mathrm{T}\right) \qquad \{6a\}$$

and, as total cost, from Eq. {4},

$$z_0 = \sqrt{2Nq\,c_L c_S} \qquad \{7\}$$

$$\left([z_0] = \sqrt{\mathrm{u\,T\,\$\$/(u.T)}} = \$\right) \qquad \{7a\}$$

2.2 Numerical example

A manufacturer receives an order for N = 120 000 parts, to be delivered in one year (q = 360 days). At what rate should he replenish his stock, if delay is not permissible in delivery ?

See "plate" http://web.ist.utl.pt/mcasquilho/compute/or/Fx-eoq.php . In the plate, the nomenclature is

Here		There	
N	demand in period	d	120e+3
c_L	setup cost	K	30e+3 $
(any)	purchase cost	c	1
c_S	holding cost	h	0,35 $/d × 360 d = 126 $

The demand in this case is at a constant rate. The costs are:

$$c_S = 0{,}35\,\$ \text{ / day} \qquad c_L = 30\,000\,\$ \qquad \{8\}$$

We have:

$$n_0 = \sqrt{2\left(\frac{120000}{360}\right)\left(\frac{30\mathrm{E}3}{0{,}35}\right)} = 7559(,3) \text{ parts} \qquad \{9\}$$

(Although it is not *a priori* important in this case, it should be numerically verified if n_0 is to be rounded down or up, examining the consequences in T_0 and, essentially, z_0).

$$T_0 = \frac{360 \times 7559}{120000} = 22{,}68 \text{ days} \qquad \{10\}$$

$$z_0 = \sqrt{2 \times 120000 \times 360 \times 30\mathrm{E}3 \times 0{,}35} = 952470 \qquad \{11\}$$

(This cost refers to one year.)

Another example, perhaps with more realistic data, is as follows (*in* Tavares *et al.* [1996], p 163), with its own nomenclature.

Annual demand	r = 1200 kg / year
Unit cost of purchase	C_1 = 20 \$ / kg
Fixed cost of ordering	A = 15 \$
Unit cost of possession	C_2 = 25 % of C_1 per year = 5 \$ / kg-year

In the notation presented above (Kaufmann's):

Total demand (per year)	N = 1200 kg
Time span	q = 1 year
Fixed cost of ordering	A = 15 \$
Unit purchase cost	C = 20 \$ / kg
Fixed cost of ordering	c_L = 15 \$
Cost of storage (per unit)	c_s= 25 % of C per year = 5 \$ / kg-year

We find, as solutions to the various variables of interest:

$$n_0 = \sqrt{2\frac{1200(\text{kg})\times 15(\$)}{1(\text{yr})\times 5(\$/\text{kg-yr})}} = 84{,}9 \text{ kg} \qquad \{12\}$$

$$T_0 = \sqrt{2\frac{q}{N}\frac{c_\ell}{c_s}} = \sqrt{2\frac{1(\text{yr})}{1200(\text{kg})}\frac{15(\$)}{5(\$/\text{kg-yr})}} = 0{,}0707 \text{ year} = 25{,}5 \text{ day} \qquad \{13\}$$

$$z_0 = \sqrt{2\times 1200(\text{kg})\times 1(\text{yr})\times 15(\$)\times 5(\$/\text{kg-yr})} = 424{,}3 \text{ \$} \qquad \{14\}$$

The annual cost of the material, not included in the model, is NC = 1200 × 20 = 24 000 \$, so (after the optimization) the maintenance charges represent 424 / 24 000, or 1,8 % of that cost. Specifically, we would lead T_0 to a reasonable value (21 days, 28, 30, "1.st day of each month", etc.). In Figure 2 is plotted z depending on the size of the order n to monitor the increase of z for non-optimal values of n.

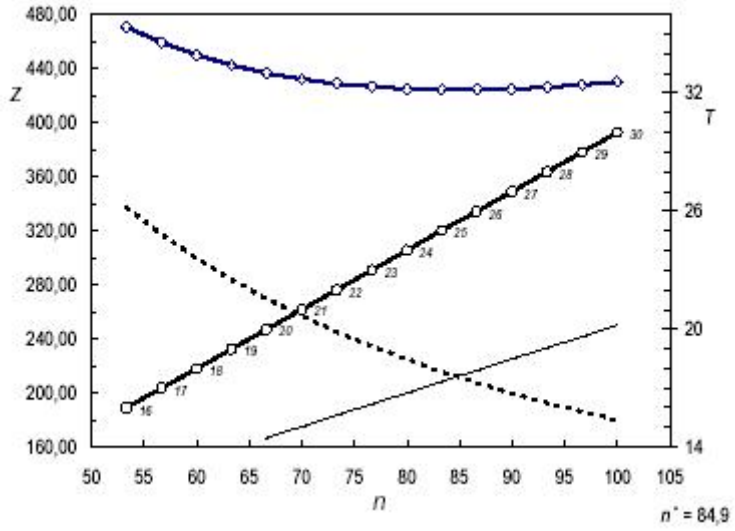

Figure 2

(2.3 ... 2.8)

2.9 Fourth case: random demand with costs of storage and shortage

Suppose that demand, for a certain time interval *T*, is random, where *p*(*r*) is the probability of a total demand *r* on the interval *T*. The demand is discontinuous, but practically it can be assumed that its rate of change is constant. The parts retain their value in the range *T*, but the cost of storage per unit time, with the interest of capital they represent, has the value c_s (cost per unit of time). It is assumed that the shortage of a part results in a loss c_p per unit of time. Consider the following example.

A factory produces cranes and has several deposits in various parts of the country. Some spare parts are very expensive, but must be made available to customer in depots since the cranes should not be unavailable too long in case of failure. Let us consider one of these parts and determine the stock to place in a depot in order to minimize the expense of the cost of storage (including income from invested amounts) and of the cost of shortage (loss of a customer, borrowing another crane, etc..).

(1) *Average Stock* corresponding to situation *"a", no-shortage:*

$$\bar{s}_a = \frac{1}{2}\left[s + (s - r)\right] = s - \frac{r}{2} \qquad \{15\}$$

(2) *Average Stock* corresponding to situation *"b", shortage:*

$$\bar{s}_b = \left[\frac{1}{2}(s + 0)\right]\frac{s}{r} = \frac{s^2}{2r} \qquad \{16\}$$

(This refers to a fraction *s/r* of the period under consideration.)

(3) *Average shortage* corresponding to the situation *"b", shortage:*

$$\bar{p}_b = \left\{\frac{1}{2}\left[0 + (r - s)\right]\right\}\left(1 - \frac{s}{r}\right) = \frac{(r - s)^2}{2r} \qquad \{17\}$$

(This refers to the remaining fraction of the period.)

The mathematical expectation of the total cost of the stock will be:

$$\begin{aligned} z(s) = c_s \sum_{r=0}^{s}\left(s - \frac{r}{2}\right)p(r) + \\ + c_s \sum_{r=s+1}^{\infty}\frac{s^2}{2r}p(r) + c_p \sum_{r=s+1}^{\infty}\frac{(r-s)^2}{2r}p(r) \end{aligned} \qquad \{18\}$$

It can be shown that the minimum of *z(s)* occurs at a value s_0 such that

$$L(s_0 - 1) < r < L(s_0) \qquad \{19\}$$

with

$$r = \frac{c_p}{c_s + c_p} = \frac{1}{1 + \frac{c_s}{c_p}} \qquad \{20\}$$

$$L(s_0) = p(r \le s_0) + \left(s_0 + \frac{1}{2}\right) \sum_{r=s_0+1}^{\infty} \frac{p(r)}{r} \qquad \{21\}$$

[Note also that $r = L(s_0)$ implies that both s_0 and $s_0 + 1$ correspond to optimum, while $r = L(s_0 - 1)$ implies optimal s_0 or $s_0 - 1$.] Of course, the determination of s_0 can be made directly numerically.

(2.10)

2.11 Numerical resolution

Let c^s = 100 \$ / month, c_p = 20 c_s = 2000 \$ / month and use the following table of the probability function $p(r)$ observed for monthly consumption, r.

r	0	1	2	3	4	5	≥ 6
$p(r)$	0,1	0,2	0,2	0,3	0,1	0,1	0

See plate http://web.ist.utl.pt/mcasquilho/compute/or/Fx-inventoryRand.php

The calculations for s = 0, 1, 2, ..., seeking a minimum value of $z(s)$ provide (in the monetary unit \$):

$$\begin{aligned} z_0 &= c_p \sum_{r=1}^{\infty} \frac{r}{2} p(r) = \\ &= 2000 \times \frac{1}{2}(1 \times 0{,}2 + 2 \times 0{,}2 + 3 \times 0{,}3 + 4 \times 0{,}1 + 5 \times 0{,}1) = 2400 \end{aligned} \qquad \{22\}$$

and successively,

$$z_1 = c_s \sum_{r=0}^{1} \left(1 - \frac{r}{2}\right) p(r) + c_s \sum_{r=2}^{\infty} \frac{p(r)}{2r} + c_p \sum_{r=2}^{\infty} \frac{(r-1)^2}{2r} p(r) =$$

$$\begin{aligned} &= 100(1 \times 0{,}1 + 0{,}5 \times 0{,}2) + \\ &+100(0{,}25 \times 0{,}2 + 0{,}167 \times 0{,}3 + 0{,}125 \times 0{,}1 + 0{,}1 \times 0{,}1) + \\ &+2000(0{,}25 \times 0{,}2 + 0{,}667 \times 0{,}3 + 1{,}125 \times 0{,}1 + 1{,}6 \times 0{,}1) = 1077{,}25 \end{aligned} \qquad \{23\}$$

Assuming monotonicity, as $z_1 < z_0$ (cost is decreasing), we must continue the calculations (and so on until it starts to increase) to detect the optimum, i.e. *minimum.*

$$z_2 = c_s \sum_{r=0}^{2} \left(2 - \frac{r}{2}\right) p(r) + c_s \sum_{r=3}^{\infty} \frac{2^2\, p(r)}{2r} + c_p \sum_{r=3}^{\infty} \frac{(r-2)^2}{2r} p(r) =$$

$$\begin{aligned} &= 100(2 \times 0{,}1 + 1{,}5 \times 0{,}2 + 1 \times 0{,}2) + \\ &+100(0{,}667 \times 0{,}3 + 0{,}5 \times 0{,}1 + 0{,}4 \times 0{,}1) + \\ &+2000(0{,}167 \times 0{,}3 + 0{,}5 \times 0{,}1 + 0{,}9 \times 0{,}1) = 479 \end{aligned} \qquad \{24\}$$

$$z_3 = c_s \sum_{r=0}^{3} \left(3 - \frac{r}{2}\right) p(r) + c_s \sum_{r=4}^{\infty} \frac{3^2\, p(r)}{2r} + c_p \sum_{r=4}^{\infty} \frac{(r-3)^2}{2r}\, p(r) =$$
$$= 100(3 \times 0{,}1 + 2{,}5 \times 0{,}2 + 2 \times 0{,}2 + 1{,}5 \times 0{,}3) +$$
$$+ 100(1{,}125 \times 0{,}1 + 0{,}9 \times 0{,}1) + 2000(0{,}125 \times 0{,}1 + 0{,}4 \times 0{,}1) = 290 \qquad \{25\}$$

$$z_4 = c_s \sum_{r=0}^{4} \left(4 - \frac{r}{2}\right) p(r) + c_s \sum_{r=5}^{\infty} \frac{4^2\, p(r)}{2r} + c_p \sum_{r=5}^{\infty} \frac{(r-4)^2}{2r}\, p(r) =$$
$$= 100(4 \times 0{,}1 + 3{,}5 \times 0{,}2 + 3 \times 0{,}2 + 2{,}5 \times 0{,}3 + 2 \times 0{,}1) +$$
$$+ 100(1{,}6 \times 0{,}1) + 2000(0{,}1 \times 0{,}1) = 301 \qquad \{26\}$$

It is now simultaneously $z_3 < z_2$ and $z_3 < z_4$, i.e., $z_2 > z_3 < z_4$, so the minimum has been found, with $z^* = 290$ \$. See Figure 3.

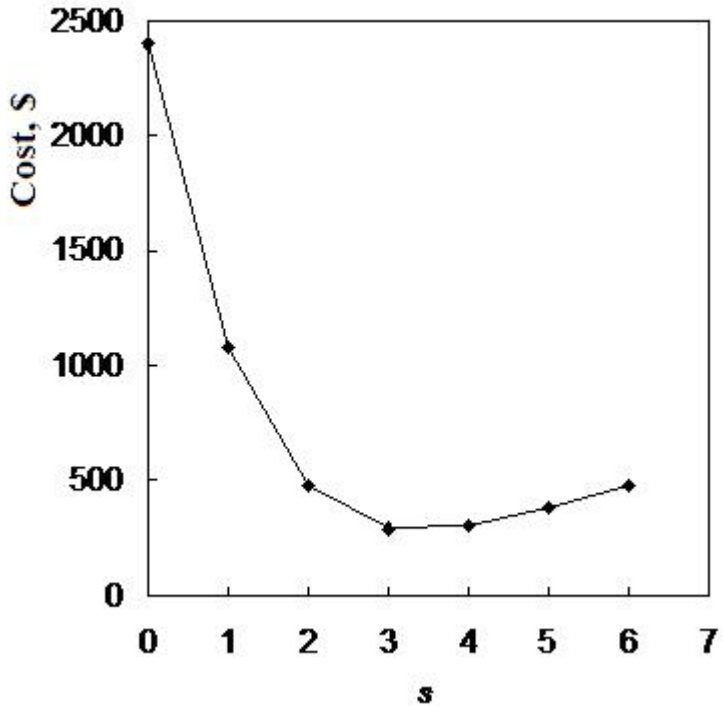

Figure 3

Appendix

$$y = ax + \frac{b}{x} \qquad \{27\}$$

$$y' = a - \frac{b}{x^2} \overset{?}{=} 0 \qquad \{28\}$$

$$a = \frac{b}{x^2} \quad \Rightarrow \quad x_0 = +\sqrt{\frac{b}{a}} \qquad \{29\}$$

In the EOQ, it is $a = \frac{1}{2} q c_S$ and $b = N c_L$, so $n_0 = 2\sqrt{\frac{1}{2}\frac{N c_L}{q c_S}} = \sqrt{2\frac{N}{q}\frac{c_L}{c_S}}$.

$$y_0 = a\sqrt{\frac{b}{a}} + b\sqrt{\frac{a}{b}} = 2\sqrt{ab} \qquad \{30\}$$

In the EOQ, $z_0 = 2\sqrt{\frac{1}{2} q c_S N c_L} = \sqrt{2Nq\, c_L c_S}$.

$$y_0'' = 2\frac{b}{x^3} = 2b\left(\sqrt{\frac{a}{b}}\right)^3 = 2ab\sqrt{a} > 0 \quad \text{(minimum)} \qquad \{31\}$$

The minimum occurs coincident with the intersection of the straight line with the hyperbola, where both contributions are $\sqrt{ab}$, i.e., $z_0/2$.

(Blank page)

Notes on *Operational Research*
March, 2008, IST

Queueing systems

MIGUEL A. S. CASQUILHO
IST, Universidade Técnica de Lisboa,
Ave. Rovisco Pais, IST; 1049-001 Lisboa, Portugal
Telephone: (+351) 21.841 7310; fax: (+351) 21.849 9242

Queueing systems are presented, with a brief introduction and formulas for usual practical cases. Some examples are solved and computer resolution is mentioned.

Keywords: *queueing systems*, *queueing theory*, *queue*, *waiting line*.

1. Fundamental and scope

The waiting phenomena, which originate the *queues*, are related to *random processes*, i.e., the models of which include random components. These are associated to probability.

The queue is almost inevitable in many situations, unless means are made available at costs possibly disproportionate to the benefits of a quick service. When circumstances impose a quick service, capable of limiting the waiting time to a reasonable level, the working conditions can be evaluated through the queueing systems theory[1].

The *queues* are frequent phenomena found in everyday life, and also in situations in economics, society, and the military. Examples: customers in a bank or post office; people waiting for a taxi or telephoning to a taxi service; cars at a (road) junction[2]; planes waiting to land or take off; broken machines waiting for repair. Several examples are given in Fig. 1. Erlang in the 1920's was one of the first to study the queueing subject applying it to the telephone system.

Arrivals	Nature of service	Servers
Customers	Sale of an article	Vendors
Ships	Unloading	Docks
Planes	Landing	Tracks
Telephone calls	Conversations	Telephone circuits
Arrival of cars	Customs control	Customs workers
Messages	Decoding	Decoders
Repair machines	Repair	Mechanics
Fires	Fire fighting	Fire brigade
Requests	Confection, repair	Repair-shop

Fig. 1 Examples of waiting phenomena.

A queue is characterized by several components: customers' population, arrival pattern, number of servers, service pattern, system capacity (size) to hold customers, and the queue discipline. Consideration of the costs of maintaining a

[1] US "waiting line"; Pt «filas de espera», «bichas»; Es «colas»; Fr «phénomènes, files d'attente»; It « fenomeni (o file) d'attesa, code»; De »Schlange(n)«.
[2] US "intersection"; Pt «cruzamento»; Fr, «carrefour».

Miguel Casquilho is Assistant Professor in the Departament of Chemical and Biological Engineering, Instituto Superior Técnico, Universidade Técnica de Lisboa (Technical University of Lisbon), Lisbon, Portugal. E-mail address: mcasquilho@ist.utl.pt.

queueing system from the supplier side and the customers' side makes it an economic optimization problem. The objective of this text is to present formulas that permit that optimization.

2. Queues structure

The structure of a queueing system is addressed based on the above mentioned parameters and characteristics. A systematization of the queueing systems by the Kendall's notation is given, as well as a nomenclature.

Customers' population

The *customers' population* may be *infinite* or *finite*. It is finite if the number of possible customers is limited and known, such as the number of machines subject to failure in a factory; infinite, otherwise.

Arrival pattern

The *arrival pattern* of customers is usually specified by the *interarrival time*, the time between successive customer arrivals to the service. It may be deterministic or a random variable with a probability distribution presumed known. [Other aspects will not be considered here, such as: arriving singly or in batches; or balking (refusal to enter) or reneging (leaving the queue because the wait is too long).]

Number of servers

The *number of servers* is the number of persons, machines, tellers, gates, etc., to attend customers. These will be considered equivalent and in parallel (other cases being series or more or less complex combinations of servers in series and in parallel).

Service pattern

The *service pattern* is usually specified by the *service time*, which may be deterministic or a random variable with probability distribution assumed known. (The service time may depend on the number of customers. The customer may be attended completely by one server or any combination of servers.)

System capacity

The *system capacity* is the maximum number of customers, both those in service and those in the queue(s). Whenever a customer arrives at a facility that is full, the customer is denied entrance to the facility and not allowed to wait outside the facility, which would increase the limited capacity, and is forced to leave. Capacity is, thus, either *infinite* or *finite*.

Queue discipline

The *queue discipline* is the order in which customers are served. This can be on a first-in, first-out (FIFO) basis (i.e., service in order of arrival, the usual one), a last-in, first-out (LIFO) basis, a random basis or a priority basis (as in hospital emergency services).

To make queue classification simpler, the so-called Kendall's notation is usually employed.

Kendall's notation

The Kendall's notation indicates ({1}): v, the arrival pattern; w, the service

$$v / w / x / y / z \qquad \{1\}$$

pattern; x, the number of servers; y, the system's capacity; and z, the queue discipline, as in Table 1. If y or z is not specified, it is taken to be ∞ or FIFO, respectively.

Table 1 Kendall's notation

	Queue characteristic	Symbol	Meaning
v, w	Interarrival time or service time	D	Deterministic
		M	Exponential
		E_k	Erlang-type (k = 1, 2 …)
		G	Any other
x	Number of servers	*Number*	∞ if not specified
y	System's capacity	*Number*	∞ if not specified
z	Queue discipline	FIFO	First in, first out
		LIFO	Last in, first out
		SIRO	Service in random order
		PRI	Priority ordering
		GD	Any other ordering

The initials are related to: D, deterministic or "degenerate" (a deterministic variable being a constant, a degenerate random variable); M, Markovian (Markovian "birth and death" process, typically with Poissonian arrivals); G, general.

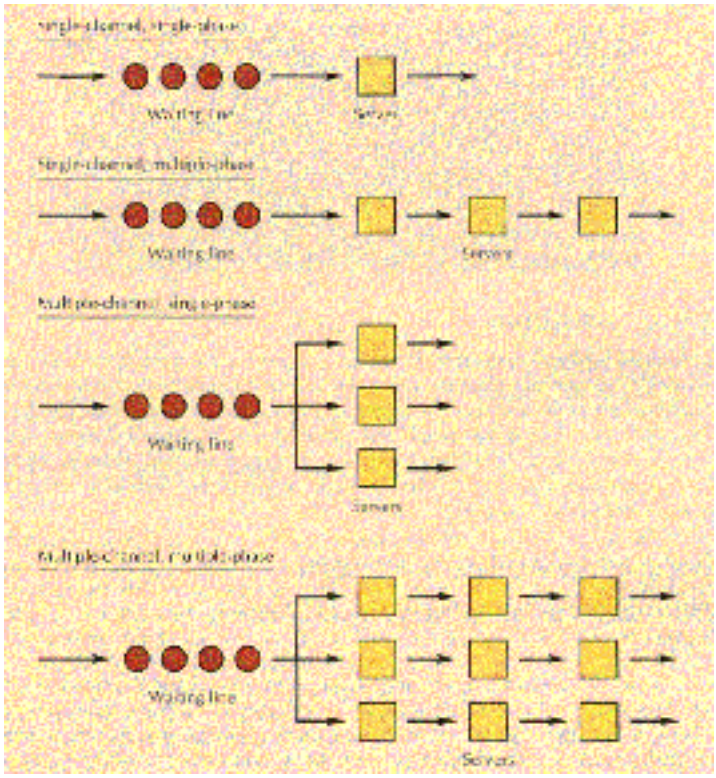

Fig. 2 Simplified queue taxonomy.

Let it be noted that in the frequent M/M/*s* case of more than one server, $s > 1$, the customers (and the selling entity) benefit from a *single* queue (which is rarely the case in large stores) [Ravindran *et al.*, 1987, 329]. This can be easily accomplished

by making available *numbered tickets* (as in post offices and usually pharmacies, in Portugal).

For a simplified taxonomy of queues, see Nemetz-Mills [2008], from whom Fig. 2 was taken. This author mentions “single or multiple channel”, i.e., single or multiple servers —here, M/M/1 and M/M/s— and “single or multiple-phase”. A multiple-phase queueing system (2.nd and 4.th rows in the figure) is a (“pure”) mixture of parallel and series servers, a complex case having a better resolution by Monte Carlo simulation.

Only M/M/1/∞/FIFO and M/M/*s*/∞/FIFO systems, i.e., for short,

$$\text{M/M/1} \qquad \text{M/M/}s \tag{2}$$

will be addressed in the following sections.

3. Single and multiple server queues

The set of a queue (or queues) and the servers constitutes the *waiting system* or simply the *system*. In the cases where it is supposed to have several queues, the *customers* place themselves either automatically in the shortest queue or according to a priority. (The term “customer” will be used instead of the more general “unit”, whether it is a person or any other entity.) These priorities make the queue discipline (hospitals, restaurants).

With the given structure of a waiting phenomenon, the notation in Table will be used:

Table 2 Notation

	Meaning
m	Number of existing customers (population size)
n	Number of customers in the system (waiting or being served)
a	Arrival rate (T^{-1}, customers / time unit)
m	Service rate (T^{-1}, services / time unit)
y	Utilization factor, or traffic intensity, $a/(sm)$
n	Number of customers in queue
j	N. of customers being served
s	N. of servers

So, it is

$$\begin{aligned} j &= n \quad \text{if} \quad n \le s \\ n + j &= n \qquad\quad n > s \end{aligned} \tag{3}$$

The values n, n and j are random. If it is

$$p_n = \Pr(n \text{ customers in the system}) \tag{4}$$

then, p_n, a probability, represents the fraction of the time the system is in state n.

3.1 Single server queues

The basic variables for a single server queue system, $s = 1$, will now be determined for the simpler and usual case of an infinite population, i.e., $m = \infty$.

The Poisson process is often used to model the situation in which a count is made on the number of events occurring in a given time, here the arrival of customers

to a service facility: $p_{\text{Poi}}(j) = [\exp(-at)](at)^j / j!$, $j = 0..\infty$ ($[a] = \text{T}^{-1}$). The time between events in a Poisson process follows an exponential[3] distribution with the same parameter a, $f(t) = \exp(-t/t)$, with mean $t = 1/a$ ($[t] = \text{T}$). The parameter t is the expected time between events.

To find p_n, consider its evolution during an instant, from time t to $t + \text{d}t$, with dt small enough so that no two (or more) events can occur.

$$p_0(t+\text{d}t) = p_0(t)\underset{\text{no change}}{(1-a)}\text{d}t + p_1(t)\underset{\text{departure}}{(m)}\,\text{d}t \qquad \{5a\}$$

$$p_n(t+\text{d}t) = p_{n-1}(t)\underset{\text{arrival}}{(a)}\text{d}t + p_{n+1}(t)\underset{\text{departure}}{(m)}\,\text{d}t + p_n(t)\underset{\text{no change}}{(1-a-m)}\text{d}t \qquad \{5b\}$$

This becomes

$$\frac{p_0(t+\text{d}t) - p_0(t)}{\text{d}t} = -ap_0(t) + mp_1(t) \qquad \{6a\}$$

$$\frac{p_n(t+\text{d}t) - p_n(t)}{\text{d}t} = ap_{n-1}(t) + mp_{n+1}(t) - (a+m)p_n(t) \qquad \{6b\}$$

Introducing the *utilization factor* [H&L, 2005, 770] or *traffic intensity* [Ravindran *et al.*, 1987, 320]

$$y = \frac{a}{Sm} \qquad \{7\}$$

which is here simply $y = \frac{a}{m}$, and in the limit as dt goes to zero, it is

$$\frac{1}{m}p_0'(t) = -yp_0(t) + p_1(t) \qquad \{8a\}$$

$$\frac{1}{m}p_n'(t) = yp_{n-1}(t) + p_{n+1}(t) - (1+y)p_n(t) \qquad \{8b\}$$

The system will be studied only in the steady state (null derivatives), so it is

$$-yp_0 + p_1 = 0 \qquad \{9a\}$$

$$yp_{n-1} + p_{n+1} - (1+y)p_n = 0 \qquad \{9b\}$$

or [from $p_2 = (1+y)p_1 - yp_0$, $p_3 = (1+y)p_2 - yp_1$, etc.]

$$p_1 = yp_0 \qquad \{10a\}$$

$$\begin{aligned} p_2 &= (1+y)(yp_0) - yp_0 = (y^2 + y - y)p_0 = y^2 p_0 \\ p_3 &= (1+y)(yp_2) - yp_1 = (y^3 + y^2 - y^2)p_0 = y^3 p_0 \end{aligned} \qquad \{10b\}$$

etc.

So, in general, it is

[3] Also called “negative exponential” [Ravindran *et al.*, 1987, 293].

$$p_n = p_0 y^n \qquad \{11\}$$

The population size is $m = \infty$. As the probabilities must, of course, add to one, and recognizing the sum of a geometric series (it is $y < 1$), it is

$$1 = \sum_{n=0}^{\infty} p_n = p_0 \sum_{n=0}^{\infty} y^n = \frac{p_0}{1-y} \qquad \{12\}$$

Thus, it is

$$p_0 = 1 - y \qquad \{13\}$$

and generally

$$p_n = (1-y) y^n \qquad \{14\}$$

[The *geometric distribution* can be recognized in Eq. {14}: $p(n) = r(1-r)^n$, $n = 0..\infty$, with parameter $r = 1 - y$, mean $m = (1-r)/r$, i.e., $y/(1-y)$.]

The probability p_0 is the fraction of time the system is ***idle*** (empty), and the parameter y can be taken as the fraction of time the server is ***busy*** [Ravindran *et al.*, 1987, 320].

The *mean* or *expected value* of the **number of customers in the system** is, by the definition of mean,

$$\begin{aligned} \bar{n} &= \sum_{n=0}^{\infty} n p_n = (1-y) \sum_{n=0}^{\infty} n y^n = (1-y) \sum_{n=1}^{\infty} n y^n = (1-y) y \sum_{n=1}^{\infty} n y^{n-1} = \\ &= (1-y) y \frac{\mathrm{d}}{\mathrm{d}y} \sum_{n=1}^{\infty} y^n = (1-y) y \frac{\mathrm{d}}{\mathrm{d}y} (1-y)^{-1} = (1-y) y (1-y)^{-2} \end{aligned} \qquad \{15\}$$

or

$$L \equiv \bar{n} = \frac{y}{1-y} \qquad \{16\}$$

The mean **number of customers in the queue,** or mean queue length (with a queue of zero if there are 0 or 1 customers in the system), is

$$\begin{aligned} L_q \equiv \bar{n}^- &= \sum_{n=2}^{\infty} (n-1) p_n = (1-y) \sum_{n=2}^{\infty} (n-1) y^n = \\ &= (1-y) y^2 \sum_{n=2}^{\infty} (n-1) y^{n-2} = (1-y) y^2 \frac{\mathrm{d}}{\mathrm{d}y} \sum_{n=2}^{\infty} y^{n-1} = \frac{y^2}{1-y} \end{aligned} \qquad \{17\}$$

The difference between L (customers in the system) and L_q (customers in the queue) should be, and is, the mean number of busy servers, y:

$$L - L_q = \frac{y}{1-y} - \frac{y^2}{1-y} = \frac{y(1-y)}{1-y} = y \qquad \{18\}$$

An equation known as **Little's formula** (cited in most queueing literature) relates L to W, the mean waiting time in system:

$$L = (1 - p_N) \, a \, W \qquad \{19\}$$

When it is $N = \infty$, as in the cases presented, the formula reduces to $L = a\,W$ (as the probability p_N obviously tends to zero). This permits easily finding the **mean time in the queue**, W, and **mean time in the system**, W_q. The formulas for the M/M/1 case are shown in Table 3. As a and m are rates (times per unit time), the expressions with $1/a$ or $1/m$ represent, indeed, time.

Table 3 Synopsis for M/M/1

Variable and formula	
Probability of 0 customers in the system	
$p_0 = 1 - y \qquad \text{with} \quad y = \dfrac{a}{m} < 1$	(a)
Probability of n customers in the system	
$p_n = (1-y)\,y^{\,n}$ $P_n = \sum_{j=0}^{n} p_j = 1 - y^{\,n+1}$	(b)
Mean of no. of customers in the queue (waiting)	
$L_q = \dfrac{y^{\,2}}{1-y}$	(c)
Mean of no. of customers in the system	
$L = \dfrac{y}{1-y} = L_q + y$	(d)
Mean of time in the queue (a customer waiting)	
$W_q = \dfrac{1}{a}\dfrac{y^{\,2}}{1-y} = \dfrac{y}{m-a}$	(e)
Mean of time in the system (a customer spending)	
$W = \dfrac{L}{a} = \dfrac{1}{a}\dfrac{y}{1-y} = \dfrac{1}{m-a} = W_q + \dfrac{1}{m}$	(f)

The probabilities of waiting at least t (with $t \geq 0$) are given [H&L, 1995, 681] by

$$\begin{aligned} \Pr(\text{wait} > t) &= \exp[-m(1-y)t] \\ \Pr(\text{wait}_q > t) &= y\ \Pr[\text{wait} > t] \end{aligned} \qquad \{20\}$$

[the first expression an exponential distribution with parameter $m(1-y)$] which lead to (and confirm) $W = 1/(m-a)$ and $W_q = y/(m-a)$.

3.2 Multiple server queues

For this case, similar but more laborious derivations can be made. The results only are presented in Table 4. A *single* queue for customers waiting and steady state are also supposed.

In the particular case of $s = \infty$, it is

$$p_0^{-1} = \frac{1}{1-y}\lim_{s\to\infty}\frac{(sy)^s}{s!} + \exp(sy) = \exp(sy) = \exp\left(\frac{a}{m}\right) \qquad \{21\}$$

Eq. {21} comes from the fact that *(i)* the sum (from 0 to $s-1$) can be recognized as the Taylor series development of the exponential function and *(ii)* the other term goes to zero. So, p_0 becomes a constant:

$$p_0 = \exp(-a/m) \qquad \{22\}$$

The remaining variables will have the following values:

$$p_n = p_0 \frac{r^n}{n!}$$
$$L_q = 0 \qquad L = L_q + r = r = \frac{a}{m}; \qquad W_q = 0 \qquad W = W_q + \frac{1}{m} = \frac{1}{m} \qquad \{23\}$$

Indeed, p_0 is not 1 (a value that might be intuitive), as there are customers arriving; L_q is zero (zero customers waiting), but L is not zero, as they are being served (spending useful time); and W_q is zero (no wait in queue), but W is the inevitable service time, $1/m$ (not zero). This may be the case of a self-service situation if there are "many" servers, enough for all the arriving customers.

Table 4 Synopsis for M/M/*s*

Variable and formula	
Probability of 0 customers in the system	
$p_0^{-1} = \frac{(sy)^s}{s!(1-y)} + \sum_{n=0}^{s-1} \frac{(sy)^n}{n!} \qquad \text{with} \quad y = \frac{a}{sm} < 1$	(a)
Remark: p_0^{-1}, not p_0	
Probability of *n* customers in the system	
$p_n = \begin{cases} \frac{(sy)^n}{n!} p_0 & 0 \le n \le s \\ \frac{s^s}{s!} y^n p_0 & n \ge s \end{cases}$	(b)
$P_n = P_{s-1} + \sum_{j=n}^{\infty} \frac{s^s}{s!} y^j p_0 = P_{s-1} + p_0 \frac{s^s}{s!(1-y)} y^n \qquad n \ge s$	
Mean of no. of customers in the queue (waiting)	
$L_q = p_0 \frac{s^s y^{s+1}}{s!(1-y)^2}$	(c)
Mean of no. of customers in the system	
$L = aW = L_q + sy = a\left(W_q + s\frac{1}{m}\right)$	(d)
Mean of time in the queue (a customer waiting)	
$W_q = \frac{L_q}{a}$	(e)
Mean of time in the system (a customer spending)	
$W = W_q + \frac{1}{m}$	(f)

The probabilities of waiting at least t (with $t \geq 0$) are given [H&L, 1995, 684] by

$$\Pr(\text{wait} > t) = \mathrm{e}^{-mt}\left\{1 + p_0 \frac{(sy)^s}{s!(1-y)} \frac{1-\exp[-mt(s-1-r)]}{s-1-r}\right\} \qquad \{24\}$$
$$\Pr(\text{wait}_q > t) = (1-P_{s-1})\exp[-sm(1-r)t]$$

which leads to (and confirms) $W = 1/(m-a)$.

4. Illustrative examples

Suppose $a = 10$ hr^{-1} and $m = 15$ hr^{-1} (data from Baker's [2006, 2] pharmacy example). For $s = 1$, $s = 2$ (in the reference), and $s = 100$, the results are given in Table 5. (In the reference, r is used for y.) In this case, with $a/m = 0.667$, they show, namely, little difference from 2 to 100 servers.

Table 5 Results for growing s (other data constant)

	$s = 1$	$s = 2$	$s = 100$
y (or r)	0.667	0.333	0.007
p_0	0.333	0.500	0.513
L_q	1.333	0.083	0.000
L	2.000	0.750	0.667
W_q	0.133	0.008	0.000
Service time, $1/m$	0.067	0.067	0.067
W	0.200	0.075	0.067

Various examples can be run on the author's Internet page [Casquilho, 2008]. Also, an economic optimization of s can be made there.

5. Conclusions

The theory applicable to queueing systems —provided that the underlying conditions are met, namely, steady state— can lead to useful results, permitting significant control on the behaviour of such systems. The calculations are cumbersome, adequate to computer treatment.

Acknowledgements

This study was made within the author's teaching and research activities in the "Centro de Processos Químicos" (Centre for Chemical Processes), Department of Chemical and Biological Engineering, Instituto Superior Técnico, Universidade Técnica de Lisboa (Technical University of Lisbon). The computing and Internet publishing was made at the "Centro de Informática do IST" (IST Informatics Centre).

References

– BAKER, Samuel L., 2006, Internet page: ("%20" means blank) (visited Feb. 2008) *http://hspm.sph.sc.edu/Courses/J716/pdf/716-8%20Queuing%20Theory%20II.pdf*.
– CASQUILHO, Miguel, 2008, Internet page: *http://web.ist.utl.pt/mcasquilho/compute/or/Fx-queues.php*.

– (H&L) HILLER, Frederick S., Gerald J, LIEBERMAN, 2005 (2001, 1995, 1990, 1986, 1980, 1974, 1967), "Introduction to Operations Research", 8.th ed., McGraw-Hill, New York, NY (USA), (xxv+1062 pp), ISBN 007-123828-X.
– NEMETZ-MILLS, Patricia, Internet page: *http://www.cbpa.ewu.edu/~pnemetzmills* [Feb, 2008]
– RAVINDRAN, A., Don T. PHILLIPS, James J. SOLBERG, 1987, "Operations Research: principles and practice", 2.nd ed., John Wiley & Sons, New York, NY (USA), (xviii+637 pp), ISBN 0-471-85980-X.

❖