On the distribution of lead time delays in supply chains

by

Connie K. Gudum
Copenhagen Business School
connie@cbs.dk

Abstract

The overall aim of this paper is to provide insight into distributions of lead times and lead time delays for use in constructing models and in testing validity of theoretical models. A case study is considered where empirical data for one focal company has been obtained to estimate distributions of lead time delays. The empirical analysis contains ABC classification analysis of customers, estimation of delay distributions on customer level and analysis of other effects on lead time from factors like order size, autocorrelation and seasonality. Then an exact expression of the delay distribution of the delay in a two-level supply chain is presented and analysed. Moreover, a simulation study of a N-level serial distribution system has been carried out. In the simulation study there is one member at each level and all members replenish according to the continuous review (s, Q) ordering policy. All levels face (compound) poisson demand from external customers and replenishment orders from the echelon downstream. There is a fixed service time between each echelon and additional delays occur when the upstream echelon is out of stock. Both the case study and the theoretical studies show that lead time delays in general can be approximated with the gamma, beta, Weibull and exponential distributions.

Keywords
Inventory control, lead time, lead time delay, case study, simulation, supply chain

1 Introduction

In most inventory and production planning problems there is an interval of time between the decision to place an order for more stock and the availability of the stock from that order to meet customer demand or a production setup. This time interval is called the lead time. Lead time can roughly be divided into time phases defined by the following time points:

- At time point 1 the customer sends the order request.
- At time point 2 the supplier receives the order request.
- At time point 3 the supplier sends the physical goods to satisfy the order.
- At time point 4 the customer receives the goods.
- At time point 5 the goods are placed on shelves ready for use or sale.

This ordering process happens both at the supply side and the customer side of a company, where the supply side is representing the input to the focal company, and the customer side represents the output of the company. When analysing empirical data, it is important to identify which time points are known in the data available. This highly depends on whether the input or the output side is analysed. Let i and o denote the input side and output side respectively, then the phases are illustrated in Figure 1.

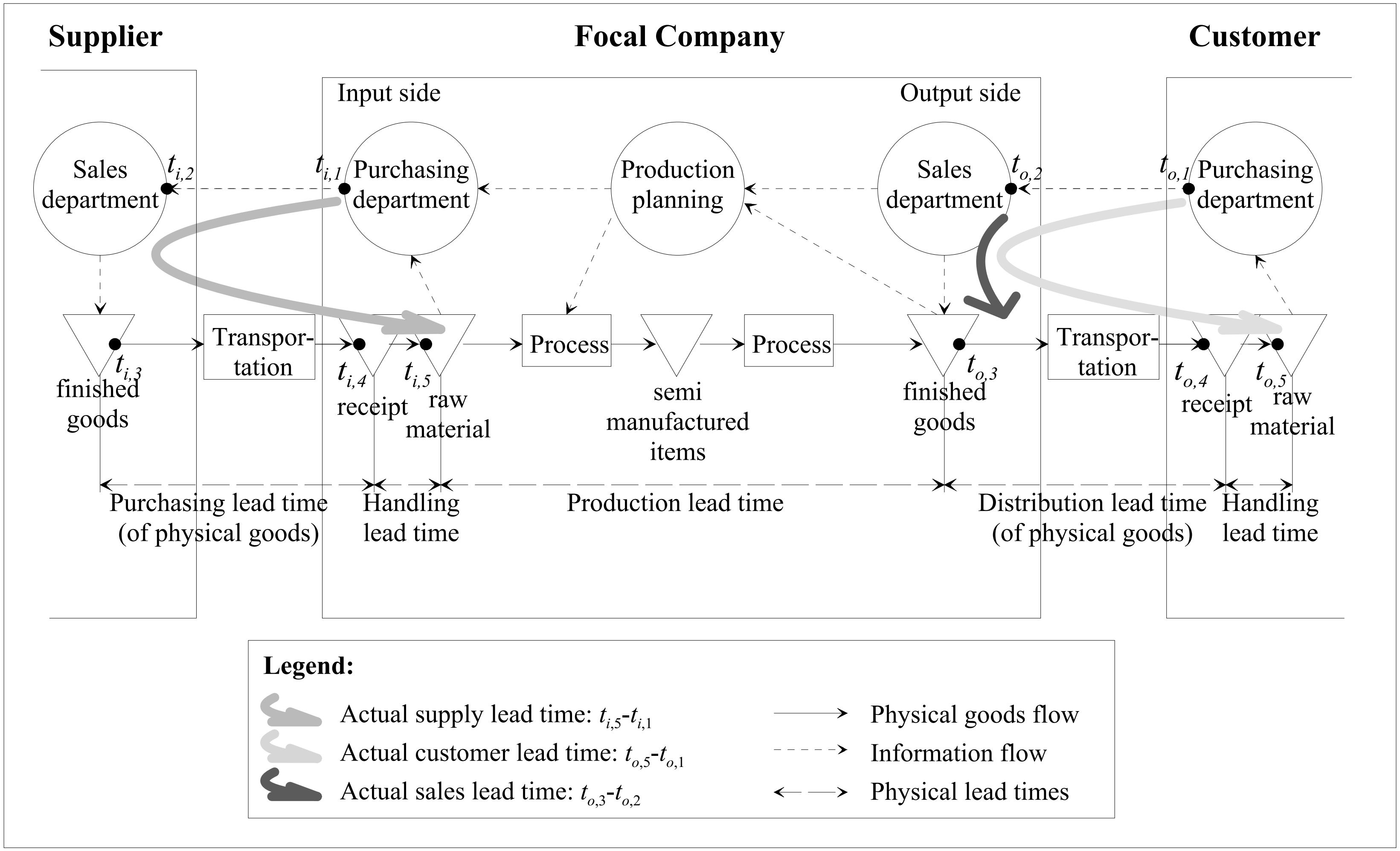


Figure 1: Phases of lead time

Naturally each of these time phases can be divided into even more phases. Prior to time point 1 the customer has to realise the need for replenishment, and a supplier has to be chosen. Between time points 1 and 2 the order is transferred to the supplier. Depending on the technology or media used this can take only seconds, e.g. by Electronic Data Interchange, or it can take days, e.g. send by post. Between time points 2 and 3 the supplier processes the order request and this phase could consist of production (if Make To Order) or picking from stock (if Make To Stock). Moreover, the supplier may be in a situation where he is out of stock, and therefore he has to refill his inventory by an order to his supplier or production unit. In this case the lead time will have an additional delay due to the stockout situation at the supplier. Also, this phase normally consists of control, packaging and loading. Between time points 3 and 4 the physical goods are being transported from the supplier to the customer. Finally, between time points 4 and 5 the customer handles the goods received, which could be actions like unloading, unpacking, control, packaging into new wrapping or smaller units. At time point 5 the goods have been placed on stock and are ready for use or sale. As Karmarkar [6] notes, lead time can also be an internal measure of the manufacturing lead time from the time a production setup is initiated to the time it is completed and material is ready for use to fill demand by either a customer or the next stage in the production process.

Inventory theory normally deals with lead time defined as above as the time between the order is placed and the physical goods are ready for use or sale. To understand the mechanisms behind the lead times, it is necessary to divide the lead time into the phases described above. Whether the phases should be divided even more depends on the data and information available and of the problem under study. By dividing lead time into phases it can be identified which phase is causing the longest part of the lead time and also which phases are responsible for very fluctuating lead times. The goal is then to determine the mean and the variance of the lead time and also to measure if there is a certain skewness. Hence, by analysing empirical lead time data it is possible to estimate the lead time distribution and the source of variation.

Other problems of interest regarding analysis of lead time are whether the lead time depends on factors like the order size, choice of supplier, type of good, type of order policy, number of suppliers in the supply chain or previous lead times. If such dependencies exist they have to be accounted for in the choice of inventory control policy.

Traditionally, lead times are either assumed fixed or independent identically distributed (i.i.d.) random variables. In the first case a delay can not occur, since it is indirectly assumed that the supplier has infinite capacity and therefore never runs out of stock. If lead times are allowed to vary the common assumption is that lead times are i.i.d. A common procedure is then to estimate mean and standard deviation of the demand distribution (assuming stationary demand) and of the lead time distribution based on historical data. Then the parameters are combined to get the parameters of the compound distribution of demand during lead time. In standard textbooks on inventory theory (see e.g. Silver, Pyke and Peterson [7] or Tersine [8]) control policies are based on the assumption that the compound distribution of demand during lead time is (approximately) normal. In their paper, Bagchi, Hayya and Chu [1] provide an extensive discussion with examples showing that this assumption is seldom reasonable and can lead to high cost penalties and incorrect customer service levels. Therefore, they provide a review of literature that suggest models based on other lead time demand distributions. The aspect of non-normal lead time demand distributions will be presented and discussed further in Section **??**.

Another question is whether it is reasonable to assume independent lead times. It often happens that lead times are longer than expected, and if this is caused by a stockout it is likely that this stockout will last for a while and thereby also cause a delay in the succeeding orders. In this case lead times of consecutive orders will be strongly autocorrelated rather than independent. The next question then is how such a correlation can be accounted for in the inventory control models.

A different approach to modelling lead time is to assume that the lead time normally takes a fixed value, but occasionally the order may be delayed. A lead time delay is therefore defined as the number of time units an order is delayed compared to the expected or promised lead time. As for the length of the lead time itself, also the delay distribution and the source of the delay is of interest. One could imagine that delays usually are caused by a stockout situation at the supplier or a manufacturing breakdown. However, also defect goods, quality disputes, missing components or problems with transportation can cause additional delays.

Hill [4] proposes a number of research questions in his paper among which is the idea of modelling the lead time with an additional delay. This has highly motivated this current research. As Hill also points out there is a need for empirical studies on observed lead time data. Therefore, in Section 2 a case study based on empirical lead time data from one focal company is considered in order to analyse empirical delay distributions. Since delays are most likely to occur due to stockouts at higher levels of suppliers the delay is related to the duration of such a stockout. Theoretical analysis has therefore been carried out to examine the impact of stockouts on delays at lower levels. In Section 3 a simple exact two-level model is presented. For a more complex N-level serial distribution system a simulation experiment has been carried out, which is presented in Section 4. In Section 5 we discuss the findings and provide some concluding remarks.

2 A case study of empirical lead time delays

In this section a case study based on empirical lead time data of one focal company is considered. The case study consists of a bi-criteria ABC classification analysis of customers. This provides the basis for the estimation of delay distributions at customer level. Also, the effect on lead time of factors like order size, autocorrelation and seasonal variation is analysed.

2.1 Data and key figures

From a middle sized Danish wholesaling company in the industry of veterinarian and agricultural products, data on all transactions regarding customer orders delivered within year 1999 are available. Hence, data represent the output side of the company according to Figure 1. The company wishes confidentiality so in the remainder of the text it will be referred to as DAN. The total annual sales volume (SV) in 1999 was 13,844,538 units. The total number of available transactions, i.e. order lines, for 1999 was 675,000, which are distributed on almost 90,000 orders. Data are analysed on order level and customer level, which reduces the number of observations in each analysis. The available data for each transaction are

- Article identification number
- Order identification number
- Customer identification number
- Order size
- Date of receiving the order request (time point 2)
- Promised date of delivery (time point 3)
- Actual date of delivery (time point 3)

Here, time points correspond to those presented in Figure 1. Date of delivery therefore corresponds to the date that the order is physically send from the company to the customer. The customer therefore has to add the transportation time to this date of delivery to get the date of receipt. The transportation time depends on the geographical location of the customer. Since ordering and delivery dates are available only for time points 2 and 3, only one phase is represented. Therefore, it is not possible to analyse the source of variation. Moreover, the delay is here defined as the difference between promised and actual date of delivery.

The number of articles for sale is 6,620. 1% of the articles account for 55% of SV, and 40% of the articles account for 99% of SV. The number of customers is 2,091. 1% of the customers account for 50% of SV, and 40% account for 99% of SV. Almost 90,000 orders are received per year each with 1-200 order lines, the average number of order lines being 7. 97% of the orders are promised immediate delivery ; however, only 59% of the orders are actually delivered immediately, i.e. the same day as the order was received. 89% of the orders are delivered within 4 days after the promised date of delivery. 1% are more than 38 days late, i.e. 38 days later than the promised delivery date. The empirical distribution of the delay of all orders is shown in Figure 2. The empirical distribution only include strictly positive delay data. Hence, the atom in delay=0 is excluded from the distribution analysis due to the idea of modelling the lead time as a probability of a fixed lead time and a probability of a delay. It is the probability distribution of the positive delays that is of interest.

The mean delay conditional on the delay being positive is 5.1 days, and the exponential distribution with the same mean is depicted in the figure with a dashed line. From this figure it appears that the empirical distribution of the delay is not approximated well by an exponential distribution. The empirical distribution may, however, vary from one customer to another, and therefore the delay distribution analysis will be carried out at customer level. Moreover, if the customers need information about lead times for use in their inventory planning, this is the level of data needed.

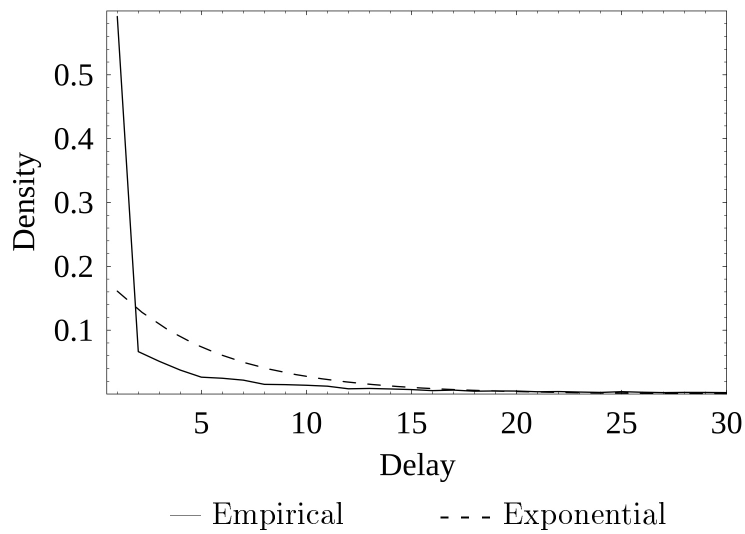


Figure 2: Empirical distribution of delays of all orders

2.2 Bi-criteria classification of customers by size and loyalty

The lead time delay distribution may depend on the size of the customer. Since the large customers are vital for the company's survival, it is important that they experience a high service level and only few delays. This suggests that customers should have different attention in terms of service and lead times depending on the size of the customer. Therefore, it is important to classify each of the customers in a so-called ABC classification analysis. This will be done according to two criteria. The first criterion is customer size measured by the purchase volume as a percentage of the sales volume. When analysing lead times and lead time delays quantitatively, a certain number of observations is needed. If a customer for example received less than ten replenishment orders during 1999, it is very difficult to conclude anything about the empirical distribution of the lead times to that customer. Therefore, the customers also need to be classified according to number of orders per year. This criterion of classification is called loyalty and is measured by number of orders received per customer during 1999. The two criteria will then be combined in a bi-criteria classification analysis of customers using the simple approach suggested by Flores and Whybark [2]. They argued that if more than one dimension of a classification needs attention, then competing criteria exist and managers must decide on how to consider all of them. One method of doing this within the ABC framework is multi-criteria ABC analysis, which employs a joint criteria matrix to reclassify the categories by weighting numerical combinations of the criteria under study. Based on the bi-criteria classification analysis, customers with a very low number of orders will therefore be omitted from the analysis of lead time delays at customer level.

For each customer the annual quantity purchased, summed over all products, is calculated and called the Purchase Volume (PV), and the percentage of total annual sales volume in units of the entire company is determined (denoted % SV). In this analysis the annual sales volume in *units* is used as the size criterion, since information about the unit costs and sales prices are not available. All customers are ranked in descending order, starting with the largest value. Then the corresponding values of the cumulative percentage of total sales volume and the cumulative percentage of the total number of customers are plotted in a graph, which can be seen in Figure 3. In a similar way, customers are ranked according to the number of orders they have received. Then the cumulative percentage of total number of orders and the cumulative percentage of the total number of customers are plotted, which can be seen in Figure 4.

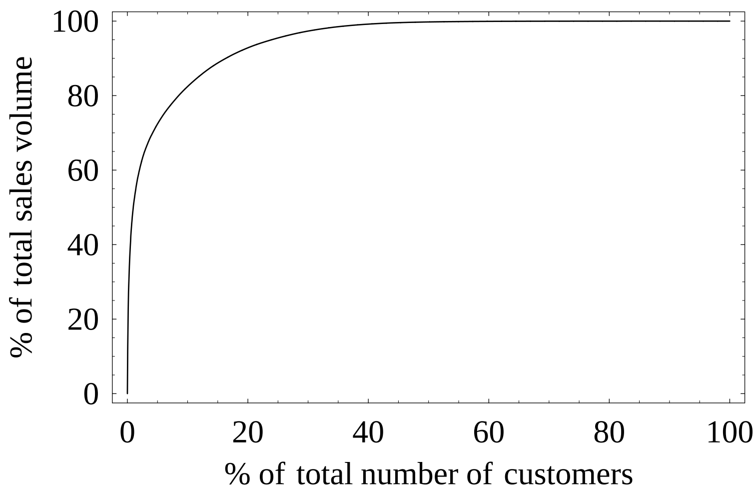


Figure 3: Classification of customers by sales volume

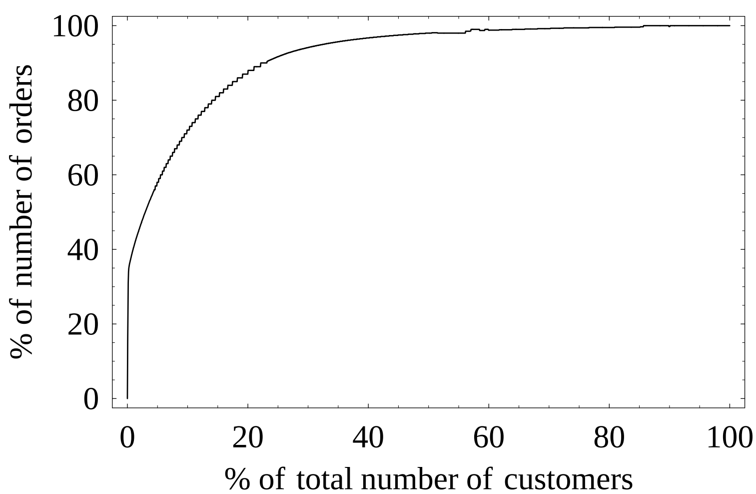


Figure 4: Classification of customers by loyalty

Based on the data in Figure 3, customers have been classified into 6 classes. 1% of the customers account for 50% of the total sales volume. These customers are classified as very large customers (class AA and A), and they correspond to 20 customers of 2091 in total. Moreover, 6 of these, corresponding to 0.33% of the customers, are classified in a separate group called extremely large customers (class AA), because they alone account for 34% of the total sales volume. Based on the data in Figure 4, it can be seen that

1% of the customers account for 40% of the orders (class AA and A), and 40% of the customers account only for 1% of the orders (class E). The loyalty criterion can also be classified into 6 classes. The individual classes of each of the two criteria are described in the Tables 1 and 2.

Class	Customer size	PV	# of cust.	% of cust.	% SV	Cum. % cust.	Cum. % SV
AA	Extremely large	$\geq$ 275,000	7	0.33	34	0.33	34
A	Very large	100,000-274,999	13	0.67	16	1	50
B	Large	10,000-99,999	189	9	32	10	82
C	Medium sized	4,000-9,999	227	10	11	20	93
D	Small	700-3,999	394	20	6	40	99
E	Very small	0-699	1261	60	1	100	100

Table 1: Description of classes in the sales volume criterion

Class	Loyalty	# of orders	# of cust.	% of cust.	% of orders	Cum. % cust.	Cum. % of orders
AA	Extremely high	$\geq$ 1,000	4	0.19	34	0.19	34
A	Very high	250-999	18	0.86	7	1.05	40
B	High	52-249	355	17	43	18	85
C	Medium	12-51	333	16	10	34	95
D	Low	3-11	545	26	4	60	99
E	Very low	0-2	836	40	1	100	100

Table 2: Description of classes in the loyalty criterion

As can be seen from Table 2, the customers in loyalty class D and E are so rare (placing less than 12 orders per year) that it makes no sense to analyse these in terms of lead times. Based on the two criteria under study, a contingency table has been made, which is shown in Table 3.

	Customer size (% of Sales Volume)					
Loyalty	AA	A	B	C	D	E
AA	4	0	0	0	0	0
A	1	1	15	1	0	0
B	1	3	83	140	127	1
C	1	2	54	36	143	97
D	0	6	29	37	86	387
E	0	1	8	13	38	776

Table 3: Bi-criteria contingency table

In general, it can be seen that the lower the % of the SV is the lower the loyalty is. However, a few of the large customers (in terms of % SV) actually order very seldom, where the purchase quantities are pooled to a few extremely large orders. From this contingency table a bi-criteria classification scheme is defined, and 7 new classes of customers are constructed. The new classes are indicated by the groupings in Table 3, and a description of the new classes are given below in Table 4.

Class	Bi-criteria combination	Orders		PV	# of cust.	% of cust.
AA	Loy=AA $\wedge$ SV=AA	$\geq$1000	$\wedge$	$\geq$ 275,000	4	0.2
A	Loy=A,B,C $\wedge$ SV=AA,A	12-999	$\wedge$	$\geq$ 100,000	9	0.4
B	Loy=A,B,C $\wedge$ SV=B	12-999	$\wedge$	10,000-99,999	152	7
C	Loy=A,B,C $\wedge$ SV=C	12-999	$\wedge$	4000-9999	177	9
D	Loy=A,B,C $\wedge$ SV=D	12-999	$\wedge$	700-3999	270	13
E	Loy=A,B,C $\wedge$ SV=E	12-999	$\wedge$	0-699	98	5
F	Loy=D,E $\wedge$ SV=AA,A,B,C,D,E	$\leq$ 11	$\wedge$	any	1381	66

Table 4: Description of bi-criteria classes

Except for the bi-criteria class F, which is omitted from further analysis, a number of customers are picked from each class for further analyses.

2.3 Distribution analysis of the delay at customer level

For the analysis of delay distributions at customer level 23 customers have been selected based on the bi-criteria classification analysis above. All 4 customers in class AA and all 9 customers in class A are included in the analysis, since these are the most important customers. Four customers have been chosen arbitrarily from class B, and from each of the classes C, D and E two customers have been chosen.

Table 5 contains the key figures of each of the 23 chosen customers: Number of units purchased (Purchase Volume), % of total annual SV, number of orders, % of total number of orders and finally the % of orders that were promised delivered immediately (PID) for each of the chosen customers. Table 5 also summarises the delay densities in % for delays equal to 0, 1, 2, 3, and 4. The zero delay density is denoted $P(t = 0)$ and corresponds to the probability of no delay, which is the same as the ready rate service measure. As can be seen from the values of $P(t = 0)$ the service level is generally very low, even for the 4 biggest customers.

Customer AA1 belongs to class AA and is the biggest customer of DAN accounting for 11% of the sales volume and 15% of the total number of orders. It is intriguing to analyse how this customer experiences lead time and delays on products bought from DAN. The date range of actual delivery dates is 4th of January 1999 to 30th of December 1999; hence, deliveries cover one full year of data. The number of orders is 13,036. For all orders placed by customer AA1, DAN has promised immediate delivery. In this case actual lead time is therefore equal to the delay time. However, only 45% of the orders are actually delivered without any delay. 88% of the orders are delivered within 4 days delay time,

Cus-tomer	PV	% of SV	# of orders	% of orders	PID	Delay densities in %				
						$P(t=0)$	$P(t=1)$	$P(t=2)$	$P(t=3)$	$P(t=4)$
AA1	1,565,768	11	13036	15	100%	45	36	4	2	2
AA2	909,500	7	3124	4	100%	11	23	16	10	5
AA3	327,848	2.3	7782	9	100%	36	52	2	2	1
AA4	278,804	2	6259	7	100%	74	13	2	2	2
A1	683,798	5	45	0.05	56%	56	4	0	7	0
A2	631,180	4.6	276	0.31	68%	39	7	3	1	1
A3	335,706	2.4	159	0.18	60%	52	1	3	8	0
A4	273,805	2	449	0.51	96%	48	29	4	2	1
A5	239,827	1.7	108	0.12	64%	48	7	1	1	0
A6	218,568	1.6	199	0.22	69%	58	3	1	1	2
A7	201,156	1.5	51	0.06	43%	51	2	8	0	8
A8	160,068	1.1	21	0.02	62%	76	24	0	0	0
A9	148,744	1.1	14	0.02	100%	100	0	0	0	0
B1	97,206	0.7	69	0.08	59%	65	3	0	3	1
B2	48,705	0.35	77	0.09	74%	49	6	0	0	1
B3	28,916	0.21	214	0.24	100%	69	12	1	2	3
B4	13,568	0.1	189	0.22	100%	70	14	3	3	2
C1	9702	0.07	114	0.13	100%	69	17	4	3	2
C2	5476	0.04	97	0.11	100%	57	29	3	1	1
D1	2830	0.02	103	0.12	100%	54	37	2	1	1
D2	1374	0.01	88	0.10	100%	49	36	0	3	1
E1	673	< 0.01	40	0.05	100%	90	3	0	0	0
E2	150	< 0.01	27	0.03	100%	22	63	0	0	4

Table 5: Empirical distributions of chosen customers

and 2% of the orders are delivered with more than 25 days delay.

An extensive exploratory distribution analysis has been conducted to see which distributions are reasonable to assume for lead time delays. A number of statistical distributions has been fitted to the empirical data of each customer and tested by Pearson's chi-square test (at a 0.98 significance level). Also the Kolmogorov-Smirnov (KS) test was used. When the number of observations gets too high such tests are seldom accepted. Therefore, for the AA customers the distributions are not accepted. Based on the squared error (S.E.) the relative precision of the distributions are also compared.

An example of the results of the distribution analysis are presented in Table 6 for customer A2. Moreover, the empirical and exponential density functions for customer A2 delays are depicted in Figure 5. As can be seen from the table, the delays to customer A2 can be fitted with the beta or the exponential distribution.

The results of the distribution analysis for all the chosen customers are summarised in Table 7 below, where $E(t)$ denotes the mean positive delay, and σ_t denotes the standard

Distribution	S.E.	χ^2	$P(\chi^2)$	KS	$P(\text{KS})$
Beta	0.0114	12	0.007	0.0994	0.0728
Exponential	0.0121	16	<0.005	0.0652	>0.15
Gamma	0.0148	34	<0.005	0.309	<0.01
Weibull	0.0290	31	<0.005	0.189	<0.01
Lognormal	0.0354	135	<0.005	0.377	<0.01
Normal	0.0774	100	<0.005	0.299	<0.01
Triangular	0.0864	138	<0.005	0.315	<0.01
Uniform	0.1400	281	<0.005	0.469	<0.01

Table 6: Distribution analysis - customer A2

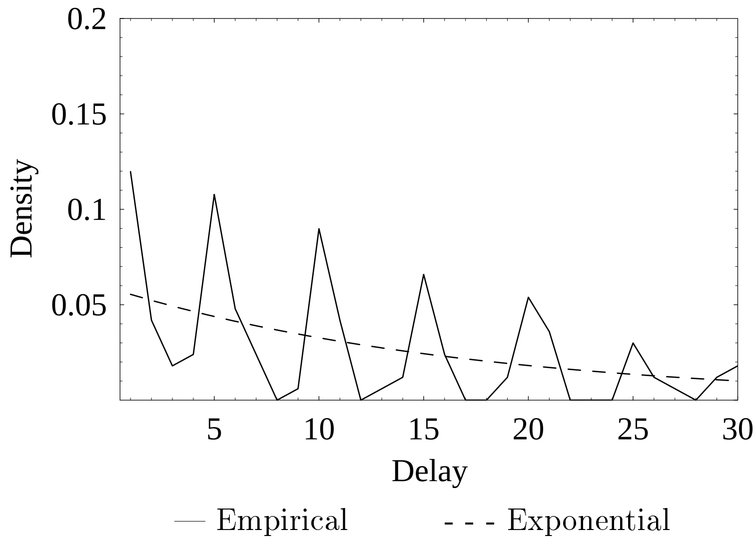


Figure 5: Empirical delay distribution of customer A2

deviation of positive delays. In the table, 'expo' refers to the exponential distribution.

The overall conclusion is that empirical delays are very much skewed to the right, and they can be approximated by the beta, Weibull or exponential distribution, which often give similar approximations. Also, the gamma distribution occasionally gives good approximations. The service level, in terms of the no-delay probability, provided from DAN is poor. Especially for class AA and A customers, who should have great attention, the amount of orders delivered without any delay or with only a small delay is very small.

Cust.	$E(t)$	σ_t	Comments on the test (Acceptance based on KS test)
AA1	4.09	7.17	Big atom in 1. None accepted, but beta + expo give low S.E.
AA2	8.04	13.40	None accepted. Beta, gamma, lognormal+expo give low S.E.
AA3	2.25	4.18	Big atom in 1. None accepted, but beta + expo give low S.E.
AA4	4.34	6.30	None accepted; however beta + expo give low S.E.
A1	12.80	9.35	No point in testing: Only 16 obs (uniformly distributed).
A2	17.00	16.40	Beta and expo are accepted and give low S.E.
A3	14.40	14.80	Expo, gamma + Weibull are accepted and give low S.E.
A4	5.77	9.22	Beta accepted. Weibull and expo also give low S.E.
A5	14.10	12.50	Erlang, expo and beta are accepted and give low S.E.
A6	12.60	10.20	Erlang, expo and beta are accepted and give low S.E.
A7	21.70	27.00	Only 22 obs. Weibull, gamma+lognormal accepted; give low S.E.
A8	1.00	0	No testing: Only 5 delays that all equal one.
A9	0.00	0	No testing: All 14 orders are delivered immediately (no delays).
B1	21.00	15.50	Only 23 observations. All distributions fit badly.
B2	17.30	14.20	Expo, triangular, beta and normal accepted; give low S.E.
B3	5.22	5.52	Beta accepted by χ^2 test. Beta and gamma give low S.E.
B4	5.22	8.90	Beta+Weibull accepted, low S.E. (Expo accepted by χ^2).
C1	2.71	2.88	Weibull + expo are accepted; however beta give lowest S.E.
C2	3.46	5.13	None accepted; however Weibull gives very low S.E.
D1	2.55	5.95	None accepted; however Weibull and expo give very low S.E.
D2	3.64	5.74	Expo accepted; however also Weibull and beta give low S.E.
E1	4.25	2.22	No point in testing: Only 4 delays (uniformly distributed)
E2	2.19	2.84	Expo accepted; however also Weibull and beta give low S.E.

Table 7: Summary of distribution analysis

2.4 Other effects on lead time

As mentioned in the introduction, it is important to analyse what factors influence the lead time. First of all, if the lead time depends on the order size, this has to be accounted for in the inventory model. Also, consecutive lead times may not be independent as is normally assumed in inventory modelling. Hence, we need to analyse what pattern of behaviour is likely to be observed in practice. If the supplier is out of stock, it is likely that the supply problem will take some time to clear. Therefore, consecutive lead times may be longer than average, and then the lead time of successive demands are strongly autocorrelated rather than independent. Finally, the lead time may vary over time in a cyclic or seasonal pattern. These three aspects will be analysed empirically in this final paragraph of the case study.

Correlation between lead time and order sizes

The size of the order may have an impact on the lead time. However, this is mostly the case when products are made to order (MTO), where the production time is related to the quantity produced. Since DAN is a wholesaling company and in general promises instant delivery, it is classified as a make to stock (MTS) company. However, to see whether there could be some correlation between lead time and order size a plot is made for orders placed by customer AA1. This is shown for actual lead times in Figure 6.

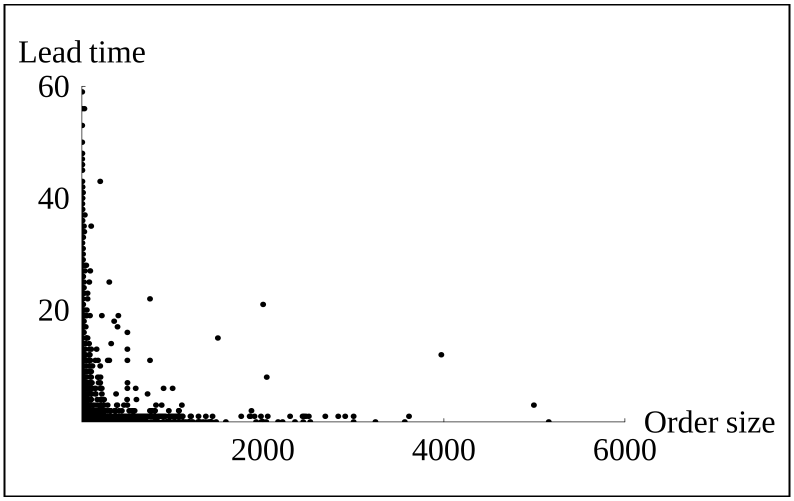


Figure 6: Order size versus actual lead time for customer AA1

In this graph, it can easily be seen that there is no positive correlation between actual lead time and order size; so there is no need for further statistical analysis. Similar plots have been made for other customers and article numbers, and they all show the same pattern. The relation has also been plotted for the delay as well as the promised lead time. These test plots also show the same pattern and will therefore be omitted here. As expected these empirical data do not support the hypothesis that the order size has a positive impact on the lead time. On the contrary, based on these data it could be argued that there is some negative correlation meaning that big orders are always delivered quickly after the receipt of the order, and the orders that are postponed are only small orders. However, there is a big mass of points with low quantities and low lead times; so to state a negative correlation would be too strong.

Correlation of lead times of successive orders

Since actual lead times and lead time delays are the same for 97% of the orders, they share the properties of dependence. The autocorrelation analysis will be done at customer level, since this is the level of analysis, which is the basis for inventory control at each customer.

Let x_t denote the actual lead time of order number t placed by a specific customer. Let $V(z)$ denote the empirical variance of variable z. If the lead time of two successive orders are independent, then $V(x_t - x_{t-1}) = 2V(x_t)$. If, on the other hand, there is dependence between the lead time of two successive orders, then $V(x_t - x_{t-1}) < 2V(x_t)$. Hence, by calculating the empirical variances of the actual lead times and of the first order differences, it can be indicated whether successive orders are independent or not. Also x_{t-1} can be plotted against x_t, and if there is a systematic slope there is some dependence. Finally,

x_t can be standardised by $v_t = \frac{x_t}{\bar{x}}$ and transformed into a new variable, $r_t = exp(-v_t)$, which is rectangularly distributed. Then, r_{t-1} can be plotted against r_t to see if there is a systematic pattern within the unit rectangle.

In the raw data available, orders within a day have been sorted by an unknown factor, and since the time of the day that the order was placed is not available, the original sequence of transactions can not be restored. Therefore, we can only analyse those customers that placed at most one order per day during 1999. 6 of the 9 customers in class A fulfil this requirement, and therefore they are chosen for the autocorrelation analysis.

Customer A1 placed 45 orders during 1999. Here $V(x_t - x_{t-1}) = 924$ and $2V(x_t) = 926$, which do not suggest that there is any correlation between successive lead times. The (x_{t-1}, x_t)-plot can be seen in Figure 7, and the (r_{t-1}, r_t)-plot can be seen in Figure 8. These figures do not reveal any clear systematic behaviour; so it can not be concluded that successive lead times are correlated. The unit sloped line in Figure 7 is only used to detect any systematic behaviour and does not represent any kind of regression line.

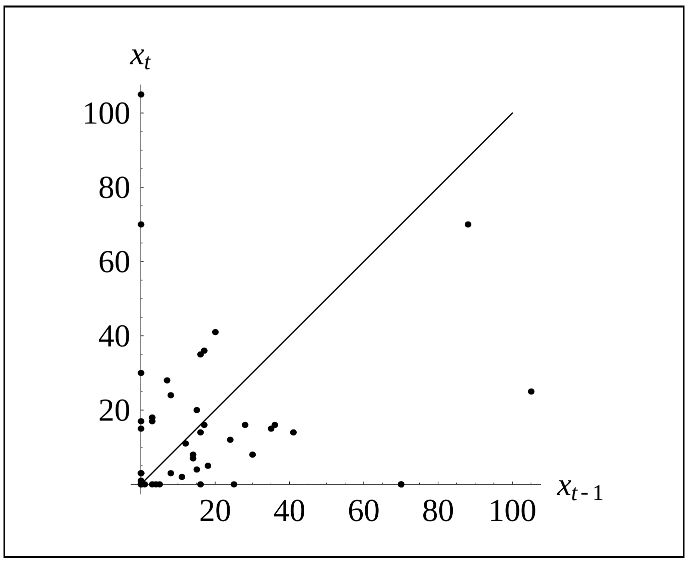


Figure 7: (x_{t-1}, x_t) plot for customer A1

Hill [4] proposes to analyse the conditional probability that an order is delayed given that the previous order placed was not delayed. Therefore, let event A correspond to a delay of an order, and let event B correspond to no delay of the previous order. Then

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

If delays of successive orders are independent, this conditional probability will be equal to the unconditional probability that an order is delayed, i.e. $P(A \mid B) = P(A)$. Therefore, both the conditional and the unconditional probability will also be calculated from the empirical data. For customer A1, the conditional probability, $P(A \mid B)$, is 0.56 and the unconditional probability, $P(A)$, is 0.42, which may suggest some autocorrelation. This

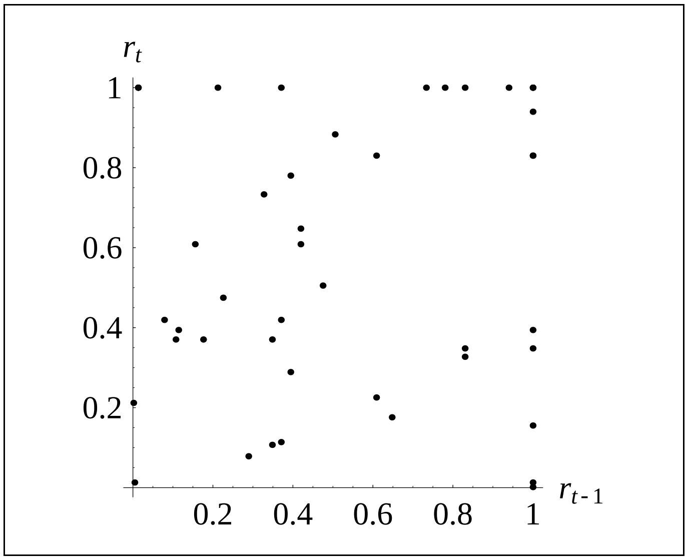


Figure 8: (r_{t-1}, r_t) plot for customer A1

correlation can be tested statistically by testing for independence in a 2x2 contingency table using Pearson's chi square test. The test probability is denoted $P(\chi^2)$.

The autocorrelation analysis for each of the 6 chosen customers is shown in Table 8.

Cus-tomer	# of orders	Actual lead times					Delay			Con-clusion
		$V(x_t\text{-}x_{t-1})$		$2V(x_t)$	(x_{t-1}, x_t)	(r_{t-1}, r_t)	$P(A\|B)$	$P(A)$	$P(\chi^2)$	
A1	45	924	$\approx$	926	NS	NS	0.56	$>$ 0.43	0.168	Indep
A3	159	282	$\approx$	280	NS	NS	0.43	$\approx$ 0.48	0.154	Indep
A5	108	219	$\approx$	263	S	NS	0.51	$=$ 0.51	0.934	Indep
A6	199	358	$\approx$	375	S	NS	0.38	$\approx$ 0.41	0.219	Indep
A7	50	1333	$<$	1956	S	S	0.21	$<$ 0.48	0.001	Dep
A8	20	0.21	$<$	0.39	TFO	TFO	0.13	$<$ 0.25	0.037	Dep

Table 8: Autocorrelation analysis of chosen customers

In Table 8 'TFO' means that there are too few observations to conclude anything, 'NS' means that there is no systematic pattern in the plot and 'S' means that there could be some systematic pattern. Based on the results in Table 8 it appears that some of the customers indeed experience autocorrelated lead times and delays of successive orders placed at DAN. Especially customer A7 experiences strong correlation in lead times, since $2V(x_t)$ is much bigger than $V(x_t - x_{t-1})$. Also, both plots show some systematic behaviour, where especially $x_{t-1} > x_t > 0$ seems to be the pattern. Finally, $P(A)$ is more than twice as big as $P(A \mid B)$ suggesting quite a high correlation between successive lead time delays. This is supported by Pearson's chi square test leading to independence ('Indep') between successive delays for customers A1, A3, A5 and A6, whereas successive delays are dependent ('Dep') for A7 and A8. The final conclusion is that customers A7 and A8 experience a supply process where the lead time of consecutive order arrivals are autocorrelated meaning that long delays follow each other.

Cyclic or seasonal effects on lead time

To see whether there could be some cyclic or seasonal effects on the lead time two figures have been made. In Figure 9 the weekly average lead time has been calculated and is depicted with the average lead time of the whole year (2.6). It does not reveal any systematic behaviour; however, a few outliers are present. In week 10, the average lead time is extremely low (1.1) and in weeks 42 and 45 the average lead time is extremely high being 5.4 and 4.8, respectively. The normal distribution fits the weekly average lead time very nicely and 96% of the values lie between 1 and 4.

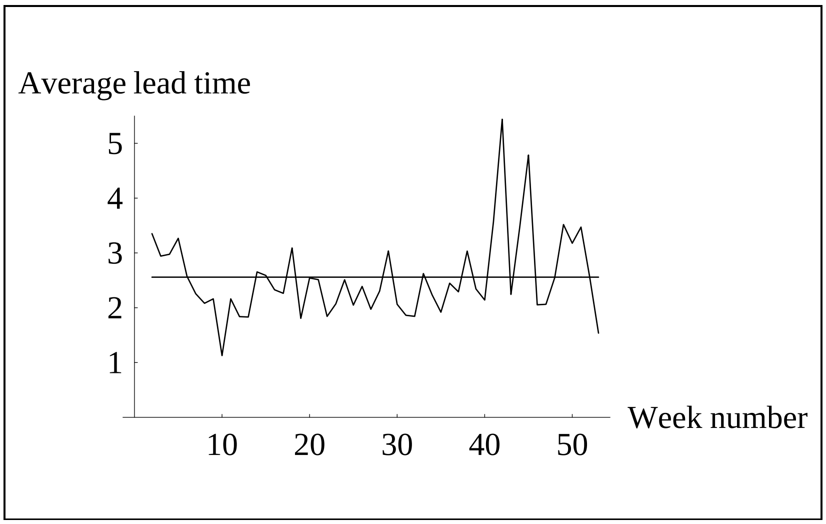


Figure 9: Seasonal effects of week number on actual lead time

To see if there is any seasonal effect the monthly average lead time has been calculated and is depicted in Figure 10 with the average lead time of the whole year.

Based on only one year of data and no additional information it is difficult to judge whether there is any seasonal effect or not. The fluctuations could be random; however, by testing the data statistically the lead times during the winter months, October, November, December and January are significantly higher than the lead time in the rest of the year. For this particular empirical data set it is not reasonable to conclude that there are seasonal changes in the lead time without more information; however, it indicates that it is important to analyse for seasonal variations in lead time data. What is really intriguing then is to find the causes for such a seasonal variation in the lead time.

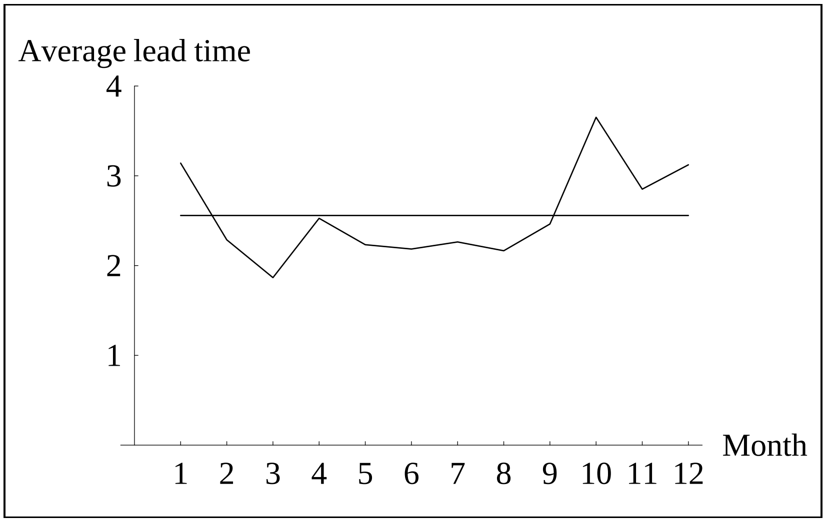


Figure 10: Seasonal effects of month on actual lead time

3 Delays in a two-level inventory system

In this section a two-level serial distribution system is modelled, and by exact analysis an expression for the delay density function is presented and analysed. The model analysed in this section is suggested by Hill [4], but some additional considerations and analyses will be presented here.

3.1 The model

Consider a two-level supply system consisting of one retailer (R) that experiences external customer demand and one warehouse (W) that receives replenishment orders from the one retailer only. W has infinite capacity, i.e. never runs out of stock. The arrival of customer demands at the retailer follows a Poisson distribution with mean λ. Customer demands are unit sized; hence, mean demand per unit time is λ. Customers are served according to a first come first served basis, and unfilled demands at R are completely backordered. If R has sufficient stock to meet a customer demand the lead time is L_R, otherwise there is an additional delay. The lead time from W to R is L_W. The inventory position (stock on hand - backorder + outstanding orders) at the retailer is maintained at the level S, hence a base-stock $(S-1, S)$ inventory control policy is applied. If $S \leq 0$ then all customer demands are passed directly on to the warehouse, and the total customer lead time is then $L_R + L_W$. Hence, the delay time faced by the external customers at R can not exceed L_W.

Based on this model it will be analysed how the delays are distributed. By extending the work of Hill [4], the mean delay will be determined, and it will also be analysed how the delay distribution depends on the lead time from the warehouse to the retailer, L_W, and the order-up-to-level, S. Moreover, a decision rule, for determining the value of S given L_W subject to a service level constraint, is developed.

3.2 The delay distribution and mean delays

This model can be viewed by the retailer as a queueing system with Poisson customer arrivals with rate λ, a fixed service time L_W and S identical servers in parallel. The state of the system is represented by the number of outstanding replenishment orders (between R and W) denoted by n. The steady state probability of the system being in state n is an L_W-fold convolution of the Poisson demand. Then,

$$p(n) = \frac{(\lambda L_W)^n e^{-\lambda L_W}}{n!}, \qquad n = 0, 1, 2, ...$$

Hence, n follows a Poisson distribution with parameter λL_W. With probability $p_0 = \sum_{n=0}^{S-1} p(n)$ the demand of an arriving customer will be met directly from stock, and the customer will experience a lead time of L_R. Hence, p_0 represents the no-delay probability, $P(t = 0)$, also corresponding to the ready rate (here also equal to the fill rate since demand is unit sized). Otherwise, the customer will experience a lead time of $L_R + t$, where t is the additional delay due to a stockout. Since R may already have orders outstanding, the delay may be less than L_W. In Hill [4], the expression of the density function of t of such a model is derived, and it is

$$f(t) = \frac{\lambda^S (L_W - t)^{S-1} e^{-\lambda(L_W - t)}}{(S-1)!}$$

The expression of the density function has the form of a truncated gamma distribution with shape parameter S and scale parameter λ when viewed backwards in time from L_W. The distribution is truncated, because t can only take values between 0 and L_W. The distribution of the delay (when viewed backwards from L_W) is therefore gamma distributed. The backward density is illustrated in Figure 11, where the top left hand graph is the density of $L_W - t$ truncated in L_W, and the top right hand graph is the density of t truncated in 0. As can be seen these two graphs are the mirror image of each other. The bottom graphs correspond to the cumulative distribution function of the density above. The truncated part corresponds to p_0.

Since S is an integer, the distribution is also called Erlang. In the special case of $S = 1$ the delay is exponentially distributed in a "mirrored" way, again meaning that the delay is viewed backwards from L_W, because this is the maximum extra time the customer order can be delayed.

In the non-truncated gamma distribution (i.e. t may be less than 0), the mean of the distribution is $E(L_W - t) = S/\lambda$. Hence, the mean of the delay time is $E(t) = L_W - S/\lambda$. In the non-truncated gamma distribution the mean delay time may be negative (if $\lambda L_W < S$). The mean of the truncated distribution, however, is

$$E(t^{trunc}) = \int_0^{L_W} t f(t) dt$$

If a customer arrives when there is no stock on hand, it would probably be nice for that customer to know the expected additional delay. Therefore, the truncated mean conditional on $t > 0$ is also relevant. This is determined by

$$E(t^{trunc} | t > 0) = \frac{E(t^{trunc})}{1 - p_0}.$$

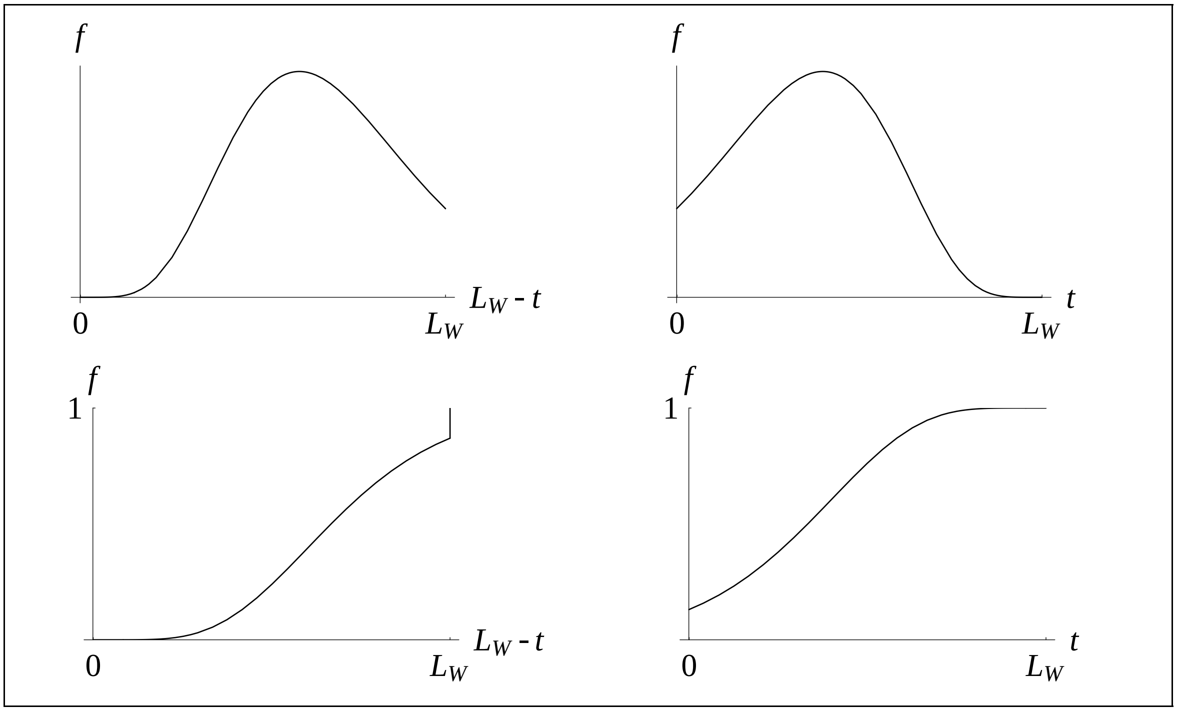


Figure 11: The delay density function - an illustrative example

3.3 Numerical results

Impact of L_W and S on the delay distribution

To see how the delay distribution depends on L_W and S, the first order partial derivatives could be derived. The expressions are quite complex, however, so distributions based on numerical examples will provide a better insight. Appendix A contains 12 figures illustrating the delay densities for 12 different combinations of L_W and S keeping L_R and λ fixed. So, $L_R = 0$ to concentrate only on the additional delay, and $\lambda = 10$. the values of L_W and S are: $L_W \in \{1, 4, 8\}$ and $S \in \{1, 5, 10, 20\}$.

For each of the 12 combinations also p_0, $E(t^{trunc})$ and $E(t^{trunc}|t > 0)$ have been calculated, which can be seen at each of the figures in Appendix A.

Obviously, a combination of low L_W and high S gives a low expected delay time and a high no-delay probability. S seems to have a greater effect on $f(t)$ than L_W, however a change in S has a greater effect on $f(t)$ for low values of L_W. As can be seen from the graphs with $S = 1$, the distribution is a mirror image of the exponential distribution.

The truncated mean is strictly positive. In the cases where p_0 is close to 0, the truncated mean is almost equal to the non-truncated mean, since there is not so much to truncate. The higher p_0 is, the closer the truncated mean is to 0, which is quite natural, since a high p_0 means that delays seldom occur, and therefore the truncated mean is pooled towards 0. So for high p_0, the conditional mean is quite interesting, whereas the conditional mean is very close to the unconditional mean for low p_0. For instance, for $S = 10$ and

$L_W = 1$ the probability of a delay, $P(t > 0)$, is 1-0.46=0.54. It can also be seen that given that the order is delayed, the expected delay will be 0.23 time units, whereas the unconditional mean delay is 0.12 time units, which is nearly the half of the conditional mean.

Trade off curves of order-up-to-level versus service level

In this model p_0 represents the service level as noted above. Let γ be the target service level, then given λ and L_W, choose S such that

$$\gamma = \sum_{n=0}^{S-1} \frac{(\lambda L_W)^n e^{-\lambda L_W}}{n!}$$

We found that $(L_W - t) \sim \Gamma(S, \lambda)$ and $n \sim P(\lambda L_W)$, and since S is a positive integer, following Hoel, Port and Stone [5, p. 130], another way of writing p_0, and thereby γ, for $S \geq 2$ is

$$\gamma = 1 - \int_0^{L_W} \frac{\lambda^S (L_W - t)^{S-1} e^{-\lambda(L_W - t)}}{(S-1)!} d(L_W - t) \tag{1}$$

Hence, from given values of λ, S and L_W, the resulting service level, γ, can be determined. Also, given a target service level, S can be determined by trial-and-error. Equation 1 is therefore the decision rule for determining the reorder point subject to a service level constraint.

The trade off between the order-up-to-level and the service level is illustrated in two exchange curves in Figure 12 for $L_W \in \{1, 8\}$ and $\lambda = 10$.

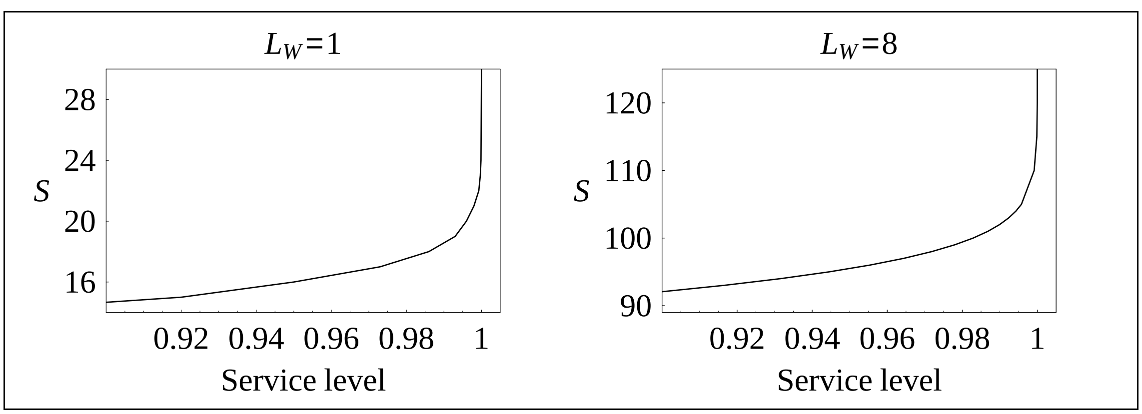


Figure 12: Exchange curves of service level versus order-up-to-level

As typical for this kind of exchange curve a unit change in the desired service level only requires a small change in order-up-to-level for low service levels, whereas a unit change in desired service level requires a big change in order-up-to-level for big service levels (above 0.99).

To see the relation between L_W and S, Figure 13 is presented for γ=\{0.90, 0.95, 0.99, 0.9999\} and $\lambda = 10$. The figure shows that the S that fulfils the target service level increases quite a lot with L_W. For instance, for $\gamma = 0.95$, $S = 16$ for $L_W = 1$, $S = 51$ for $L_W = 4$ and $S = 95$ for $L_W = 8$. The curves appear to be approximately linear, and a unit increase in L_W for this particular example ($\gamma = 0.95$) leads to an increase in S of 10.4 units. The target service level does not have a big impact on the slope of the curve, and it has only a minor impact on the level when $\gamma \leq 0.99$, which is precisely what the exchange curves above indicate. For target service levels above 0.99, there is a shift in the level meaning that, for a given wholesaler lead time, S increases a lot with the target service level, which is also indicated by the exponential increase in the exchange curves above.

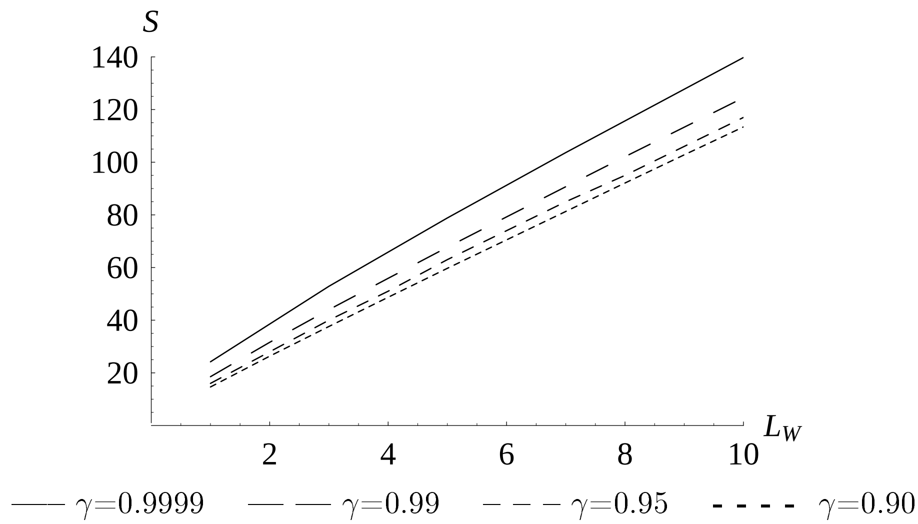


Figure 13: (L_W, S)*-plot for* $\lambda = 10$

Correlation between successive orders

Hill [4] examines the association between successive lead times of this two-level queueing model. Therefore, this will not be repeated here, but the results will merely be summarised. The conditional probability that a customer order is delayed given that the previous customer order was not delayed is derived. If delays of successive orders are independent, this conditional probability will be equal to the unconditional probability that an order is delayed as argued in the previous section. Based on numerical examples both the conditional and the unconditional probability are calculated, and Hill concludes that there is a high serial correlation between consecutive lead times.

As for the autocorrelation analysis of the empirical data, this suggests a need for adaptations in inventory control models to account for dependence between lead times of consecutive orders.

4 Delays in an N-level serial distribution system

In this section an N-level serial distribution system is modelled for the (s, Q) ordering policy. By simulation, the delay distribution at the final echelon is determined. A distribution analysis is then conducted to see which distributions fit well.

4.1 The model

Consider a serial supply chain consisting of N consecutive suppliers. Each supplier i faces Poisson demand (with mean λ=1) from external customers and replenishment orders from the downstream supplier, $i+1$. There is a fixed service time between all suppliers, but if supplier i is unable to fill the entire order from supplier $i+1$ the order is delayed until this is possible. The realised lead time is therefore a stochastic variable. Customers are served on a first come first served basis, and all demands and orders that are not satisfied immediately are backordered. All echelons use the (s, Q) ordering policy to replenish their stock. Whenever the inventory position (defined as the net stock + outstanding orders - backorders on hand) is lower than the reorder level, s, an order of size Q is placed. The order size, Q, is assumed the same for all echelons. The reorder level at supplier i is determined by

$$s_i = (1+k)\tau\lambda(N+1-i)$$

where k is interpreted as the safety factor, and τ is the service time, which for convenience is assumed the same between all suppliers. Let μ_i and σ_i denote the mean and standard deviation of demand during lead time at supplier i. Then for $i = N$, and also for $Q = 1$, $s = (1+k)\mu_i = \mu_i + k\sigma_i$, since $\mu_i = \sigma_i$ for Poisson demand, which is the traditional formula for determining the reorder level used in classical continuous review policies. Although Q can take values greater than 1, we will also use this decision rule for the reorder point at echelon 1 to $N-1$ to obtain an easy approximate value.

Economic considerations are not in focus here, and therefore no optimisation of k and Q takes place. Instead, these two values are varied as experimental factors to see the effect on the delay distribution faced by the external customer at the final supplier (N). The multi-level distribution system is illustrated in Figure 14 below.

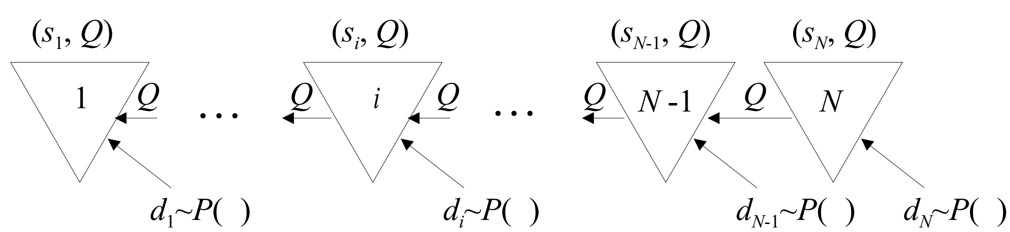


Figure 14: Illustration of the N-level distribution system

The system is modelled by simulation. The simulation is run for 15 combinations of experimental factors, which are the service time, τ, the number of suppliers, N, the safety factor, k, and the order size, Q. The 15 scenarios of the experimental design are represented in Table 9 with some of the results. Scenario 2 is the base case, which is used for comparison with groups of other scenarios.

Each simulation is run for 100,000 time units. Each time an external customer arrives at the final supplier the waiting time is recorded. If the waiting time is greater than the service time a delay has occurred. Hence, only the delay distribution faced by the external customers at echelon N is analysed in detail.

4.2 Numerical results

The results, in terms of the no-delay probability and the mean conditional delay, of the simulations for the 15 scenarios can be seen in Table 9.

Scenario	τ	N	k	Q	$P(t=0)$	$E(t)$
1	4	10	0.2	10	0.767	2.14
2	4	10	0.5	10	0.944	1.05
3	4	10	1	10	0.992	0.61
4	4	10	2	10	0.999	0.45
5	4	10	0.2	1	0.654	3.88
6	4	10	0.5	1	0.859	2.05
7	4	10	1	1	0.965	1.37
8	4	10	2	1	0.999	0.78
9	4	10	0.5	2	0.859	1.40
10	4	10	0.5	4	0.894	1.01
11	4	2	0.5	10	0.947	1.03
12	4	4	0.5	10	0.943	1.06
13	4	20	0.5	10	0.945	1.05
14	1	10	0.5	10	0.783	1.48
15	10	10	0.5	10	0.980	1.14

Table 9: Experimental factors, probability of no delay and conditional mean

The mean delay, $E(t)$, is conditional on the delay being strictly positive. Hence, given there is a delay the mean delay is given by $E(t)$. The probability of no delay, $P(t=0)$, is measured as 1 minus the fraction of total customers facing a positive delay. This no-delay probability is actually the same as the probability of no stockout at any arbitrary time period, which is the same as the ready rate service measure.

By comparing scenarios 1 to 4, it can be seen that the probability of no delay, $P(t=0)$, is quite high for high safety factors. As expected, the probability of no delay increases with the safety factor, which implies higher safety stocks. The same applies for scenarios 5 to 8, however the difference is more severe due to the low order size. In fact, scenarios 2 and 6 can be compared with scenarios 9 and 10 to see the effect of increasing the order size as a means for protection. Obviously, the probability of no delay increases as Q increases, which is the nature of this kind of service measure.

To see if there is any effect of the number of suppliers in the chain on the delays faced by the final customer, scenarios 2, 11, 12 and 13 are compared. By comparing these 4 scenarios, it can be concluded that there is no difference between the no-delay probabilities and there is no difference betwen the conditional mean delays. Therefore, from the end-item customer's point of view there is no difference as to whether there are 2, 4, 10 or 20 consecutive suppliers in the chain, which is quite important from a supply chain management perspective.

By comparing scenarios 2, 14 and 15 in Table 9, it can be concluded that the service time also has impact of the no-delay probability; however, in a somewhat surprising way. The no-delay probability actually increases with the service time. The reason is that the reorder level is over-compensating for this lead time increase, and therefore the high service time influences the safety stock directly, which then serves as extra means of protection.

It should be emphasised that this simulation is an exploratory study to analyse delay distributions. Also, we did not attempt to optimise the system parameters, such as the reorder level and the order size, for all echelons. Therefore, the interpretations and conclusions made above regarding the comparison of scenarios should not be generalised to other systems.

4.3 Distribution analysis

The simulated delay distributions of each scenario can be seen in 15 separate graphs in Appendix B. These distributions only include strictly positive delay data. Hence, the atom in delay=0 is excluded from the distribution analysis due to the idea of modelling the lead time as a probability of a fixed lead time and a probability of a delay. It is the probability distribution of the positive delays that is of interest.

As for the empirical delay data in Section **??**, an extensive distribution analysis has been conducted of the simulated delay data. A number of statistical distributions have been fitted and tested by Pearson's chi-square test and the Kolmogorov-Smirnov (KS) test (at a 0.98 significance level). However, when the number of observations gets too high such tests are seldom accepted. In general, the fits of the distributions are not accepted, which is not surprising, but based on the squared error (S.E.), the relative precision of the distributions are evaluated and compared. The beta, Weibull and gamma distributions give the best approximations in most of the 15 scenarios, and often these three distributions are close to each other. Also, the exponential distribution gives reasonably good approximations in many of the scenarios. Therefore, for each scenario, the beta, Weibull, gamma and exponential distributions are also shown in the graphs in Appendix B having the same mean and standard deviation as the simulated delay data.

An example of the results of the distribution analysis is presented in Table 10 for scenario 13.

The results of the distribution analysis for all the scenarios are summarised in Table 11 below, where $E(t)$ is the mean conditional delay, and σ_t is the standard deviation of the conditional delay.

Distribution	S.E.	χ^2	$P(\chi^2)$	KS	P(KS)
Beta	0.00027	105	<0.005	0.017	0.0963
Weibull	0.00036	51	<0.005	0.015	>0.15
Gamma	0.00059	77	<0.005	0.026	<0.01
Exponential	0.00195	187	<0.005	0.056	<0.01
Erlang	0.00195	187	<0.005	0.056	<0.01
Lognormal	0.00547	650	<0.005	0.081	<0.01
Normal	0.01450	2100	<0.005	0.118	<0.01
Triangular	0.02520	4120	<0.005	0.372	<0.01
Uniform	0.04970	1100	<0.005	0.568	<0.01

Table 10: Distribution analysis - scenario 13

Scenario	$E(t)$	σ_t	Comments on the test
1	2.14	1.790	None accepted; however beta, gamma+Weibull give very low S.E.
2	1.05	0.906	Beta+gamma accepted. Weibull+expo are close and give low S.E.
3	2.75	0.605	Beta, gamma, Weibull + exponential are accepted.
4	0.45	0.330	Only 11 observations, no point in testing. Looks uniform.
5	3.88	2.020	Atom in 4. None accepted; beta, gamma, Weibull+normal: low S.E.
6	2.05	1.170	None accepted; however beta, gamma+Weibull give very low S.E.
7	1.37	0.825	Beta is accepted. Weibull + gamma close and give low S.E.
8	0.78	0.554	Only 63 obs.: All accepted; lognormal+Weibull lowest S.E.
9	1.40	1.020	None accepted; however beta, gamma+Weibull give very low S.E.
10	1.01	0.840	None accepted; however beta give very low S.E.
11	1.03	0.867	Beta accepted. Weibull, gamma+expo are close and give low S.E.
12	1.06	0.924	None accepted; however beta, gamma+Weibull give very low S.E.
13	1.05	0.888	Beta+Weibull are accepted. gamma+exponential also give low S.E.
14	1.48	1.210	None accepted; however beta, gamma+Weibull give very low S.E.
15	1.14	1.140	Beta, gamma + Weibull are accepted and give very low S.E.

Table 11: Summary of distribution analysis

As can be seen from Table 11 and the graphs in Appendix B, the empirical delay distributions are generally fitted by a beta distribution skewed very much to the right. The gamma and the Weibull distributions are often close to the beta distribution, and they also give a low S.E. The exponential distribution has the same intuitive appearance as the empirical data. However, often the exponential distribution overestimates the number of low delays and underestimates the number of medium and high delays. The conclusion of the distribution analysis is that delays can be approximated well with the beta, gamma or Weibull distributions.

5 Summary and conclusions

In this paper a number of analyses has been carried out to find an appropriate way of modelling lead time delays. An idea was proposed of modelling lead time as the mixture of a fixed lead time occurring with a certain probability, p_0, and an additional delay occurring with probability $1 - p_0$.

A case study was first conducted to explore how empirical delays are distributed. The distribution analysis was carried out at customer level, and therefore customers were classified based on a bi-criteria ABC classification analysis. The two criteria applied were customer size, measured by purchase volume as a percentage of total annual sales volume, and loyalty, measured by the number of orders. 23 customers, 13 of which were classified as important, were chosen for detailed analysis. Then the distribution analysis of these 23 customers was carried out, which showed that the empirical delay distribution is skewed to the right with a long tail. The beta, Weibull, gamma and exponential distributions gave good approximations with the beta and the exponential as the two dominant ones. It was also concluded, that even though the company promises instant delivery to its customers, the orders are often delayed. Hence, the service level, p_0, which represents the no-delay probability, was much lower than promised. The case study also contained an analysis of other factors that were expected to have an impact on the lead time delay. The order size did not have any impact, however some customers were found to experience autocorrelated lead times and delays.

Since delays are most likely to occur due to stockouts at higher levels of suppliers, the delay is related to the duration of such a stockout. Theoretical analyses were therefore carried out to examine the impact of stockouts on delays at lower levels. An exact expression of the delay density of a two-level queueing system using a base-stock policy was presented. This expression was a truncated gamma density function. Numerical results showed that increasing the order-up-to-level or decreasing the lead time between retailer and wholesaler increases the service and decreases the mean waiting time. This is obvious; however, the numerical results also showed that an increase in the order-up-to-level has bigger impact on service and waiting time for low values of wholesaler lead time than for high values of the wholesaler lead time. When analysing exchange curves between the order-up-to-level and the service level, it was found that a unit increase in the service level requires a small increase in the order-up-to-level for service levels below 0.99, whereas a high increase in the order-up-to-level was required for service levels higher than 0.99. It was also found that for a given target service level, the order-up-to-level increases approximately linearly with the wholesaler lead time. Furthermore, based on the work of Hill [4], it was concluded that high correlation exists between successive lead times.

Finally, an N-level serial distribution system using the (s, Q) ordering policy was modelled. Through a simulation experiment, it was concluded that the no-delay probability, i.e. the service level, increases with the safety factor and the order size, whereas the mean waiting time (conditional on a positive delay) decreases. The number of levels in the chain does not have any impact on the service nor the waiting time faced by the end-user customer, which is a very important result from a supply chain management perspective. However, the internal service levels differ for different horizontal positions within the chain. A distribution analysis was carried out on the simulated delays faced by the end-user customer. The delay densities look exponential, but the beta, gamma and Weibull distributions provide better approximations.

One important conclusion is that autocorrelation indeed appears between consecutive lead times suggesting a need for adaptation in inventory control models to account for this dependence.

Another conclusion is that both the empirical lead time data and the theoretical models, in terms of the 2-level queueing model and the N-level distribution system, supported the idea of modelling lead time as a mixture of a fixed lead time and the occurrence of a delay. Moreover, the delay was generally approximated well with the beta, Weibull, gamma or the exponential distributions. In Gudum [3], we use the information gained on the delay distribution to propose a new way of modelling lead time. This is used to develop a new compound lead time demand distribution approach in order to compare performance of various decision rules in continuous review inventory control models.

References

[1] U. Bagchi, J. C. Hayya, and C.-H. Chu. 1986. The effect of lead-time variability: The case of independent demand. *Journal of Operations Management*, 6(2), 159–177.

[2] B. E. Flores and D. C. Whybark. 1986. Multiple criteria ABC analysis. *International Journal of Operations and Production Management*, 6(3), 38–46.

[3] C. K. Gudum. 2002. *Managing variability in a supply chain: An inventory control perspective*. PhD thesis, Copenhagen Business School. Denmark.

[4] R. Hill. 1999. Lead time behaviour in inventory modelling. In *Proceedings of the Fourth ISIR Summer School on Inventory Modelling*, pages 119–125, Exeter, UK.

[5] P. G. Hoel, S. D. Port, and C. J. Stone. 1971. *Introduction to Probability Theory*. Houghton Mifflin Company, Boston.

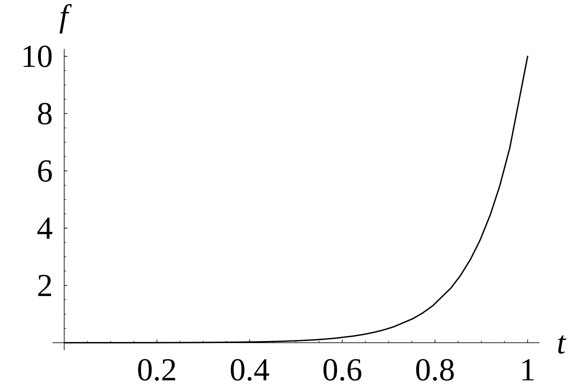
[6] U. S. Karmarkar. 1993. Manufacturing lead times, order release and capacity loading. In S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4 of *Handbooks in Operations Research and Management Science*, pages 287–329. Elsevier Science Publishers, Amsterdam.

[7] E. A. Silver, D.F. Pyke, and R. Peterson. 1998. *Inventory Management and Production Planning and Scheduling*. John Wiley and Sons, New York, third edition.

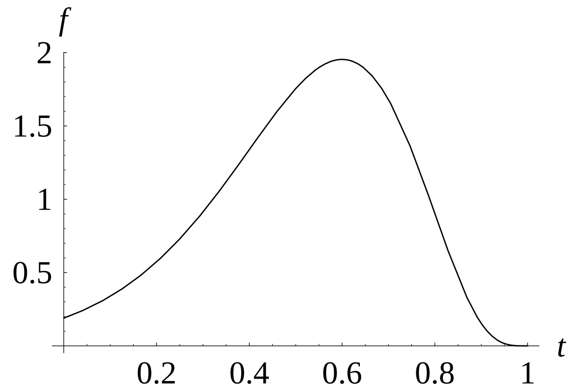
[8] R. J. Tersine. 1994. *Principles of inventory and materials management*. Prentice Hall, New Jersey, fourth edition.

Appendix A Truncated gamma densities of delay

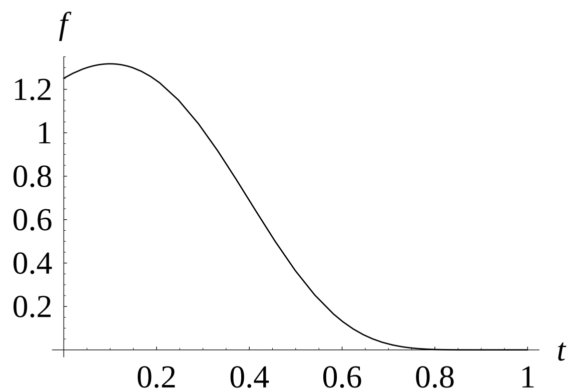
Illustrations of the delay distribution in the 2-level queueing model.



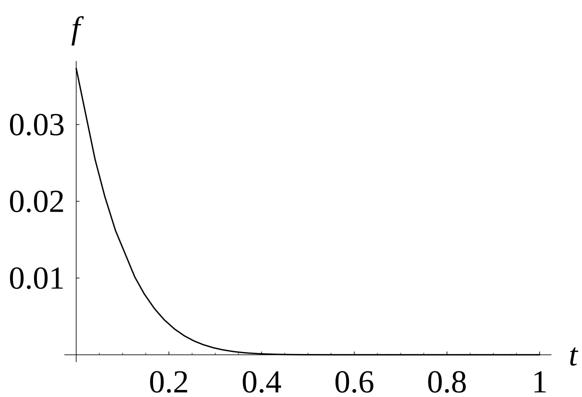
$\lambda = 10$
$L_W = 1$
$S = 1$
$E(t^{trunc}) = 0.9$
$E(t^{trunc}|t > 0) = 0.9$
$p_0 = 0$



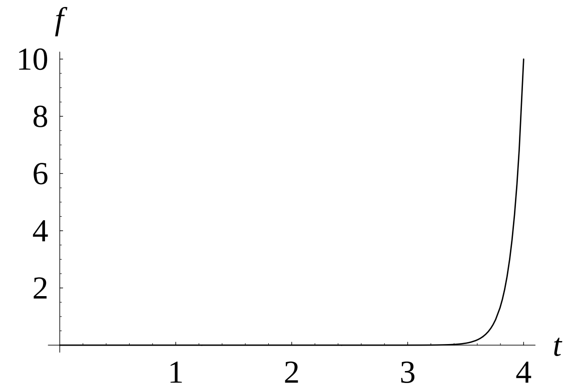
$\lambda = 10$
$L_W = 1$
$S = 5$
$E(t^{trunc}) = 0.5$
$E(t^{trunc}|t > 0) = 0.52$
$p_0 = 0.03$



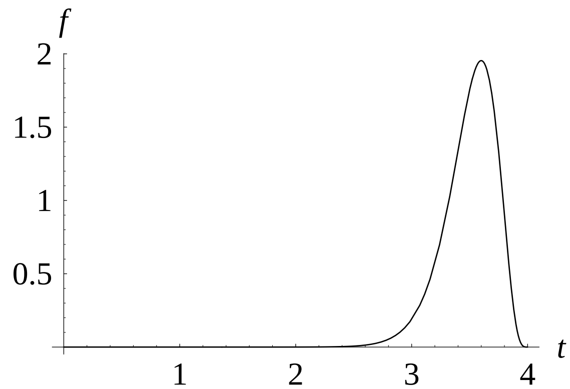
$\lambda = 10$
$L_W = 1$
$S = 10$
$E(t^{trunc}) = 0.12$
$E(t^{trunc}|t > 0) = 0.23$
$p_0 = 0.46$



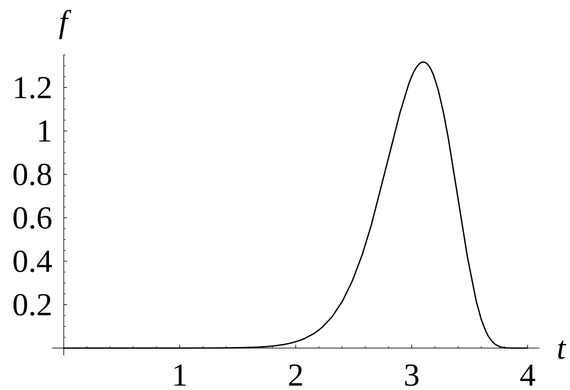
$\lambda = 10$
$L_W = 1$
$S = 20$
$E(t^{trunc}) = 0.003$
$E(t^{trunc}|t > 0) = 0.08$
$p_0 = 0.996$



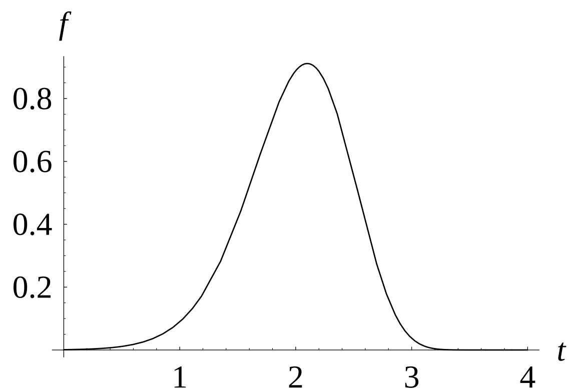
$\lambda = 10$
$L_W = 4$
$S = 1$
$E(t^{trunc}) = 3.9$
$E(t^{trunc}|t > 0) = 3.9$
$p_0 = 0$



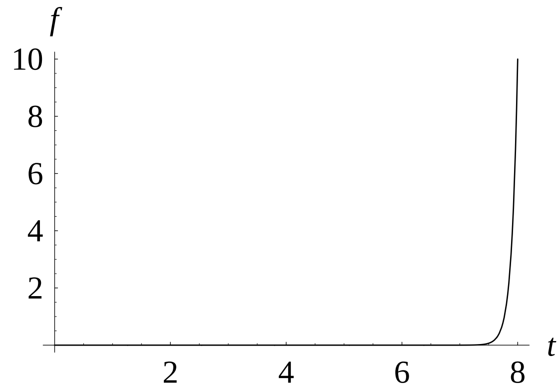
$\lambda = 10$
$L_W = 4$
$S = 5$
$E(t^{trunc}) = 3.5$
$E(t^{trunc}|t > 0) = 3.5$
$p_0 = 0$



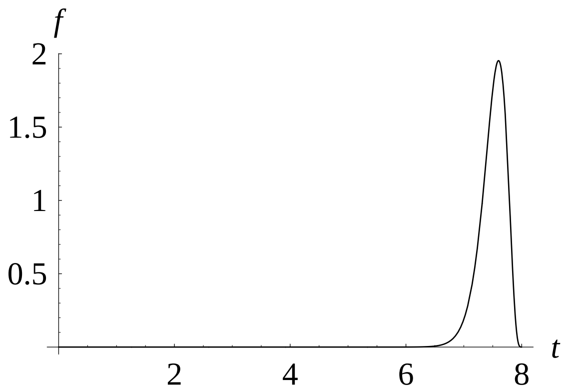
$\lambda = 10$
$L_W = 4$
$S = 10$
$E(t^{trunc}) = 3.0$
$E(t^{trunc}|t > 0) = 3.0$
$p_0 = 0$



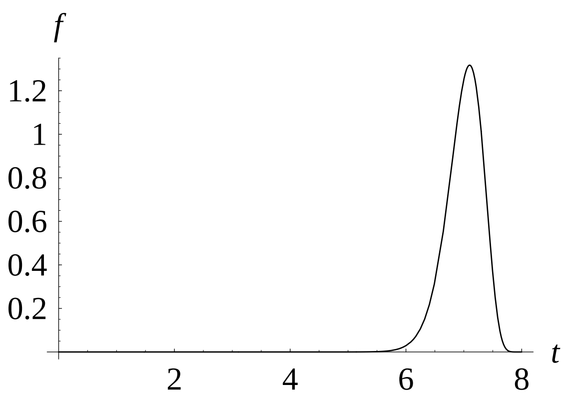
$\lambda = 10$
$L_W = 4$
$S = 20$
$E(t^{trunc}) = 2.0$
$E(t^{trunc}|t > 0) = 2.0$
$p_0 = 0.00018$



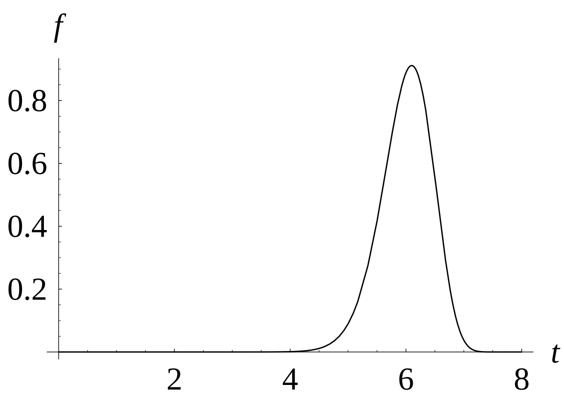
$\lambda = 10$
$L_W = 8$
$S = 1$
$E(t^{trunc}) = 7.9$
$E(t^{trunc}|t > 0) = 7.9$
$p_0 = 0$



$\lambda = 10$
$L_W = 8$
$S = 5$
$E(t^{trunc}) = 7.5$
$E(t^{trunc}|t > 0) = 7.5$
$p_0 = 0$



$\lambda = 10$
$L_W = 8$
$S = 10$
$E(t^{trunc}) = 7.0$
$E(t^{trunc}|t > 0) = 7.0$
$p_0 = 0$



$\lambda = 10$
$L_W = 8$
$S = 20$
$E(t^{trunc}) = 6.0$
$E(t^{trunc}|t > 0) = 6.0$
$p_0 = 0$

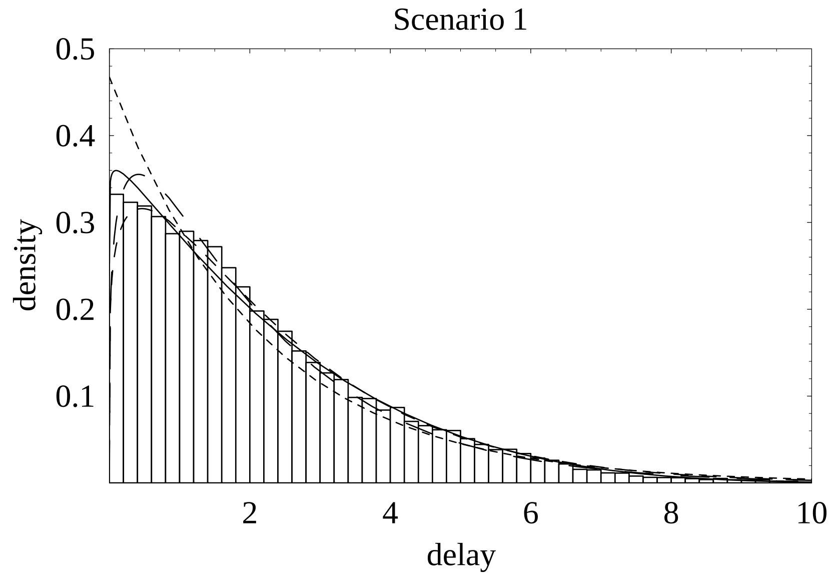
Appendix B Simulated and fitted delay densities

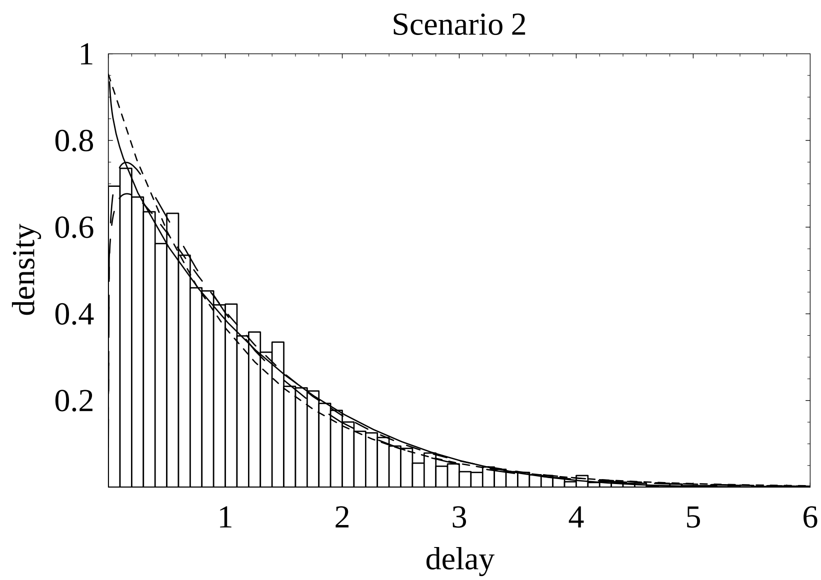
In this appendix, 15 figures will be presented corresponding to the 15 scenarios. Each graph contains the histogram of the observed delays from the simulation of the N-level distribution system. The histogram is normalised so that the mass is unity. Fitted statistical density functions are also depicted in each figure.

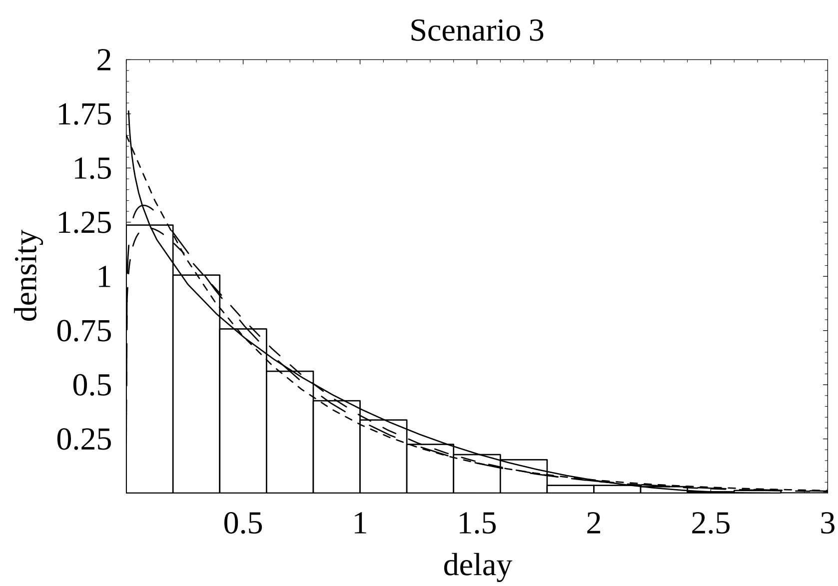
The legend for the density functions is:

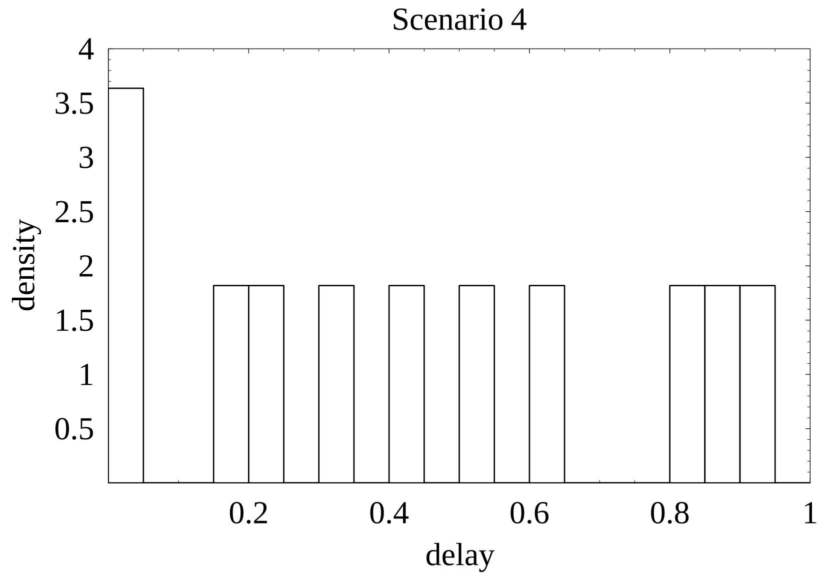
———	Beta density function
— — —	Gamma density function
– – – – –	Weibull density function
- - - - - -	Exponential density function

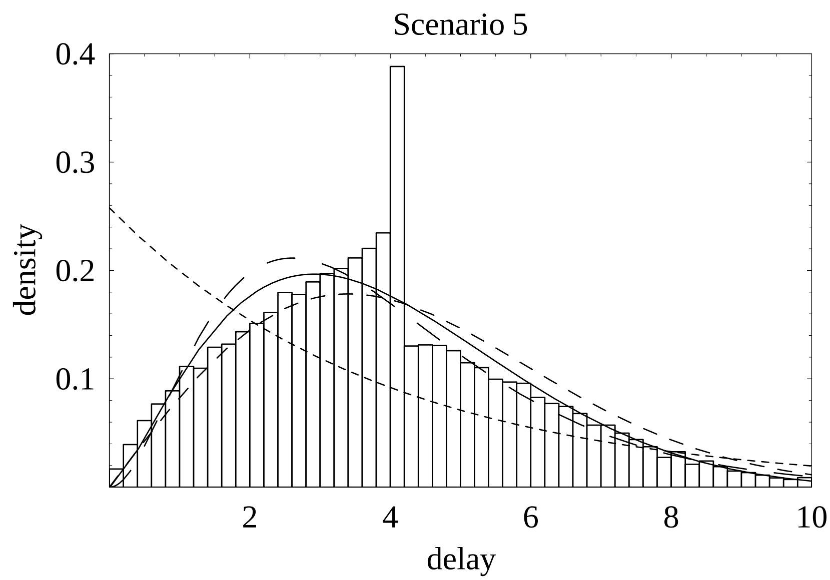
The graph of scenario 4 does not contain any density functions, since there is only 11 observations and they seem uniformly distributed. The graph of scenario 8 does not contain the gamma density function since it fits poorly.

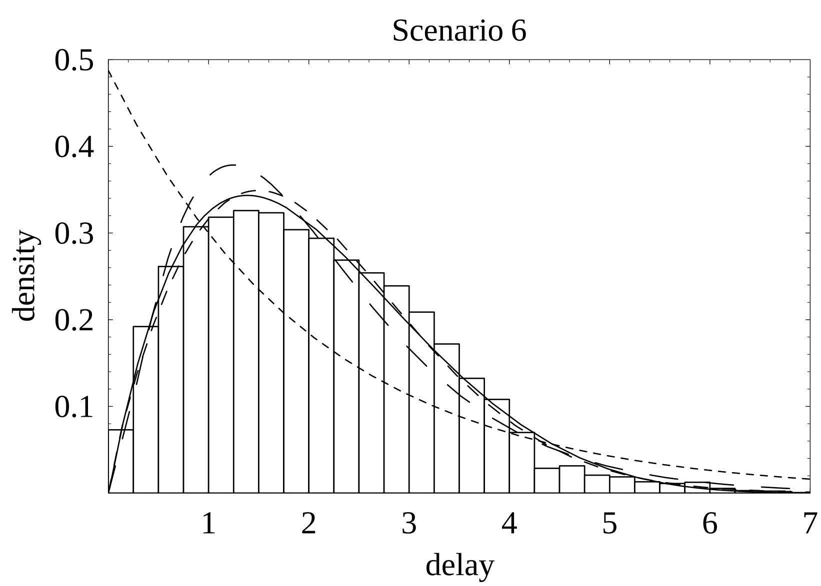

Scenario 1
0.5
0.4
0.3
0.2
0.1
density
2
4
6
8
10
delay


Scenario 2
1
0.8
0.6
0.4
0.2
density
1
2
3
4
5
6
delay

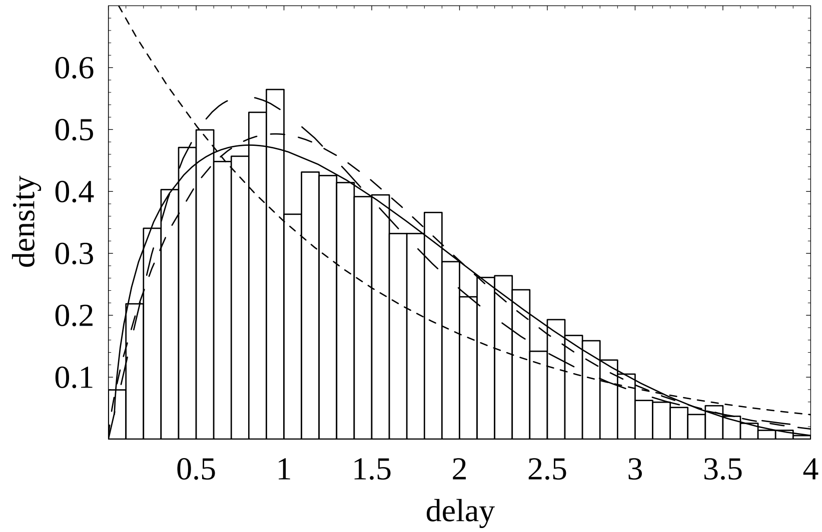

Scenario 3
2
1.75
1.5
1.25
1
0.75
0.5
0.25
density
0.5
1
1.5
2
2.5
3
delay


Scenario 4
density
delay
0.5
1
1.5
2
2.5
3
3.5
4
0.2
0.4
0.6
0.8
1

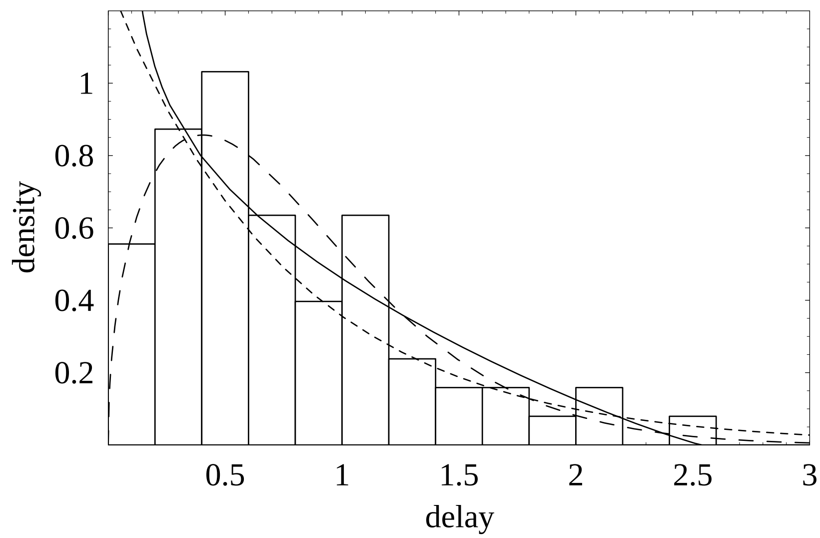

Scenario 5
density
delay
0.1
0.2
0.3
0.4
2
4
6
8
10


Scenario 6
density
delay
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7

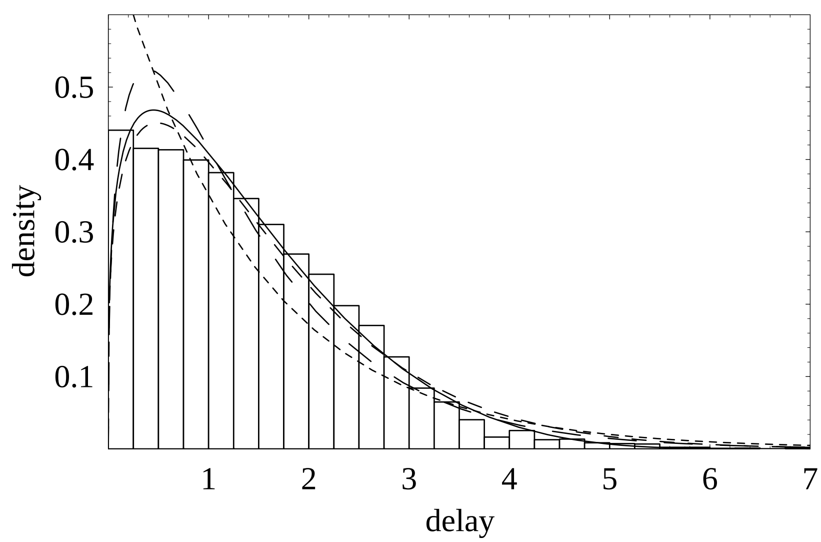
Scenario 7



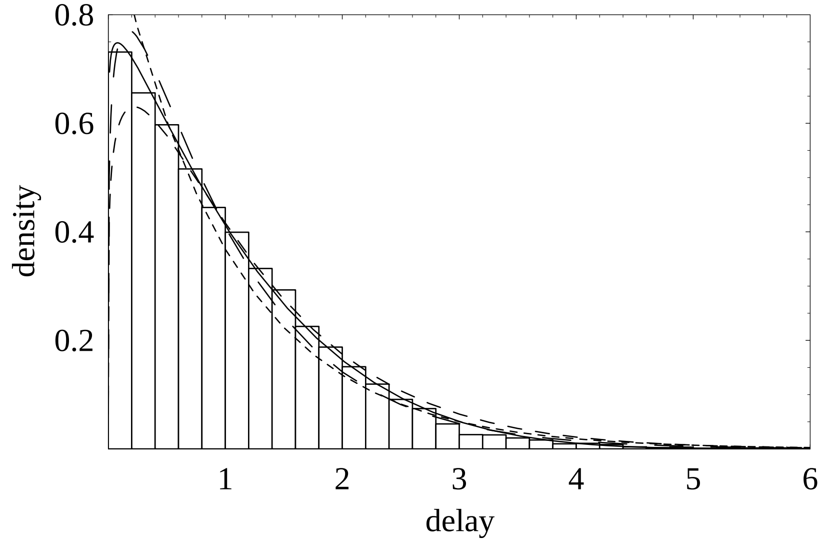
Scenario 8



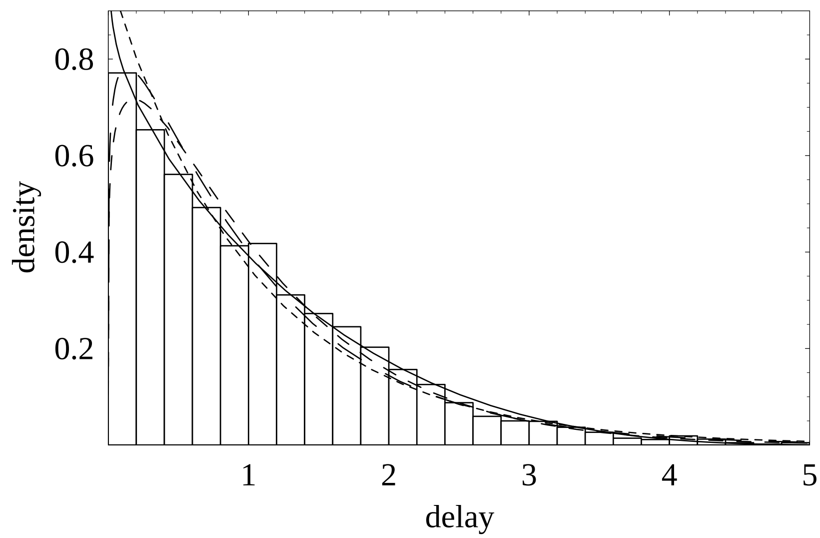
Scenario 9



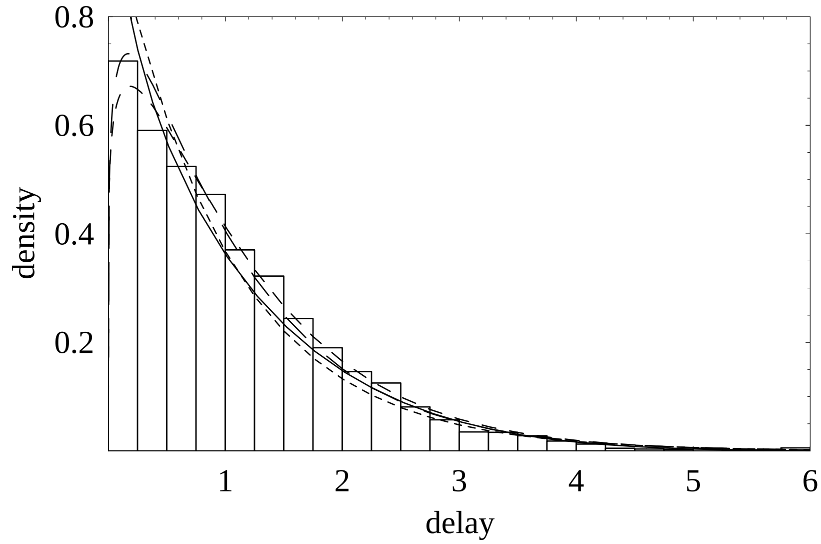
Scenario 10

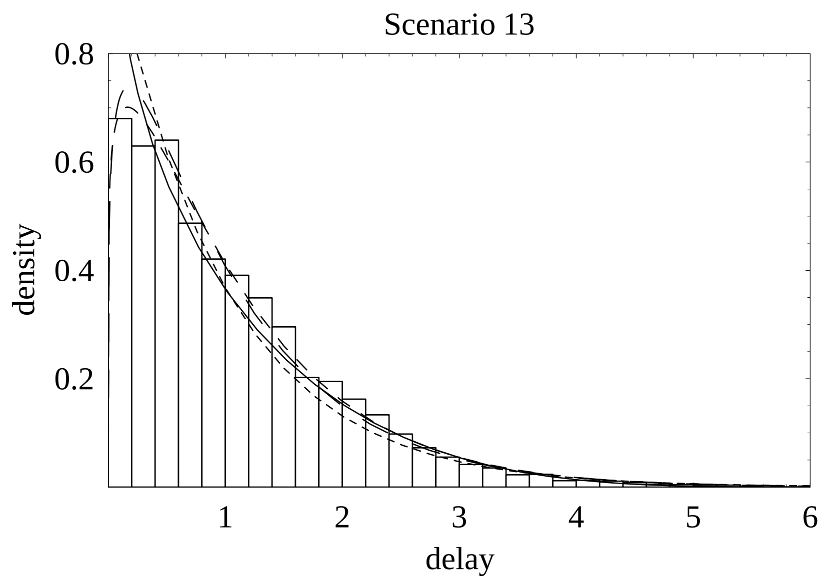


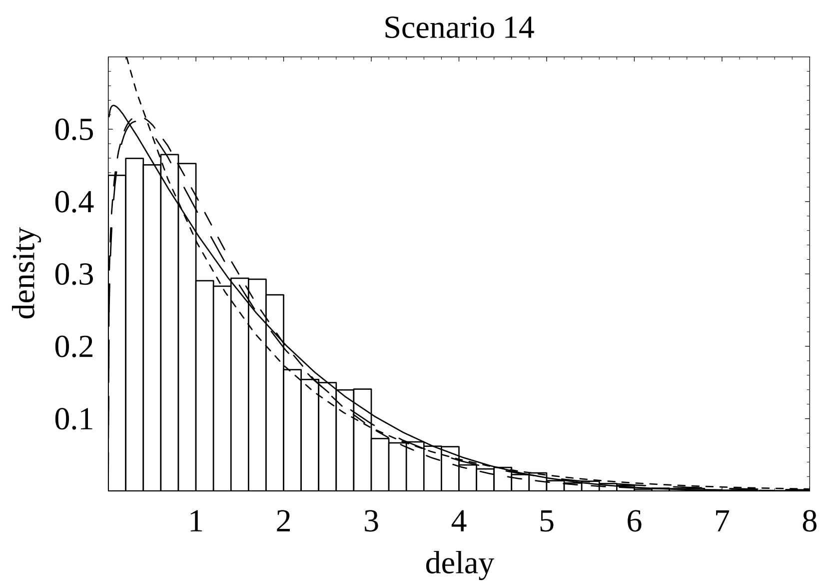
Scenario 11

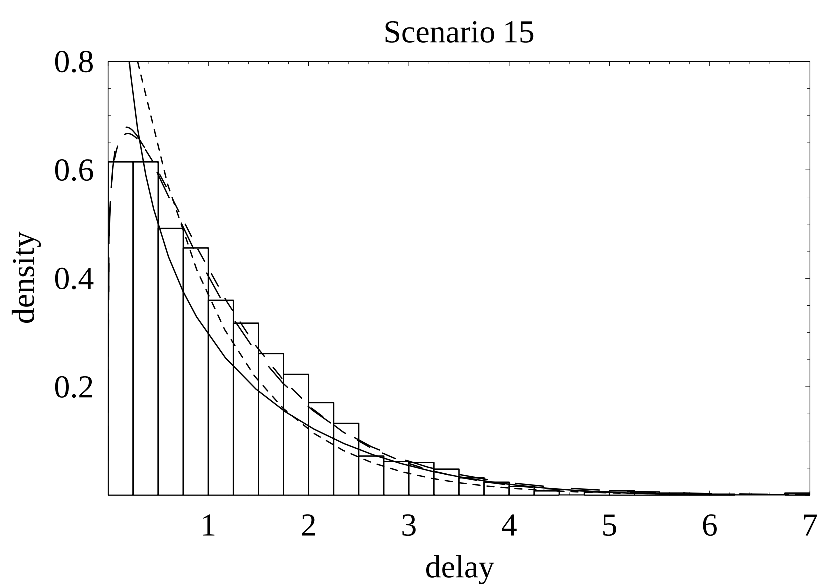


Scenario 12




Scenario 13
density
0.8
0.6
0.4
0.2
1
2
3
4
5
6
delay


Scenario 14
density
0.5
0.4
0.3
0.2
0.1
1
2
3
4
5
6
7
8
delay


Scenario 15
density
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
delay