Measuring the effect of the lead-time on the target level performances in a finite horizon

Erasmus School of Economics
Bachelor Thesis [programme Econometrics and Operational Research]
Aletta Heemskerk
Student number: 431309
Name supervisor: Eruguz Colak, A.S.
Name Second assessor: Oosterom, C.D. van

July 3, 2018

Abstract

The research focuses on the fill rate and waiting time performance measure in the finite horizon. The two measures will be investigated in a periodic review order-up-to policy. Investigating these performance measures in a finite horizon is relevant because suppliers have contracts regards the supply of goods and obtaining a certain service level for supplying these goods. This research will first verify the results of Thomas (2005) regards the fill rate performance in a finite horizon. Afterwards the focus will shift towards what implication a deterministic positive lead-time has on these results. On top of that, the waiting time of customers is analyzed as a second performance level. The research shows that most results found by Thomas (2005) still hold to an extent. In the discussion it is reviewed to what extend these results hold.

Preface

In the three years of the Bachelor Econometrics and Operational Research at the Erasmus University Rotterdam, I became interested in inventory management, as specially the performance of such inventory policies. The course Quantitative Methods for Logistics and the course Simulation were one of my favorites. This BSc thesis allows me to combine these two topics, to learn more about inventory management while using the theory of simulation in practice.

First of all, I would like to thank Dr. van Oosterom for taking the time to read this research. In addition, I would like to thank my supervisor Dr. Sena Eruguz Colak for her clear guidance and useful insights, without which none of the research I conducted would have been possible. Furthermore, I am grateful to Douglas J. Thomas, for his research regarding the fill rate in a finite horizon. Last but not least I would like to thank my friends, fellow students and parents for their moral support.

Contents

1 Introduction

For a supplier stocking goods and delivering these goods to customers always has a trade off between potentially not meeting customers demand and having excess inventory. Having an excess inventory brings inventory costs but it is also definitely not desired to have an inventory below the demand since than opportunity loss increases. By opportunity loss is meant that you could have sold more goods if there was more inventory present at that certain time or satisfied more customers by being able to deliver goods within time. In addition, suppliers often have contracts and get penalized for not meeting the set requirements. To ensure that a certain part of the demand can be satisfied by in-stock goods a target level for the fill rate is introduced. The fraction of demand that is satisfied by in-stock goods is known as the fill rate. This research focuses on the fill rate performance in a finite horizon, in particularly focusing on the effect of the lead-time on the target level performance of for example the fill rate. Reason to review the performance over a finite horizon could be that a supplier has a contract regarding the supply of goods. Besides, the service level and efficiency of a company is often analyzed over a finite period. Therefore it is important for the management of a supply business to have a sense of what the probability is of meeting those requirements regarding the fill rate. The finite period over which the performance measures are analyzed is referred to as review horizon. First the results found by Thomas (2005) will be verified and afterwards the effect of the lead-time will be investigated.

In the research of Thomas (2005) a periodic review model in a finite horizon is studied. The study is restricted to the order-up-to policy with the assumptions of having zero lead-time, stationary demand and no ordering costs. The order-up-to policy means that at the beginning of every time period the current inventory is being reviewed and restocked up to s units. Demand that could not be satisfied by the current in-stock goods is being backlogged. The s is calculated by a formula given in Chen (2003) which is stated in Section 3. As mentioned above, the study of Thomas (2005) assumes zero lead-time. Since in practice it is hardly ever the case that the lead-time is zero an extension is being introduced that relaxes this assumption. The research question will be how the lead-time effects the probability of meeting target levels in a finite horizon. The fill rate performance and the waiting time of customers is analyzed for different deterministic lead-times. In addition to the fill rate, waiting time performance is of interest in service level agreements. It is worthwhile to investigate these two performance measures (i.e. fill rate and waiting time) are related to each other. In Section 3 a more detailed outline of the procedure is given.

The global outline of this article is as follows. Section 2 presents the related literature. Section 3 formulates the model and introduces notation. Section 4 provides the research analysis to verify the results found by Thomas (2005). Section 5 presents the set-up of the extension model as well as the corresponding numerical experiments. On top of that, an analysis regarding the performance measures is provided and insight in those results is given. In Section 6 a conclusion is drawn, in addition the limitations of the research are mentioned combined with future research directions.

2 Literature review

Different analysis on the periodic performance review system with order-up-to policy have been done. Chen (2003) defines a formula for the fill rate in a finite horizon. Thomas (2005) uses this formula to examine how the length of the finite-horizon influences the target performance of the fill rate. Furthermore, there are also papers that based their research on the results found by Thomas (2005). For example Liang et al. (2013) investigate what an appropriate penalize function to penalize underperformance in a periodic review system with an order-up-to policy would be, including a positive constant lead-time. Evidence is found that it is preferable that penalties should be proportional to the underperformance. The research also shows that a supplier benefits less from the chosen strategic behavior in case of a positive lead-time. A possible cause that is named is the information delay that is present, a supplier cannot respond to the performance history as effectively as in the case of having a zero lead-time. In addition Katok et al. (2008) consider fill rate performance in a finite horizon, using a static periodic-review base-stock model with zero lead-time. It is found that overall longer period performance reviews are more effective in inducing higher stocking levels. Which is related to one of the statements made by Thomas(2005). Thomas (2005) states that for smaller horizons it is more likely to achieve very high or very low fill rates. Such that by increasing the horizon length the chance of deviating substantially from the average becomes less likely. Thomas (2005) also shows that by increasing the horizon length the required inventory level increases initially, yet after a while it decreases to a lower level. On top of that, Katok et al. (2008) notice is that for higher target levels it takes longer to recover from a large demand realization.

In the extension, the focus will be on the deterministic lead-time as well as the performance measure waiting time. Chopra et al. (2004) focus their research on how lead-time uncertainty influences the safety stock level. They investigate that if a business is interested in lowering their inventory levels while maintaining their service levels, it should focus on reducing the lead-time rather than the lead-time variability. This result holds for a business who has a fill rate target between 97 and 99% or a cycle service level target between 50 and 70%. Chopra et al. (2004) assume an indivisible period of analysis and a gamma distributed lead-time with varying parameters. In addition, some research has been conducted regarding the waiting time performance measure. Houtum et al. (2015) found that maintaining a certain aggregate mean waiting time is closely related to having a target level on the total amount of backorder demands.

3 Model and notation

In this section, the inventory model is presented, as well as the inventory model that is used for the extension. In addition, some notation is introduced.

The study focuses on a periodic evaluation performance system with an order-up-to policy. This model assumes that there is stochastic demand (which is i.i.d. in each period), no reordering costs. Demand is assumed to be Erlang(k,1) distributed what implies that on average k customers arrive per period. The values for parameter k that are considered in this research are 1, 3, 5 and 9. In addition, to verify the result found by Thomas (2005) zero lead-time is assumed. In the extension part of this research the lead-time is assumed to be deterministic and non zero.

The order-up-to policy is combined with a periodic review period. For simplicity the same notation as in the research of Thomas (2005) is used. The fill rate random variable for T periods is given in Equation 1 were X_t is defined as the demand random variable of period t. Besides Y_t represents the filled demand which is equal to the minimum of s and X_t were the level of the inventory is restocked up to s. The definition of the expected fill rate is given in Equation 2 for a infinite horizon.

$$\alpha_T(s) = \frac{Y_1 + ... + Y_T}{X_1 + ... + X_T} \tag{1}$$

$$\lim_{T\to\infty} E[\alpha_T(s)] = \frac{E(Y)}{E(X)} \tag{2}$$

The performance measure waiting time is defined as the number of periods which elapse between the receipt of the customer's order and the delivery of the goods. There are two types of average waiting time. The first one includes the customer that is served directly and the second option is the waiting time given that the customer could not be served directly. It is chosen to display the waiting time including customers that were served directly. The reason being that in a service contract it is usually stated that a customer has the right to receive his goods within a certain time. The way it is calculated is the following. If a customers could not be served directly the arrival time of this customers is recorded. Then when the system is able to serve this customer, the waiting time is calculated. In the end, the waiting times of customers are added and this is divided by the number of customers.

The long-run order-up-to level which is required to obtain a certain fill rate is computed by using Algorithm 3. For this calculation a horizon length of 100,000 (this is considered as long-run) and a 100 replications are used.

The alternative method for calculating the long-run s is done by deriving this s from the ESC which is the number of shortages (per cycle), the interested reader is referred to Copra and Meindl (2004). The formula is given in Equation 3 were s stands for the order-up-to level that is required and $f(x)$ is density function of demand distribution during the lead-time (which is Erlang distributed in this research). Equation 4 shows the relation between the ESC and the fill rate. Q is the expected demand within a cycle.

$$ESC = \int_s^\infty (x - s) \cdot f(x), dx \tag{3}$$

$$fillrate = 1 - \frac{ESC}{Q} \tag{4}$$

3.1 The extension model

Thereafter the deterministic lead-time is implemented in the simulation. Several variables must be recorded during the simulation. The on-order inventory is defined as the number of units that have been ordered, but still have to arrive (from now on referred to as the OO-inventory). The on-hand inventory (OH-inventory) which is equal to the number of units that are at that point ready to serve the demand. Backorder is defined as the total amount of demand that has not been satisfied yet. If new goods become available, the backorder demand is first satisfied. With use of these variables the inventory level (IL) is set equal to the on-hand inventory minus the backorder and the inventory position (IP) is set equal to the on-order inventory plus the inventory level. How to update these variables in the simulation can be found in algorithm 4 and Equations 5 - 8. In Section 5 the decisions regards the extension are explained.

$$OH_t = max(0, IL_t) \quad (5)$$

$$B_t = max(0, -IL_t) \quad (6)$$

$$IL_{t+1} = IL_t + OO_{t+1-L} - d_t \quad (7)$$

$$IP_{t+1} = IL_{t+1} + \sum_{i=t+2-L}^{t} OO_i \quad (8)$$

As described before, the level of inventory should be restocked up to s units, making the order quantity (Q) per period equal to s minus the inventory position. Since there is a non zero lead-time, a warm-up period is introduced. The warm-up period is determined by running the system for a series of replications and calculating the mean fill rate for the first period after the (differing) warm-up period. If the warm-up length does not effect the mean fill rate anymore, then the conclusion is drawn that warm-up period is sufficiently long. By not effecting is meant that the relative difference of the mean fill rate should not exceed 1%. This method is further explained in Section 5.1. In case of zero lead-time a warm-up period is unnecessary because it is known that at the beginning of each period the OH-inventory is equal to s. Whereas if the lead-time is for example three days, in the first three days of the simulation no replenishments arrive whereas from that day on goods could arrive every day. On top of that, is the OH-inventory at the beginning of each period not necessary equal to s.

4 Verifying the Results of Thomas (2005)

In this section, the results found by Thomas (2005) are compared to the results found in this research. An analysis of the similarities and differences is done. Furthermore, is insight given into why some results may differ.

To determine the long-run s Algorithm 3 is used with horizon length of 100,000 and 100 replication. Thomas (2005) does not specify a long-run horizon length. The long-run s was obtained with two decimal precision. After obtaining the long-run s, it is checked if the performance of this long-run s in a finite horizon are the same as Thomas (2005) finds. It is checked by using 1,000,000 replications. There appears to be a small difference between the results found by Thomas (2005) and the results found in this research. For instance in the case with a horizon length of $T = 20$ and demand being Erlang(1) distributed, the algorithm finds an s equal to 3.00 which is presented in the last row of Table 1. Thomas (2005) provides a long-run value of s rounded by two decimals. To check whether these results are in line with the results found by Thomas (2005) a 95% confidence interval is used. The confidence intervals can be found in the last column of Table 1. The mean that is found by Thomas (2005) in this case is equal to 0.9567 which is outside the found confidence interval. A possible explanation can be that rounding s to two decimal precision is not sufficient to obtain the same results. To check whether this is indeed the case, the long-run value of s for each case was computed with a three decimal precision. Table 1 provides the distribution characteristics when using a three decimal precision for s and Figure 2 represents the corresponding distribution plot.It can be concluded that by increasing the number of digits of the long-run s the results become closer to the results of Thomas (2005). For the case previously mentioned, a s equal to 2.996 is obtained and for this long-run value of s the mean found by Thomas (2005) is inside the confidence interval. Concluding that by increasing the number of digits even further the results will become even closer to the results found in Thomas (2005), however these results are already within the confidence interval.

An alternative method for calculating the long-run s is to derive this s from the ESC, which is the number of shortages as described in Section 3. Algorithm 1 is used to determine the long-run s. To compute the distribution characteristics corresponding with these s values Algorithm 2 is used. By allowing a gap of 0.0001 in Algorithm 1 and using 1,000,000 replications in Algorithm 2. The results can be found in Table 4. The results found by this method are also in line with the results found by Thomas (2005).

Furthermore in Figure 3 the distribution of the mean fill rate with different arrival intensities is shown. The obtained results correspond with the results found by Thomas (2005). In addition the results provided in Figure 2 also correspond with Thomas (2005). Thomas (2005) states: "Here we observe, not surprisingly, that with very short horizons, very high and very low realized fill rates are more likely". The standard deviation in Table 1 and the plots in Figure 2 help to support this claim. By increasing the horizon length, deviation from the average becomes less likely. When examining Figure 2, the conclusion is drawn that the results are in line Thomas (2005).

Table 1: Descriptive statistics for distributions shown in Figure 1 and Figure 2. By using a three decimal long-run s.

s	m-erlang	fill rate %	Length (T)	mean	std. dev.	median	skewness	CI 95%
7.262	5	95	5	0.9577	0.0585	0.9890	-1.5886	[0.9576 0.9578]
7.262	5	95	10	0.9541	0.0449	0.9647	-1.0968	[0.9540 0.9542]
7.262	5	95	20	0.9521	0.0276	0.9543	-0.5948	[0.9513 0.9515]
7.262	5	95	100	0.9504	0.0153	0.9512	-0.3248	[0.9503 0.9505]
2.996	1	95	20	0.9568	0.0549	0.97901	-1.5452	[0.9567 0.9569]
5.188	3	95	20	0.9531	0.0388	0.9601	-0.9348	[0.9530 0.9532]
7.262	5	95	20	0.9521	0.0331	0.9569	-0.7526	[0.9520 0.9522]
11.284	9	95	20	0.9514	0.0276	0.9543	-0.5948	[0.9513 0.9515]
3.00	1	95	20	0.9574	0.0543	0.9794	-1.5516	[0.9573 0.9575]

Minimum required order-up-to level

To find the minimum order-up-to level for which a fill rate of 95% or 99% is obtained Algorithm 5 is used. The required minimum stock level are obtained for horizon lengths between 1 and 100. The horizon step length to obtain s for each horizon length was 5. A different step size than Thomas (2005) is chosen to obtain a more precise and detailed plot, that was based on more data points. After examining those results, there were relatively high changes of s for small horizon lengths (T between 1 and 25). Therefore, the step size between those horizon lengths was set smaller to a step size of 1. Decreasing the step lengths for these intervals resulted in the Figures shown in 4 and 5.

The Figures 4 and 5 are less smooth than the results provided by Thomas (2005). A possible cause could be that the number of replicates was not sufficient to create a smooth line (100,000 replicates are used and a gap of 0.05 was allowed). However, from the figures provided in the appendix the same conclusion as Thomas (2005) can be drawn, namely: that small horizon lengths benefit from potentially avoiding large demand realizations is greater for higher variability demand. Such that the order-up-to level is lower for small horizon lengths.

To examine if the roughness in the lines could be explained by an insufficient number of replicates, the number of replicates for the Erlang(5) with a fill rate target of 95% was increased to 1,000,000. The results are shown in Figure 6, by increasing the number of replications the lines become smoother. However it is still not entirely smooth, by reducing the allowed gap between the lower- and upperbound and increasing the replications to 10,000,000. Expecting that the figures will be even more similar. A specially because most of the bumps appear for longer horizon lengths. This could be due to the gap of 0.05 that is allowed. If the actual value of two different horizon lengths is smaller than 0.05, then those two horizons would probably end up with the same order-up-to level even though the actual required order-up-to levels differ slightly.

5 Numerical experiments

In this section, the research executes numerical experiments in order to test the performance of the target level measures. The performance is compared for different lead-times. First, the long-run value of s is computed for the different positive lead-times with an Erlang(5) demand distribution. Here the performance measures are analyzed when using this long-run s in a finite horizon setting. Secondly, these performances are analyzed for varying horizon lengths, to investigate the effect of a positive lead-time on the order-up-to level over varying horizon lengths. In third, the effect of a positive lead-time in combination with varying $Erlang(k, 1)$ is investigated for differing horizon lengths. Besides the set-up for the extension given in Section 5.1.

5.1 Set-up of the extension

The effect of having a positive lead-time and its effect on the performance measures is examined. The lead-time lengths that will be considered are: 1, 2, 3 and 4. By use of simulation and a bisection algorithm the lowest value for s for which we can guarantee in $x\%$ of the cases a 95% fill rate performance is searched. The algorithm works in the following way and is presented by Algorithm 5. The function that is called inside this algorithm is the simulation that calculates the performance measure given: the order-up-to level (s), horizon length, $Erlang(k, 1)$. The algorithm uses a bisection method. If the target level is not met for a given value of T and s. It divides the interval in two and checks again if this new s will meet the requirements. The algorithm allows a gap between the lower- and the upperbound. Once the gap becomes sufficiently small, the order-up-to level is saved and the waiting time of customers is examined given this s.

As described in Section 3 the average of the waiting time is calculated. The clarification of the calculation of the waiting time is the following. Since there is a warm-up period prior to the review horizon, it can occur that there are customers who are already in line before the review horizon starts. The way this is dealt with is the following. Only customers who arrive in the review horizon are taken into account when calculating the waiting time average. In the same way it can occur that at the end of the review horizon there are still customers waiting for their order. For those customers the after period is introduced. The after period is an extension at the end of the simulation, such that it is possible to know when those customers eventually get served (and there waiting time can be calculated).

Before simulating the warm-up length has to be determined. This was already briefly discussed in Section 3. The simulation length that is a 100 periods, besides the number of replications is set to 1,000,000. The mean fill rate for the first period within the review horizon is plotted. This is done for different warm-up length to examined when the mean fill rate remains approximately constant. The results are shown in Figure 7. In this Figure there is clearly shown that longer lead-times require a longer warm-up period. For instance if the lead-time is 1 the fill rate remains relatively constant for warm-up periods equal to 40 periods or more. However for a lead-time of 4, 40 periods does not seem sufficient. Furthermore, Figure 8 shows for different order-up-to levels the mean fill rate over differing warm-up lengths (given a lead-time of 2 periods). This figure shows that the relative difference in the mean fill rate does partly depend on the chosen order-up-to level. For higher order-up-to levels the difference in the mean fill rate is smaller. An explanation can be that for higher order-up-to levels it is less likely to have a waiting list at the beginning of a period which influences the fill rate performance. In case of low order-up-to levels, not including a warm-up period will result in absence of a waiting line at the start of the simulation. Introducing a warm-up period increases the likelihood of having a waiting line at the start of the review horizon that influences the performance measures.

In the end, it is decided that the warm-up period for a lead-time of one period should be 40 periods, for a lead-time of two 50 periods. The warm-up period for a lead-time of three should be 60 and lastly that the warm-up period for a lead-time of four is 70. This decision was based on examining the plots for different order-up-to levels and the benchmark that was explained in Section 3.1. Namely that the relative difference of the mean fill rate should not exceed 1%.

Since the research is restricted by time constraints, the precision of the extension part is decreased. There are a few options, for example to decrease the number of replications or allow a bigger gap in the bisection method. The Erlang(9) with a lead-time of 2 periods is plotted, for various gaps and number of replications. This shows that by decreasing the number of replications, there appeared more variation in the minimum required order-up-to than is preferred. Therefore the decision is made to still accept a gap of 0.05 and to use 100,000 replications. However, because Section 4 shows that the results of this research are in line with the result of Thomas (2005), the step size is changed to the step size that Thomas (2005) uses. The step size is changed due to time constraints, by changing the step size to the step size Thomas (2005) uses the number of order-up-to levels that have to be computed decreases from 40 to 20.

5.2 Case 1: Effect of lead-time on performance measures

For this Section the focus will be how the two target measures perform over different lead-time lengths. The distribution characteristics that will be investigated are: the standard deviation, mean, skewness, the required order-up-to level and lastly the minimum and maximum waiting time of customers. First the long-run s for the varying lead-times is determined for which demand is Erlang(5) distributed. This is done by using the about the same Algorithm 3 as for verifying the results found by Thomas (2005). The structure of the algorithm is kept the same, however line 6 to line 13 are replaced by the extension Algorithm 4. Such that the algorithm can calculate a long-run s for positive lead-times. Next, the focus shifts towards examining this long-run s in a finite horizon and its distribution characteristics within this finite horizon.

The long-run s values that the algorithm finds for the lead-times 0, 1, 2, 3 and 4 are respectively $s = 7.262$, $s = 13.759$, $s = 19.920$, $s = 25.902$ and $s = 31.770$. In Table 2 the distribution characteristics are shown for the fill rate at a given horizon length of 20 periods and with demand Erlang(5) distributed. Table 3 provides the results of the waiting time measure for the same setting as mentioned above. Both tables show that by increasing the lead-time length the uncertainty of both measurements increases. This can be seen by the standard deviation that increases for longer lead-times. A larger standard deviation indicates that the obtained results are more spread. For instance if the lead-time is zero the standard deviation of the fill rate is 0.0330 and if the lead-time is 4 periods the standard deviation is 0.07223. Which is roughly twice as large.

Secondly, The amount of stock that is required increases if the lead-time increases. Which can be explained by the fact that if you order goods at period t these goods do not arrive until the beginning of period $t + L$. In the mean time a business still has to maintain there target levels and therefore has to increase it safety stock, meaning there is a delay in replenishment. This can be explained by the same kind of information delay that was also mentioned in Liang et al. (2013).

Table 2: The fill rate descriptive statistics for which a fill rate of 95% is obtained on average in the long-run. Demand is assumed to be Erlang(5) distributed and the review horizon is equal to $T = 20$.

lead-time	s	mean	std. dev.	skewness
0	7.262	0.9520	0.0330	-0.7477
1	13.759	0.9530	0.0470	-1.2796
2	19.920	0.9536	0.0572	-1.6383
3	25.902	0.9541	0.0652	-1.9062
4	31.770	0.9543	0.0723	-2.1385

Table 3: The waiting time descriptive statistics for which a fill rate of 95% is obtained on average in the long-run. Demand is assumed to be Erlang(5) distributed and the review horizon is equal to $T = 20$. The minimum and maximum that are provided is the absolute minimum and maximum over all replications.

lead-time	mean	std. dev.	skewness	[min,max]	long-run mean
1	0.0475	0.0480	1.3618	[0,2]	0.0505
2	0.0479	0.0609	1.8381	[0,3]	0.0518
3	0.0487	0.0729	2.2512	[0,4]	0.0531
4	0.0496	0.0841	2.6085	[0,5]	0.0544

Third, even though the oder-up-to level increases for longer lead-times (such that the required fill rate level is still achieved on average), the average waiting time increases. By further investigating this appears to be due by large demand realizations. If the lead-time is longer the quantity that compensates for the large demand realization arrives later, making the system less efficient. As a result, customers may wait longer. This observation is also supported by examining the maximum waiting time. For instance if the lead-time is equal to 1 period in the 100,000 replications that were done, the longest a customers had to wait was 2 periods. Whereas if the lead-time is 4 periods there appeared a waiting time of 4 periods. The maximum length of waiting time can be related to the length of the lead-time. If a customer can not be served at time t, than a order is placed at the beginning of period $t+1$. This order will arrive at $t + 2$ provided that the lead-time is 1 period. In the end this customer has to wait 2 periods. It makes sense that this does not occur often because this will mean that during $t + 1$ there was not enough inventory to serve the outstanding backorder let alone serve the customers arriving in period $t+1$ (since backorders are satisfied first). If this is the case there is a low fill rate obtained and the average fill rate should be equal or higher than 95%. Meaning this cannot occur often. The maximum number of periods a customers has to wait becomes equal to $1 + L$.

In fourth, when examining Figure 9 there are a few things that stand out. The first thing is the spread of the fill rate distribution. For instance if the lead-time is 4 periods there are fill rates obtained of 75%. Whereas if the lead-time is zero this does not occur. Also lower fill rates (fill rates below the 95% benchmark) are obtained more frequently. On the other hand the number of times the fill rate was equal to 1 increases (meaning every customer was served directly). For instance achieving a fill rate of a 100% given that the lead-time is zero, is roughly around 4%. This in comparison to 42% chance if the lead-time is 4 periods (using interval width of 0.01). This is related to the first statement of this case that for longer lead-time the uncertainty in achieving a required target fill rate level increases, such that deviating from the average becomes more likely.

Moreover, it is clear that the waiting time and the fill rate have some correlation. Namely, if a high fill rate target is set it becomes less likely that a customer has to wait. Not only the chance of ending up in the waiting line decreases, but also the average waiting time of customers that have to wait decreases. Because if there is a high chance of serving a customer directly for every period (by having a high enough up-to-order level). It means that the backorder also has to be served quickly otherwise you can not serve new customers directly.

Although, Table 2 shows that by increasing the lead-time the fill rate average increases. While if Table 3 is examined, the average waiting time also increases. This indicates that it cannot be said with certainty that by maintaining a particular fill rate performance a certain waiting time performance is also obtained.

Lastly, the mean fill rate in the finite horizon is higher than the fill rate that is obtained in the long-run (0.95 for each case). Chen (2003) proves that the expected fill rate of the finite horizon is greater of equal than the expected fill rate of the infinite horizon given that the lead-time is zero. To check whether this result also holds for positive lead-time a simulation is used. In Figure 10 the average fill rate is provided for different horizon lengths using the long-run s, this was done for 10,000,000 replications. The figure shows that the mean fill rate for horizon lengths between 1 and a 100 periods is higher than the 0.95 that is obtained in the long-run. When comparing the average waiting time in a finite horizon with the case of an infinite horizon, it shows that for the finite horizon the average is lower. Which can be partly explained by the fact that the fill rate (fraction of customers that is served directly) is higher for the finite case. The same was executed for horizon lengths between 1 and 100,000 periods given that the lead-time was 4 periods. However, the figure did not provide additional insight.

5.3 Case 2: Lead-time Behavior over varying Horizon Lengths

In this case, the research examines the minimum order-up-to level for which a fill rate of 95% is obtained for a certain percentage of the replications. First it is assumed that demand is Erlang(5) distributed and that the lead-time varies between 0 and 4 periods. The algorithm that this case uses to find the minimum order-up-to level is the same as used for verifying the minimum order-up-to level obtained by Thomas (2005), namely Algorithm 5. The number of replications is equal to 100,000 and the gap that is allowed between the lower- and upperbound is 0.05.

The first thing to notice when examining the behavior of the order-up-to levels in Figure 11 is that (particularly for the line that represents the 99% certainty that a 95% fill rate is obtained) the lines become more flat when the lead-time is longer. A possible explanation is that if the lead-time is 0 and there is a short horizon length (for instance $T = 1$) there is a high chance of avoiding a large demand realization making it more likely to meet the target level. Whereas if the lead-time is 4 periods, a large demand realization before the review horizon starts (so that it occurs at the end of the warm-up horizon) can still affect the target level in the review horizon. For longer lead-times it takes longer to recover from a large demand realization since the system becomes less efficient. The system becomes less efficient because the goods that compensates for the large demand realization arrive later for longer lead-times, this longer recovery will reflect in the fill rate performance. Therefore, it becomes less likely for long lead-times to totally avoid a large demand realization even for small horizons. When the lead-time is 0 the OH inventory at the beginning of each period is equal to the order-up-to level. This is not the case for positive lead-times. If a large demand realization occurs at time t, than this can effect the number of customers that can be served at time $t + 1$. How long this large demand realization is effecting other periods depends on the length of the lead-time as explained in Case 1.

It seems that for a fill rate of 95% that is guaranteed to be obtained in 99% of the cases, might convert slowly towards a straight line. If that is indeed the case then, there would be no difference between a short and a long review horizon. To test whether this is true, the minimum order-up-to levels is computed for a lead-time of 10 periods, which is also shown in Figure 11.

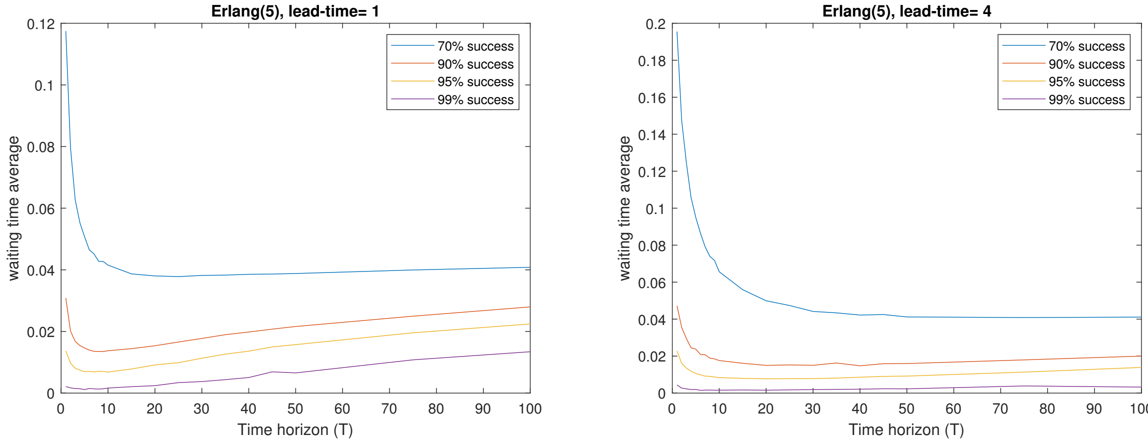
However, when examining the lead-time of 10 periods there is still a difference between the short horizon and the long horizons. This raise the suspicion that it will always eventually be the case that by increasing the horizon length the required order-up-to level will be lower. For shorter lead-times it appears that the order-up-to level decreases earlier and more rapidly. From the figure right below in Figure 11 with a lead-time of 10 periods it is not entirely clear whether the order-up-to level decreases at the end or not. However, if the horizon is increased length to $T = 200$ the required order-up-to level indeed decreased (shown in Figure 12).

Furthermore, it seems to be the case that for small lead-time (0 and 1 period) the relative difference in the minimum order-up-to level decreases. Such that the difference in required order-up-to level is smaller for differing horizons. For instance taking the case for which you can guarantee a 95% fill rate in 95% of the cases. If the lead-time is 4 periods the maximum level of order-up-to levels is around 36.59 and the lowest order-up-to level is around 33.38, the relative difference is 1.096. If the lead-time is 1 the highest order-up-to level is about 23.88 and the lowest is around 21.59 with a relative difference of 1.106, which is slightly larger. This again shows that for longer lead-times you benefit less from having a short review horizon.

Next the research analyzes the the waiting time performance measure for the minimum order-up-to level that has to be present to require a certain fill rate performance. Figure 1 shows the average waiting time matching the minimum order-up-to level of Figure 11 for 1,000,000 replications. A 1,000,000 replications were used because when computing the same for a 100,000 replications the figure was less smooth than was preferred.

The first observation that is made is that the average waiting time is somewhat correlated to the minimum order-up-to level. For short horizons the required order-up-to level is lower because it has to potential benefit from avoiding a large demand realization. A consequence is that by lowering this order-up-to level it becomes more likely that customers have to wait (or have to wait for longer). The same relation is shown for long horizons (around 100 periods), longer review horizons have more time to recover from a large demand realization as mentioned by Thomas (2005). Therefore,

Figure 1: The figure presents the waiting time average over varying horizon lengths with the corresponding minimum order-up-to level which was found in Case 2.



the required order-up-to level can be lower, although this does increase the average waiting time.

Another observation that can be made by examining Figure 1, is that for longer lead-time the average waiting time is higher. Besides, it does show that for higher fill rate targets the average waiting time is lower. This can be seen by the fact that in case were the requirement is to hit the fill rate target 70%, the average waiting time is higher than in the case were the requirement is 90%. The figure overall shows that maintaining a specific fill rate target over a finite horizon does not give you a guarantee that you will obtain a certain waiting average. Because, even though the fill rate remains constant over the differing horizon lengths the waiting time average does not.

5.4 Case 3: Lead-time behavior for varying Erlang distributions

This case examines the fill rate performance of 95% for which we can guarantee either 99% or 95% of the cases for varying Erlang demand distributions. This is done by using the same simulation as in Case 2. The 90% and the 70% are not included in this case due to time constraints. The reason being that the overall behavior can still be examined by just examining the 99% and the 95% case. Also, the preference was given to examine those two cases with more precision rather than examining all four cases with decreased precision. The main reason to investigate the lead-time behavior under different Erlang distributions is that the uncertainty in demand increases. A statement that is made by Thomas (2005) is that short-horizon benefit from potentially avoiding large demand realizations is greater for higher variability demand. This case will investigate if this statement is still valid for positive lead-times.

The $Erlang(k, 1)$ distribution that this research considers are the same as Thomas (2005) namely: $k = 1$, $k = 3$, $k = 5$ and $k = 9$. First, the long-run value of s for these Erlang distribution is determined by using again Algorithm 3 with the modifications described in Section 5.2.

The long-run values for s are in Table 5. The long-run values that are found relate to the findings of Case 1, by increasing the lead-time the required order-up-to level increases. Figure 13 shows the required order-up-to level for varying lead-time lengths, given that demand is Erlang(1) distributed. Figure 14 shows the same but for Erlang(3) and Figure 15 for Erlang(9).

The figures show overall the same behavior. For longer lead-time, short review horizons benefit less, due to the fact that avoiding a large demand realization becomes less likely as explained in Case 2. Besides, it shows that the point at which lengthening the horizon reduces the necessary stock level occurs much later for longer lead-times. The benefit from a long review horizon was that the system is not doomed by one single large demand realization. If the lead-time increases, the recovery time is longer due to the delay in replenishment mentioned before. It therefore makes sense that the system takes longer to recover and the necessary order-up-to level starts to decreases for longer review horizons. Katok et al. (2008) find that for higher target levels it takes longer to recover from a large demand realization. So there are two important factors that can influence the speed of the recovery, i.e. the target level itself and the lead-time.

Furthermore, it continues to be the case that the length of the review horizon significantly impact the required order-up-to level. For instance if Figure 13 were the lead-time is 3 periods. The minimum order-up-to level (meeting the 95% case) for $T = 1$ is about 7.7 and for $T = 20$ is approximately equal to 10.1.

6 Discussion

In this section a summary of the results of Thomas (2005) is given, next to the discussing of the extend to which these results till hold in cases with a positive lead-time. Afterwards, the research discusses the most important findings of the extension part.

For all cases Thomas (2005) considers, the necessary order-up-to level increases for short review horizons but eventually decreases. If the demand is relatively low in variability the order-up-to level decreases faster, due to potentially avoiding large demand realizations. In the case of having a deterministic positive lead-time this remains true. Yet, for longer lead-times the benefit from having a short review horizon decreases. As discussed in Section 5 this is partly due to the delay in replenishment that increases if the lead-time increases. The longer the lead-time, the longer recovery the system needs from a large demand realization. So, if a large demand realization occurs in the warm-up period this could still affect the performance measures in the review horizon. Making short review horizons less beneficial.

Secondly, Thomas (2005) finds that long review horizons increase the chance that large demand realizations are seen but also give the supplier more opportunity to recover from these large realizations. This statement is given to explain why the required order-up-to level eventually decreases. Again, the overall behavior remains the same, but because of this delay in replenishment the recovery is more slowly, meaning that the horizon length should be longer to be beneficial. For instance in Figure 1 of Case 2 the fill rate is shown given a lead-time of 10 periods. The required order-up-to level nearly stays constant for long review periods. Although it does decrease, the decrease occurs beyond the range of the plot. Whereas if the lead-time is zero, the decreasing of the order-up-to level starts around a review period length of 13.

Third, Chen (2003) proved that if the lead-time is 0 the expected fill rate mean in a finite horizon is higher than the one for the infinite horizon. In Case 1 Table 2 suggested that this might be the case as well for positive lead-times. To confirm these suspicions, the mean fill rate was computed for different review horizon lengths. These results show that the mean fill rate that was obtained in a finite horizon was indeed higher.

Next, the target performance waiting time of customers was added. For a supply business it is important to realize that the two target performance measures are not one-to-one related. A business cannot assure that the average waiting time will stay below a certain benchmark just by reaching a certain fill rate target. To illustrate this, Case 2 showed that by setting the order-up-to level in such a way that a 95% was obtained, did not result in a constant waiting time average. They should carefully consider the requirements that are set in these contracts and base their performance measures based on those requirements.Besides, an increasing lead-time results in an increase in the maximum number of periods a customer had to wait increases. For instance in Case 1, for a lead-time of 1 period the longest a customer had to wait was 2 periods. If we compare this, a customer could be waiting up to 4 periods if the lead-time is 4.

7 Conclusion and Future Research

This research observes the fill rate performance as well as the waiting time of customers in a finite horizon with a periodic performance review system. The motivation behind this research was to investigate what effect a deterministic lead-time had on meeting target levels that might be set by service contracts. First, the results found by Thomas (2005) were replicated. Although there were sometimes slight deviations, the conclusion could be drawn that the statements made by Thomas (2005) still hold. Secondly, it was investigated to what extent these results would hold for a non-zero deterministic lead-time. Short review horizons still benefit the supplier, as large demand realizations may not occur during the review horizon. This increases the chance of meeting both performance measure requirements. However, if the lead-time increases this advantage decreases since large demand realizations in the warm-up period can still affect the small review period.

Another observation that is made by Thomas (2005) is that long horizon length do not have this advantage. Nevertheless, the long horizon also gives the supplier the chance to recover from this large demand realization. This observation still holds for a deterministic lead-time, although a note of caution has to be made here. For longer lead-times the recovery is more slowly, meaning that the horizon length should be longer to be equally beneficial. The numerical experiments further show that the expected fill rate mean in a finite horizon is higher than in the infinite horizon.

An important consequence of having a positive deterministic lead-time will be state once more. For longer lead-time the uncertainty in the performance measures increases, making it more likely to "overshoot" but also to "undershoot" in a finite review period. A manager of a supply business needs to be aware of this uncertainty, since it becomes more likely to get penalized for not meeting the required performance measures. The same holds for higher demand variability, for a lead-time of zero higher demand variability benefit more from a short review period because it can avoid possible large demand realizations. However if the lead-time increases this benefit decreases faster than for low demand variabilities. Another important consequence of increasing the lead-time is that the maximum periods a customer has to wait increases. A manager of a supply business has to realize this consequence when setting service level contracts.

This research can be further extended by introducing a stochastic lead-time. It can be expected that the overall uncertainty increases and it will be harder for a supply business to guarantee a certain level in a finite horizon setting. Besides, the research is restricted to demand being $Erlang(k, 1)$ distributed. It would be interesting to see how these performance measure behave under different demand distributions.

References

Chen, J., D. K. Lin, D. J. Thomas. 2003. On the item fill rate for a finite horizon. Oper. Res. 31(1) 199-123

Chopra, S., Reinhardt, G., Dada, M., 2004. The Effect of Lead Time Uncertainty on Safety Stocks. Decision Science. 35(1) 1-24

Chopra, S. and Meindl, P., 2004. Supply Chain Management: Strategy, Planning and Operation, Pearson Education, Inc., Singapore, Second Edition.

van Houtum, G., B. Kranenburg. 2015. Spare Parts Inventory Control under System Availability Constraints.

Katok, E., D. Thomas, A. Davis. 2008. Inventory Service-Level Agreements as Coordination Mechanisms: The Effect of Review Periods. Manufacturing Service Oper. Management. 10(4) 609-624

Liang, L., D. Atkins. 2013. Designing Service Level Agreements for Inventory Management. Production and Operations Management. 239(2) 377-388

Thomas, D. J. 2005. Measuring item fill-rate performance in a finite horizon. Manufacturing Servervice Oper. Management. 7(1) 74–80

Tyworth, J. E., L. O'Neill. 1997. Robustness of the normal approximation of lead-time demand in a distribution setting. Naval Res. Logist. 44(1) 164-186

Appendix

Algorithm 1 The algorithm that is used to determine the long time s by mean of the ESC.

procedure S APPROXIMATIONTHEORETICAL($precision, \lambda, benchmark$)
    bigS=100;
    smallS=1;
    while $(bigS - smallS) < precision$ **do**
        $\{$
        $middle = \frac{bigS+smallS}{2}$
        ESCmiddle $= (x - s) \cdot gampdf(x, \lambda, 1)$
        fillrate $= 1 - \frac{ESCmiddle}{\lambda}$
        if fillrate $\geq$ benchmark **then**
            $\{bigS = middleS\}$
        else
            $\{smalllS = middleS\}$
        $\}$
    Return: bigS

[1]

Algorithm 2 The algorithm that is used to determine the long time s that is required to obtain a certain fill rate target.

procedure S CALCULATION($s, \lambda, numrep, horizonlength, benchmark$)
    for $replications = 1 : numrep$ **do**
        $\{$
        sumx=0
        sumy=0
        for currenttime = 1:horizonlength **do**
            $\{$ Generate X
            Y = min(s,X)
            sumx = sumx +X
            sumy = sumy +Y
            $\}$
        fillrate(replications) $= \frac{sumy}{sumx}$
        $\}$
    succesfull = (fillrate > benchmark)
    fracsucces $= \frac{sum(succesfull)}{numrep}$

[2]

Algorithm 3 The algorithm that is used to determine the long time s that is required to obtain a certain fill rate target.

```
procedure S APPROXIMATION(precision, λ, numrep, horizonlength)
    while decimal < precision do
        {
        initialize all variables
        for s = Begin : 1/10^decimal : Begin + 1 do
            {
            for replications = 1 : numrep do
                {
                sumx=0
                sumy=0
                for currenttime = 1:horizonlength do
                    { Generate X
                    Y = min(s,X)
                    sumx = sumx +X
                    sumy = sumy +Y
                    }
                fillrate(replications) = sumy/sumx
                }
            if mean(fillrate) ≥ 0.95 then
                { begin = s − (1/10^decimal)
                decimal = decimal +1
                Break
                }
            }
        }
```

[3]

Table 4: Descriptive statistics, using the alternative method (the ESC method).

s	m-erlang	fill rate %	Length (T)	mean	std. dev.	median	skewness
7.2639	5	95	5	0.9578	0.0584	0.9890	-1.5910
7.2639	5	95	10	0.9541	0.0448	0.9547	-1.0918
7.2639	5	95	20	0.9521	0.0330	0.9567	-0.7523
7.2639	5	95	100	0.9504	0.0153	0.9513	-0.3276
2.9957	1	95	20	0.9567	0.0550	0.9789	-1.5408
5.1863	3	95	20	0.9530	0.0388	0.9600	-0.9342
7.2639	5	95	20	0.9521	0.0330	0.9567	-0.7523
11.2858	9	95	20	0.9514	0.0276	0.9543	-0.5961

Algorithm 4 The simulation model including a positive lead-time. The after period is included such that if a customer arrives in the review horizon but could not be served within this horizon the waiting time of this customer can still be calculated.

procedure S APPROXIMATION($T, S, \lambda, numrep, L, warmup$)
  for $i = 1 : numrep$ **do**
    {Initialize variables and set variables empty
    $extra = L + 1$ ▷ to prevent the code from calling non positive index in vectors
    for t = 1:(T+warmup+afterperiod **do**
      {
      if $length(arrivaltime) > 0$ and OO(t-L+extra)>0 **then**
        { taken =0
        while length(arrivaltime)>0 and OO(t-L+extra)-taken>0 **do**
          {
          if frequency(1) > OO(t-l+extra)-taken **then**
            {
            if arrivaltime(1)>(warmup) and (T+warmup)≥arrivaltime(1) **then**
              {servedwaitingtime =[servedwaitingtime; (t-arrivaltime(1)]
              servedfrequency =[servedfrequency; (OO(t-l+extra)-taken)]}
            taken = OO(t-L+extra)
            frequency(1) = frequency(1) - (OO(t-l+extra)-taken)
            }
          else
            {
            if arrivaltime(1)>(warmup) and (T+warmup)≥arrivaltime(1) **then**
              {servedwaitingtime =[servedwaitingtime; (t-arrivaltime(1)]
              servedfrequency = [servedfrequency; frequency(1)] }
            taken = taken + frequency(1)
            delete arrivaltime(1), frequency(1)
            }
          }
        }
      $IL(t + extra) = IL(t - 1 + extra) + OO(t - L + extra) - d(t - 1 + extra)$, set OO(t-L+extra) =0
      $IP(t + extra) = sum(OO) + IL(t + extra)$
      $OO(t + extra) = Q = S - IP(t + extra)$
      $OH(t + extra) = max(0, IL(t + extra))$
      $B(t + extra) = max(0, -IL(t + extra))$
      $d(t + extra) = X = gamrnd(lapda, 1)$ ▷ play the actual period
      $Y = min(OH(t + extra), X)$
      if $t > warmup$ and $(warmup + T) \geq t$ **then**
        { $sumx = sumx + X$, $sumy = sumy + Y$
        if $Y == X$ **then**
          { $arrivaltime = [arrivaltime; t]$
          $frequency = [frequency; (X - OH(t + extra)]$ }
        }
      }
    $fillrate(i) = sumy/sumx$
    }

[4]

Algorithm 5 The algorithm that is used to determine the minimum required order-up-to level. For the verifying the results of Thomas (2005), the standard algorithm (partly described in Algorithm 3 is used instead of the *extensionsimulation* function that is called in the pseudo code. A general comment: extensionsimulation returns more than just these variables but only the one relevant for this algorithm are stated.

procedure FIND REQUIRED ORDER-UP-TO LEVEL($percentage, \lambda, L, warmup, benchmark, gap$)
  $stepsize = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100]$ ▷ The stepsize used for the extension part
  $stepsize = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100]$ ▷ The stepsize used for verifying Thomas (2005)
  for $counter = 1 : length(stepsize)$ **do**
    $\{T = stepsize(counter)$
    small = 1
    big = 100
    get fillratemeansmall = extensionsimulation($T, small, \lambda, L, warmup, benchmark$)
    get fillratemeanbig = extensionsimulation($T, big, \lambda, L, warmup, benchmark$)
    while big - small > gap **do**
      $\{middle = (small + big)/2$
      get fillratemeanmiddle, fillratefractionsuccesmiddle = extensionsimulation($T, middle, \lambda, L, warmup, benchmark$)
      if $fillratefractionsuccesmiddle \geq percentage$ **then**
        $\{big = middle\}$
      else if then
        $\{small = middle\}$
      $\}$
    get waitingtimemean = extensionsimulation($T, big, \lambda, L, warmup, benchmark$)
    $OrderUpToLevel = [OrderUpToLevel; big]$
    $waitingaverage = [waitingaverage; waitingtimemean]$
    $\}$

[5]

Figure 2: Distribution of fill rate with the same stocking level and per-period demand distribution but different horizon lengths, T.

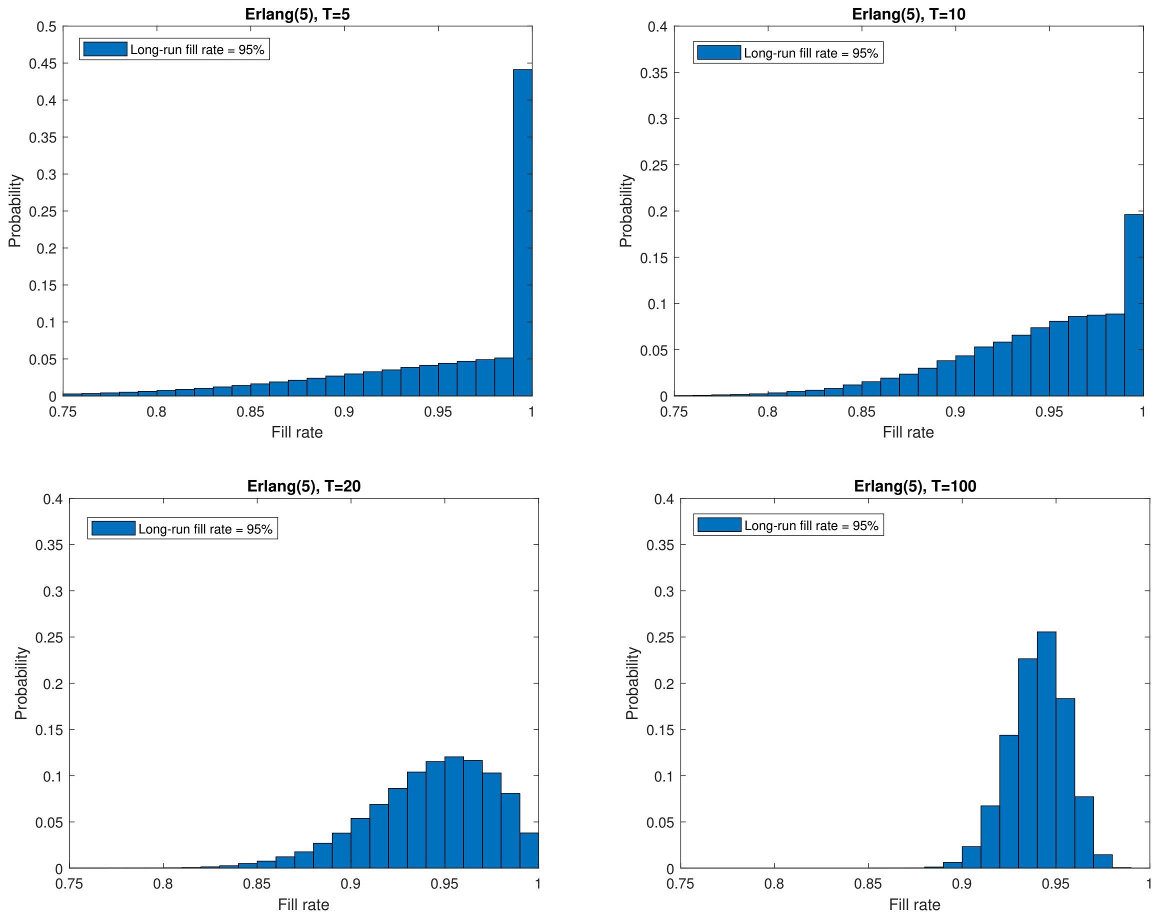


Figure 3: Distribution of fill rate with the same stocking level and per-period demand distribution but different horizon lengths, T.

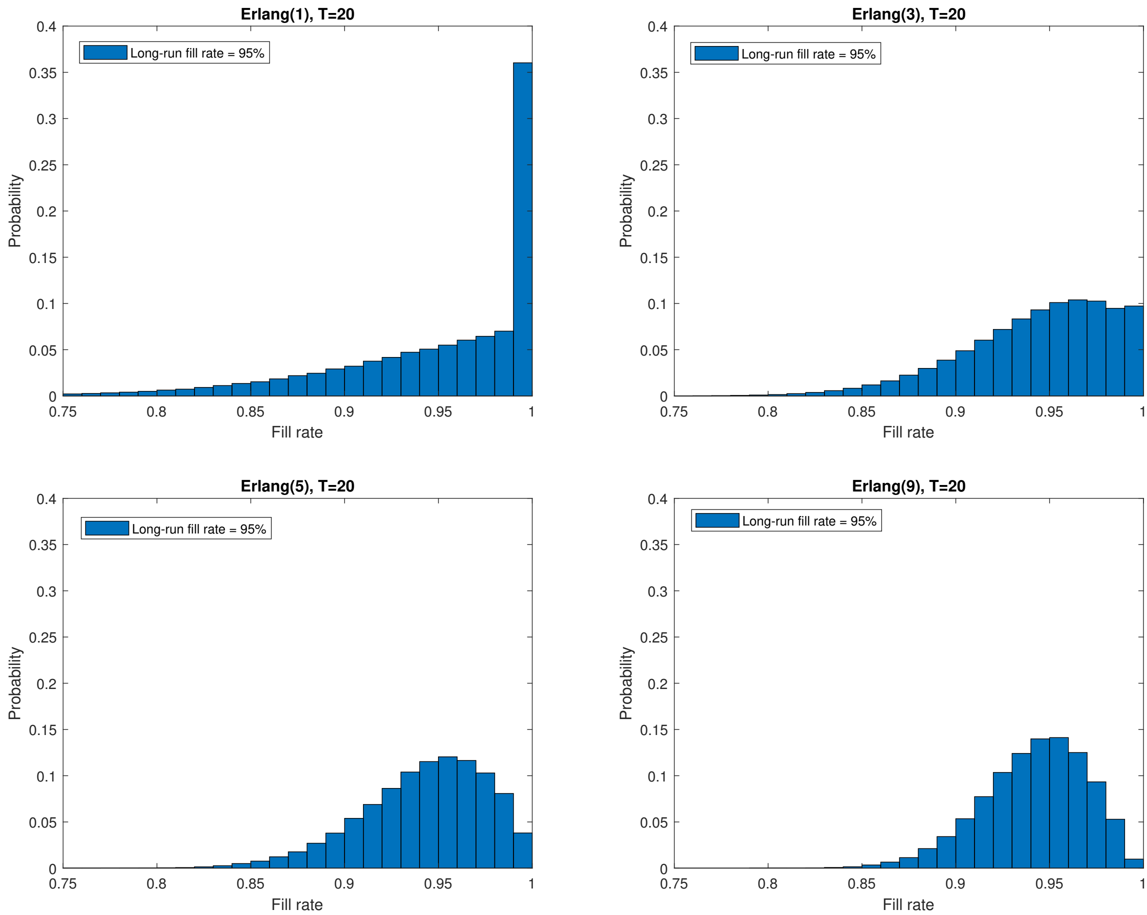


Figure 4: Stock levels required to have a specific chance of meeting a 99% target fill rate, with demand Erlang distributed. The stripped green line represents the long-run value of s.

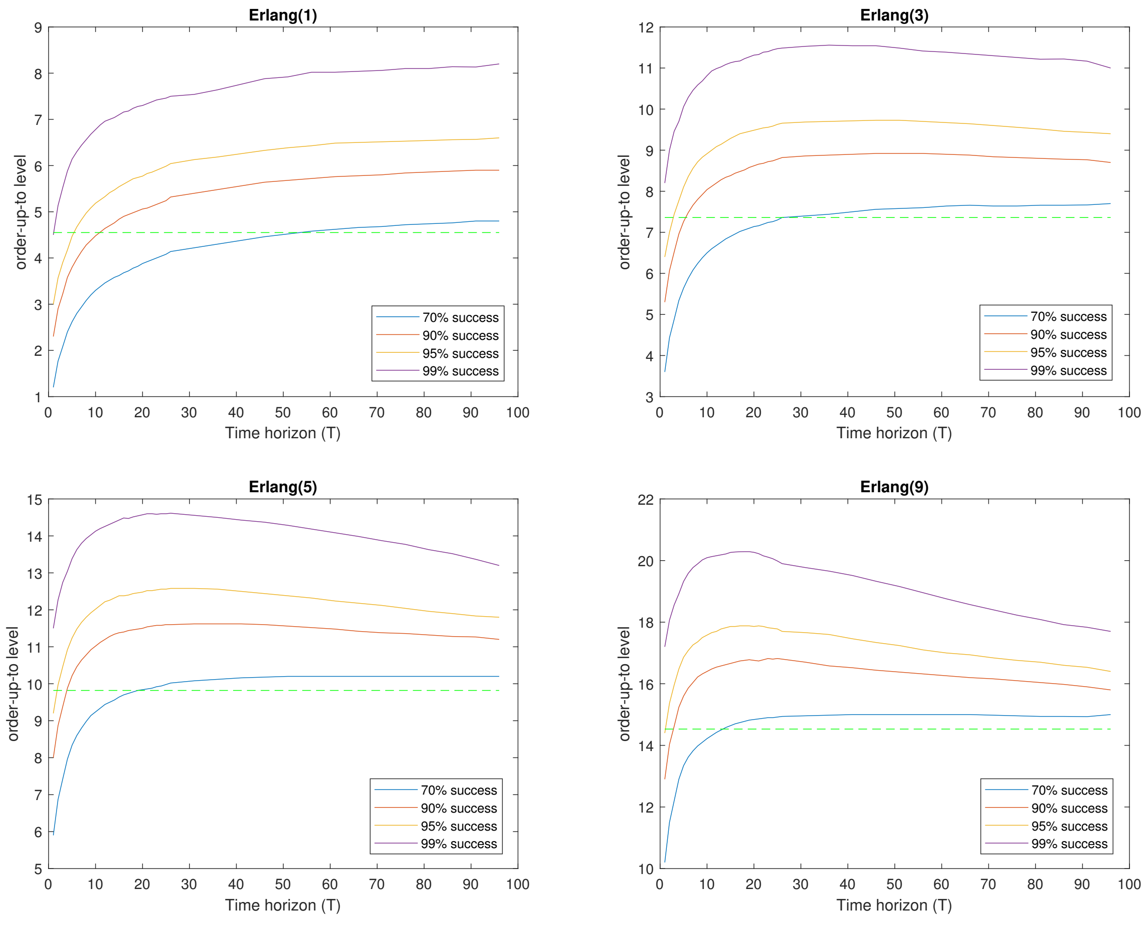


Figure 5: Stock levels required to have a specific chance of meeting a 95% target fill rate, with demand Erlang distributed. The stripped green line represents the long-run value of s.

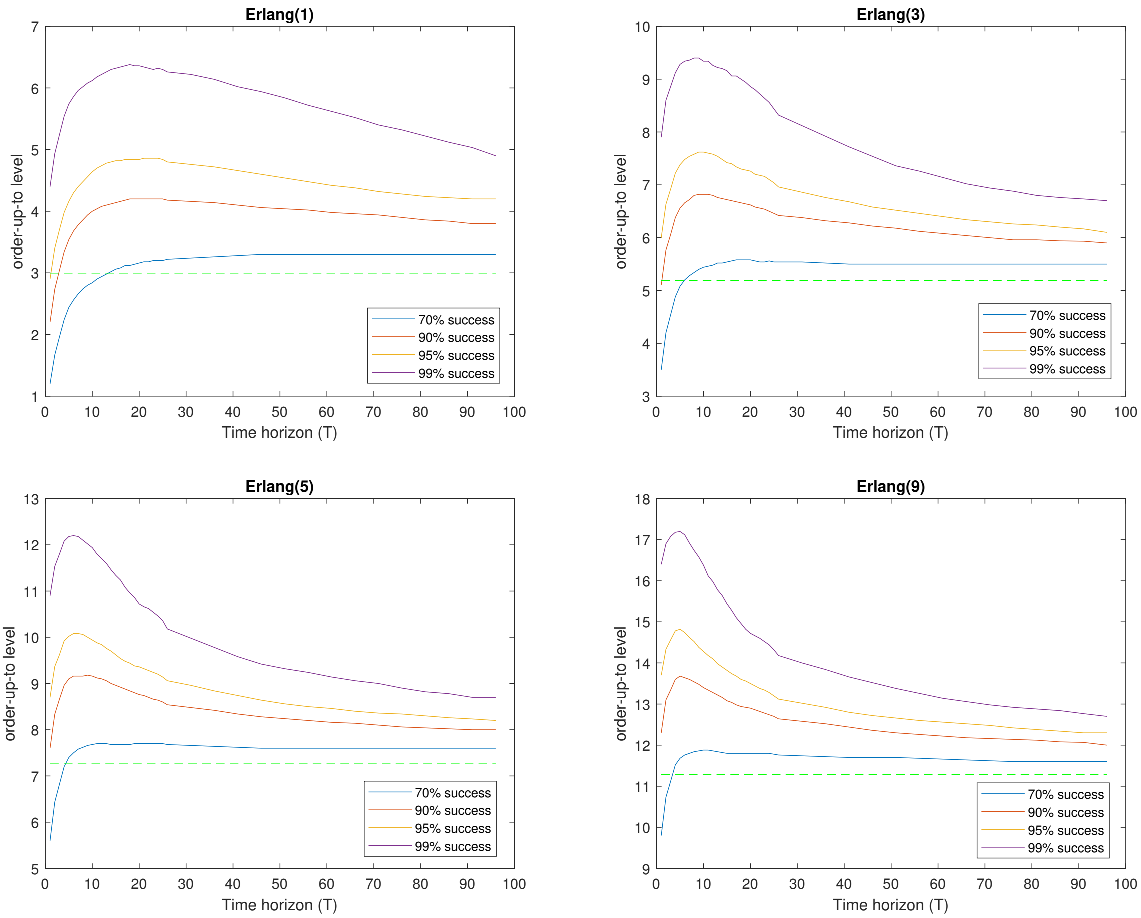


Figure 6: Erlang(5) distribution with an increased number of replications for a 95% fill rate requirement.

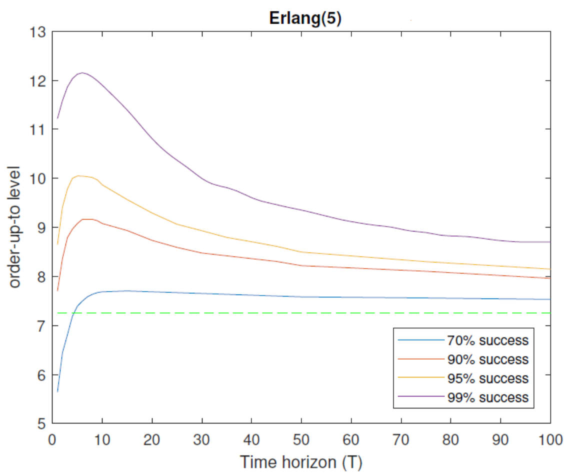


Figure 7: The mean fill rate for an Erlang(5) distribution with differing warm-up periods (T=1) and for different leading times plotted.

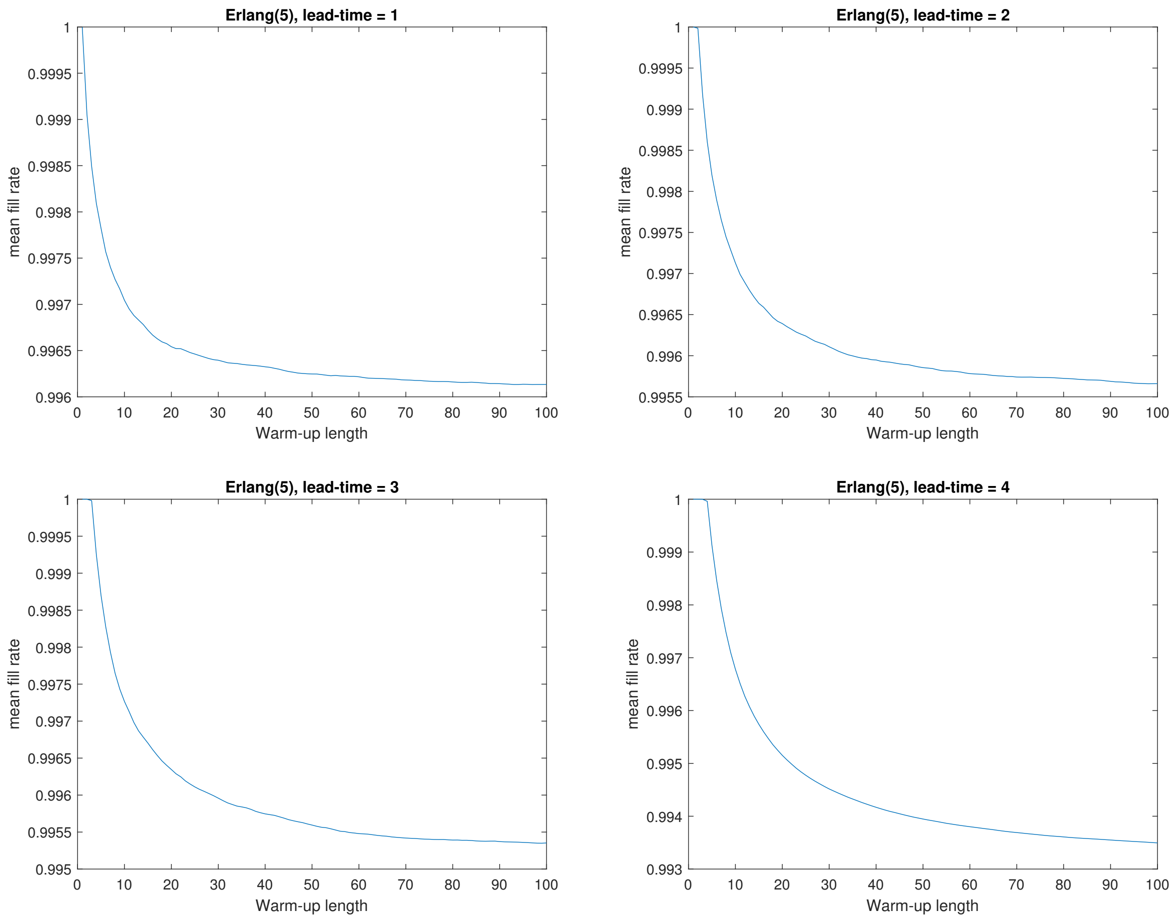


Figure 8: The mean fill rate for $T = 1$ for different order-up-to levels and a lead-time of 2 periods.

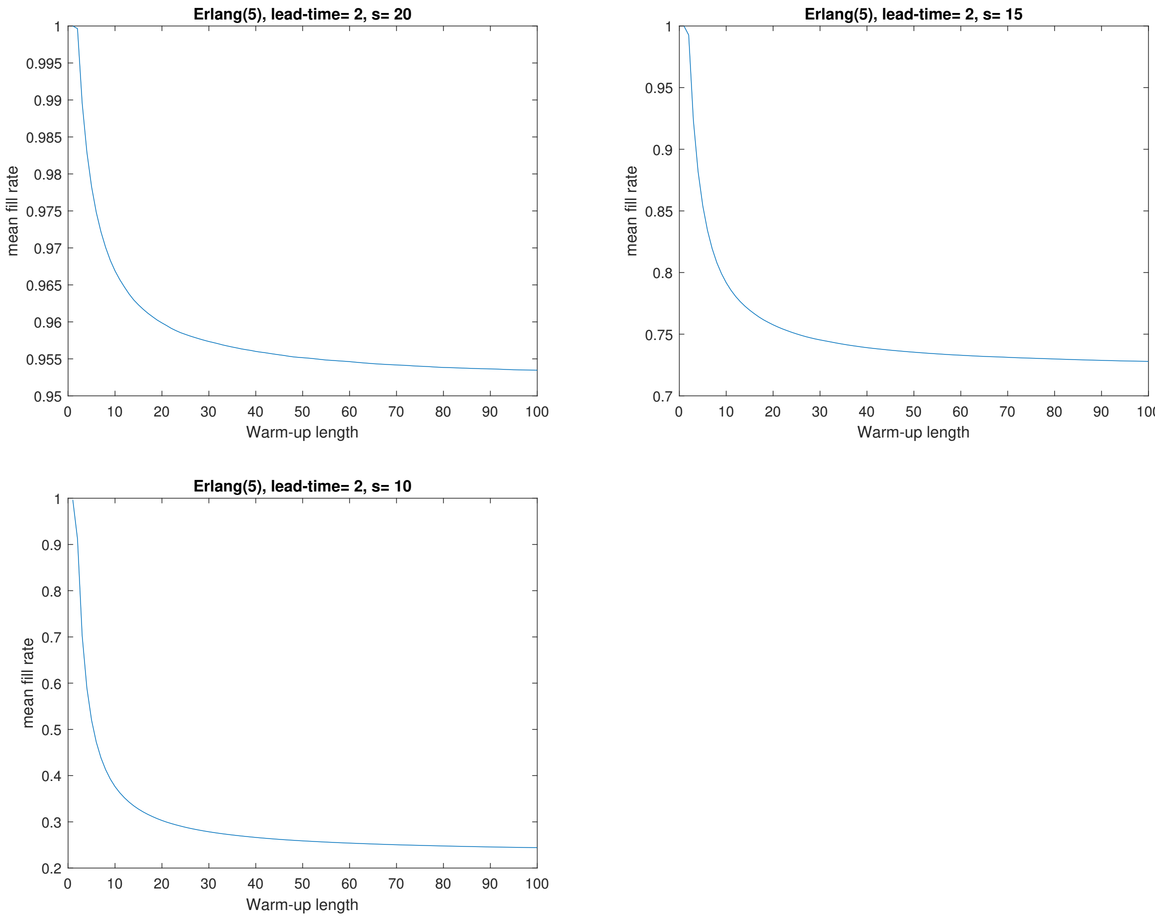


Table 5: The Table shows the required long-run value of s to obtain a 95% fill rate.

Erlang(k,1)	lead-time 1	lead-time 2	lead-time 3	lead-time 4
1	4.741	6.300	7.750	9.150
3	9.4250	13.3610	17.1430	20.821
5	13.759	19.920	25.902	31.770
9	22.098	32.562	42.810	52.9410

Figure 9: The distribution of the mean fill rate for an Erlang(5) and a horizon length of $T = 20$.

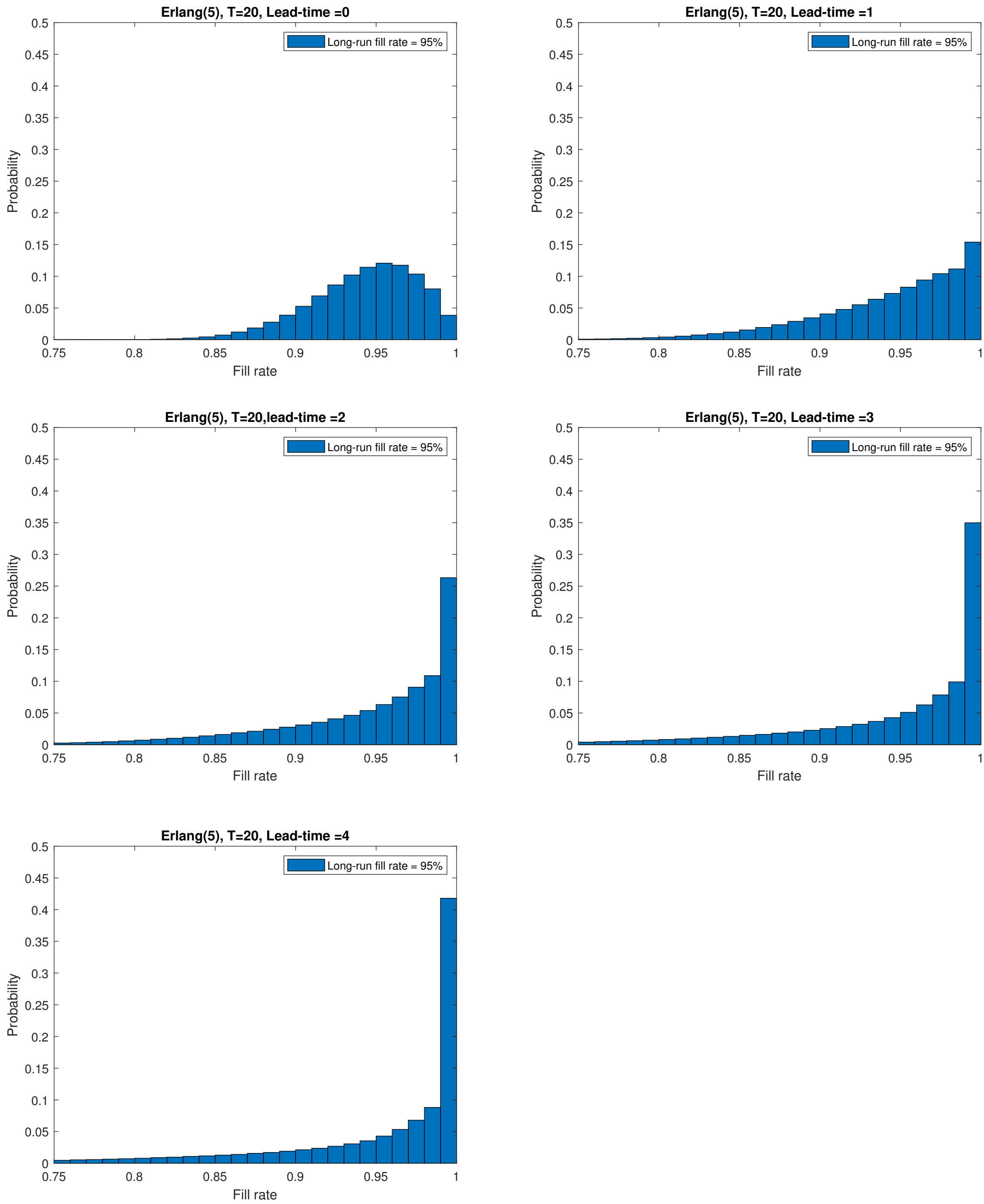


Figure 10: The mean fill rate for an Erlang(5) demand distribution over differing horizon lengths. The stripped green line represents the mean fill rate for an infinite horizon.

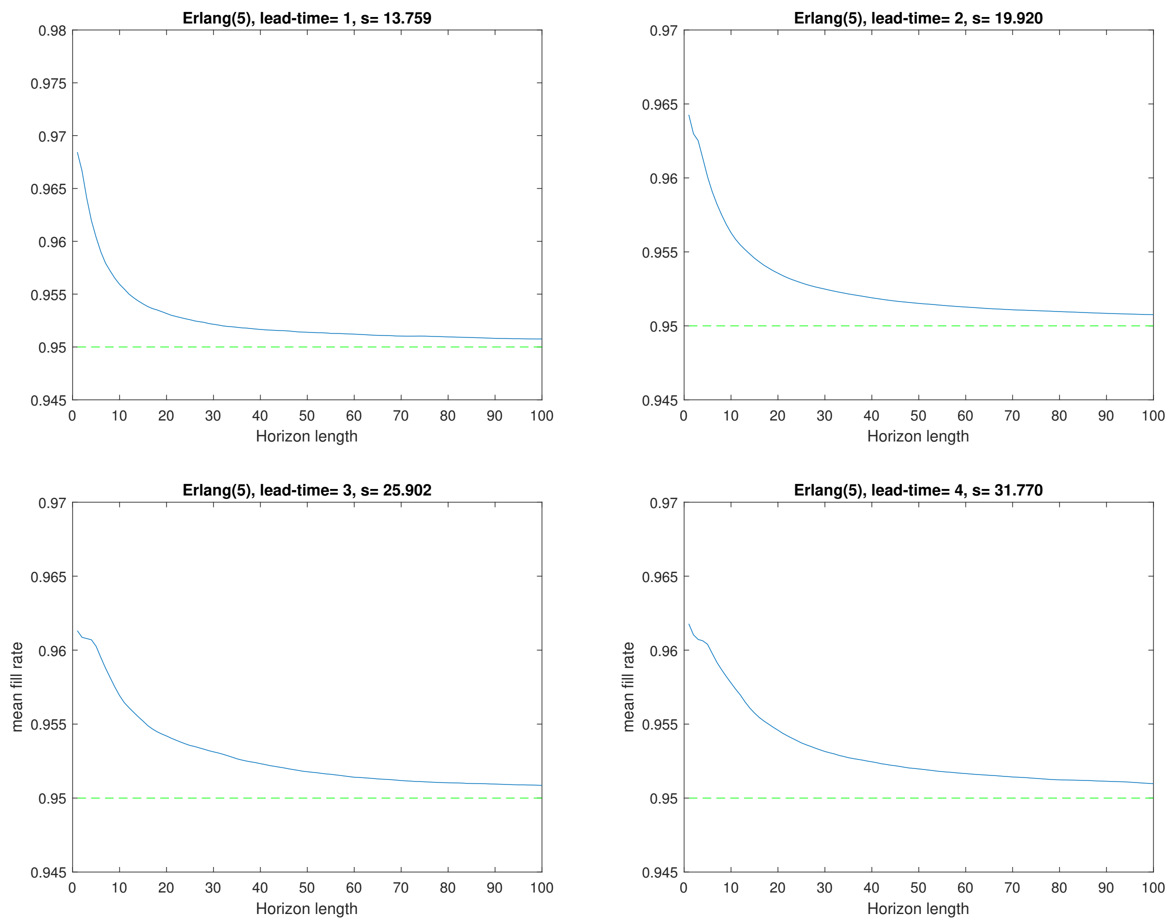


Figure 11: The stripped green line represents the long-run value of s for a fill rate requirement of 95% is obtained for Erlang(5) demand. The figure shows the required order-up-to levels for various lead-time lengths.

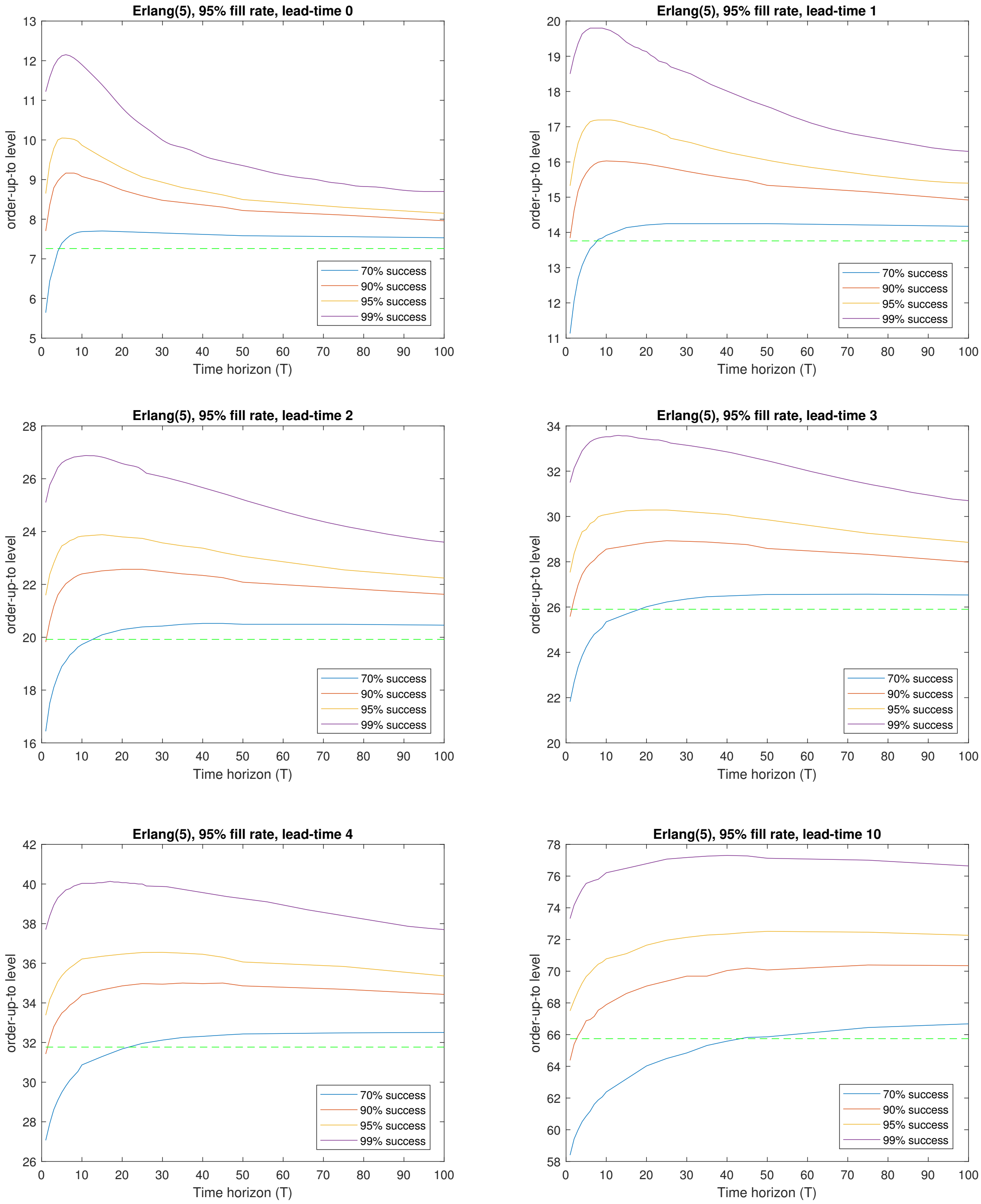


Figure 12: The stripped green line represents the long-run value of s for a fill rate requirement of 95% is obtained. The figure shows the lead-time 10 behavior for a horizon length up to $T = 200$.

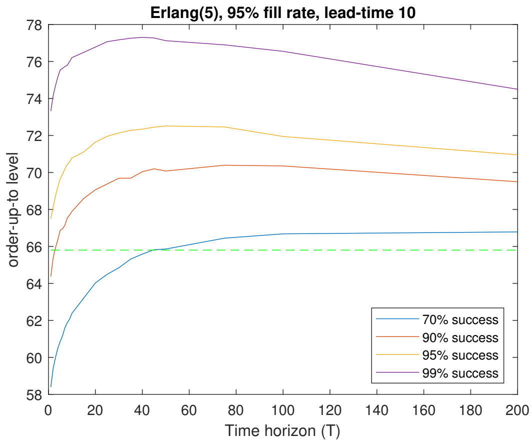


Figure 13: The stripped green line represents the long-run value of s for a fill rate requirement of 95% with demand being Erlang(1) distributed.

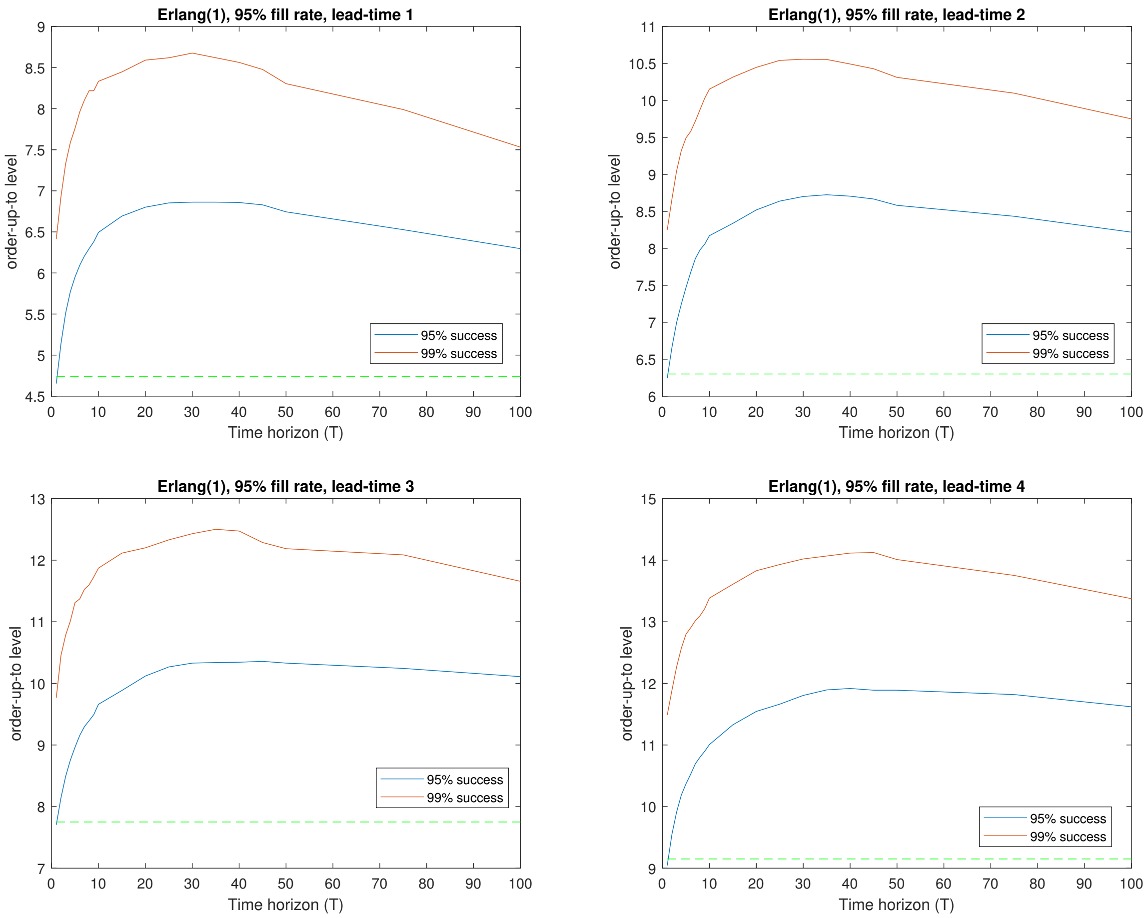


Figure 14: The stripped green line represents the long-run value of s for a fill rate requirement of 95% with demand being Erlang(3) distributed.

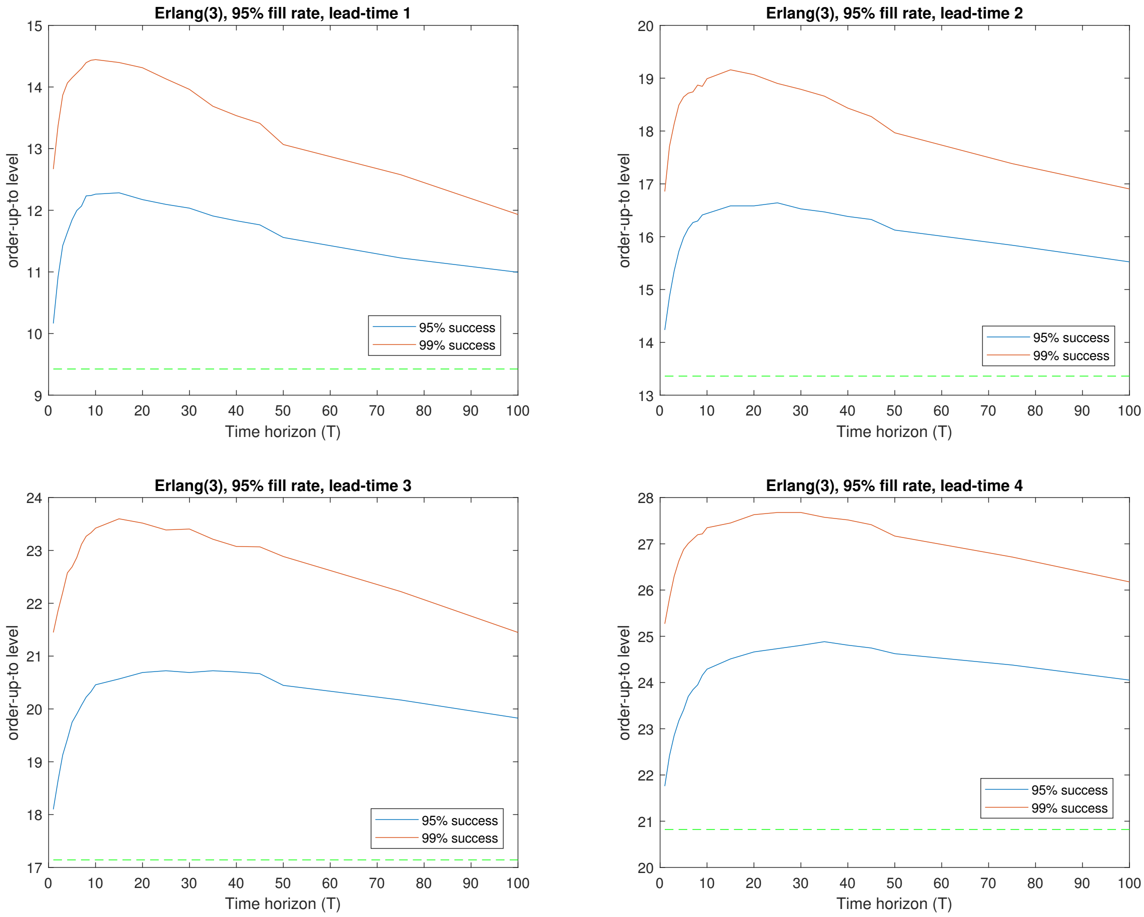


Figure 15: The stripped green line represents the long-run value of s for a fill rate requirement of 95% with demand being Erlang(9) distributed.