Chapter 2
Concepts and Definitions

We first give a brief introduction to the objects and phenomena that we will examine in this study. The key object of our interest is an inventory system within a non-changing, static environment. In order to make any statements about this inventory system, we will first define how to describe the *state* of such a system in terms of the relevant characteristics. From an epistemological point of view, we will introduce the vocabulary here that enables us to talk about the object being examined. Next, we propose a framework to classify the structure and relevant elements of an inventory system within a static environment. This should help us to later on declare the specific nature of the system being analyzed and its variants. To evaluate the specific configuration of a system and allow for a preferential order with alternative configurations, the last section summarizes common metrics for evaluating the performance of inventory systems.

2.1 State of an Inventory System

From the pure perspective of keeping inventory, the following seven basic concepts are generally used in the literature to characterize the state of an inventory system at a certain moment t in time. In discrete time, these states commonly refer to the end of each period. Instead of inventory, a significant fraction of the literature uses the term *stock*, at least in some of the definitions. In the following, we use the term inventory to describe amounts of the goods in question, whereas stock refers to the installation (e.g., a warehouse) in which the goods are kept.

- *Physical inventory.* Physical inventory (I_t) is the amount of inventory that is immediately available to satisfy incoming demand. I_t is sometimes also referred to as inventory on hand.
- *Number of backorders.* The number of backorders (B_t) or simply backorder(s) is the volume of customer orders that currently cannot be satisfied due to inventory unavailability, and that are not lost. Backorders occur whenever customers are

T. Wensing, *Periodic Review Inventory Systems*, Lecture Notes in Economics and Mathematical Systems 651, DOI 10.1007/978-3-642-20479-1_2,

willing to wait for their orders, and the physical inventory is insufficient to cover their demand. The term *backlog* is often used as synonym for B_t in the literature.
- *Outstanding orders.* The (number of) outstanding orders K_t is the random number of orders that have been issued before time t, and have not yet arrived.
- *Inventory on order.* Inventory on order (IO_t) is the amount of material that has been ordered from the supplier but has not yet arrived. In other words, it is the total amount of material covered by outstanding orders.
- *Inventory shortfall.* The inventory shortfall (SF_t) is the difference between planned and the actual inventory levels. It is equal to the inventory on order plus the demand that has occurred since the last replenishment order was placed (see also Sect. 5.2.2.).
- *Net inventory.* Net inventory (NI_t) is defined as physical inventory minus backorders ($NI_t = I_t - B_t$).
- *Inventory position.* The inventory position (IP_t) is defined as physical inventory minus backorders plus inventory on order ($IP_t = I_t - B_t + IO_t$).

The state of an inventory system in terms of the concepts defined above may be changed by three events: order initiation, order arrival and demand occurrence. We thus have the following three concepts to describe the change of an inventory system at moment t in time, as compared to the previous moment. In our case of a discrete time axis, the previous moment is $t - 1$, so we have chosen to illustrate the definitions according to this paradigm:

- *Demand.* Demand (D_t) is the customer demand that is requested from the system in t. The demand may effect all of the state concepts described above.
- *Issued Inventory.* Issued inventory (II_t) is the amount of inventory that is ordered from suppliers in t. The issued inventory may effect the outstanding orders, the inventory on order and the inventory position (e.g. $IP_t = IP_{t-1} + II_t - D_t$).
- *Arriving Inventory.* Arriving inventory (AI_t) is the amount of inventory on order that finally arrives in t. The arriving inventory may effect all of the state concepts described above except for the inventory position (e.g., $IO_t = IO_{t-1} + II_t - AI_t$, $I_t = max\{0, NI_{t-1} + II_t - D_t\}$).

Finally, we have the concept of replenishment lead time which somewhat connects the issued and arriving inventory.

- *Replenishment lead time.* Replenishment lead time (L_k) is the time that passes between issue and arrival of an order indexed with k.

The concept of *safety stocks* (SS) is also frequently used in inventory management. It is commonly referred to as the amount of physical inventory that would never be undershot if replenishment lead times and demand were deterministic according to their expected values. In other words, *safety stock* is the amount of stock that is kept beyond expected requirements. From the perspective of analyzing a policy, SS is more of a performance indicator, even if it is often perceived as a parameter for configuring an inventory system. We do not adhere to this perspective in the following, instead, we will always define the configuration of an inventory system by declaring when and how much to order.

2.2 Classification

In the literature on inventory theory, various different systems are analyzed using diverse methods. Comparing the approaches on different systems reveals that even minor changes in the assumptions may lead to dramatic changes in systems behavior. It is therefore essential for the analysis to carefully define the nature of the system being examined.

This section presents a classification scheme for inventory systems with special consideration of the systems we analyze in Chap. 6. The organization of the scheme basically follows the proposal of Hollier and Vrat (1978). On the top level, we distinguish between the system structure, its environmental parameters and the replenishment policy that is applied. Similar schemes with less aggregation are given by Aggarwal (1974), Silver (1981) and Silver (2008). An even more detailed scheme is proposed by Prasad (1994). Deviating from Hollier and Vrat, we do not consider inventory-related costs as part of the system classification. The performance aspects (including costs) are therefore treated separately in Sect. 2.3.

We will briefly indicate for each item to what extent it is considered in the analysis in Chap. 6. These model limitations are also summarized in Sect. 2.2.4.

2.2.1 Structure

In terms of the overall structure, we primarily distinguish between *single-level* (or single-echelon) and *multi-level* models, where the latter may have a linear, converging, diverging or general structure. In a linear structure, each stock may have one predecessor (supplier) and one successor (receiver) at most. In converging systems, each stock may have multiple predecessors but only one successor at the most, whereas this is vice versa in diverging systems. Finally, a system is said to have a general structure if a stock may have multiple predecessors as well as multiple successors.

Furthermore, inventory systems may hold a *single item* or *multiple items*. In the event of multiple predecessors, the inventories of certain items may be replenished from a *single source* or from *multiple sources*.

In the following we will limit the analysis to the single-item and single-source case.

2.2.2 Environmental Parameters

Every inventory system is emptied by a demand process and refilled by a replenishment process, both of which significantly influence the systems behavior. Furthermore, the nature of the stored item may be of importance. The remainder of this section takes a closer look at these three aspects.

2.2.2.1 Demand Process and Customer Order Fulfilment

An inventory system can either face *stationary* or *dynamic* (time-varying) demand. Furthermore, the *interarrival time* between two demand occurrences as well as the *volume* of a single demand occurrence can be either *deterministic* or *stochastic*. In case of stochastic interarrival times and/or demand volumes, the underlying distribution may be fully known, only be known in type but not in parameterization or be completely unknown. Both, interarrival times and demand volumes may be *independent* from or *dependent* on the previous occurrences.

In the event that demand exceeds the physical inventory, customers are either utterly willing to wait (*full backordering*), may or may not be willing to wait (*partial backordering*), or may immediately lose interest (*lost sales*). If backordering is possible and physical inventory is only sufficient to partly cover the demand, the system may allow for *split deliveries* to the customer, or may only fulfill customer orders with *full deliveries*.

In the following, we will only consider stationary demand occurring with a constant interarrival time. Demand may be deterministic or stochastic, with the distribution completely known and independent from previous occurrences. Demand may always be backordered without limitation, and we address the case of split deliveries as well as that of full deliveries only.

2.2.2.2 Replenishment Process

Analogously to demand arrivals, the replenishments arrive after a certain (lead) time and consist of a certain amount. The lead time may be *stationary* or *dynamic*. It may furthermore be *deterministic* or *stochastic*, where in the latter case the distribution may be known, be known in type but not in parameterization, or may be completely unknown. Stochastic lead times may be considered *independent* from or *dependent* on the previous occurrence or even a history of occurrences.

Replenishment deliveries may either arrive in the exact amount as issued (full reliability), or may deviate with known, partially known or unknown distribution. Finally, there may or may not be dependencies between lead times and order amounts.

Please also refer to Chap. 5 for further distinctions concerning replenishment processes.

In the following we will assume that replenishment lead times are stationary, independent from order volumes and either deterministic or stochastic with known distributions. Amounts delivered may not deviate from what has been ordered. We also consider independent lead times as well as a case of lead time dependencies.

2.2.2.3 Stored Goods

Stored goods may be considered *non-perishable* or *perishable*. Perishable products may either be completely lost or decrease in value or usefulness, if kept in stock for too long.

We will always assume non-perishable goods in the following.

2.2.3 Replenishment Policies

In the introduction we outlined that inventory management always revolves around the two fundamental questions of when to replenish the inventory and how much to order. In the literature we commonly find two decision parameters for each of the two questions that may be combined to form an inventory policy. On behalf of the first question, we can either place an order every fixed period r or as soon as the inventory position falls below a particular value s. The order volume may either be a fixed quantity q or may be determined as the difference between a value S (the so-called order-up-to-level) and the inventory position.

Considering the task of determining current inventory levels, we may add the further decision of how often the inventory status should be determined. (See Silver et al. 1998, for example.) In that context, one may either want to establish a fixed review period t or continuously review the inventory levels ($t \to 0$).

From the combination of these possible decision parameters, we may derive five useful inventory policies as boldly highlighted in Table 2.1. Note that the possible (t, r, S) policy is not explicitly regarded in the literature as it would be useless in terms of the replenishment rule to review the inventory levels without having the option to place an order. Thus, $t := r$ is commonly assumed, where this (special) case is covered by the (r, S) policy. Furthermore, a non-adaptive policy such as (t, r, q) or (r, q) is not appropriate in a stochastic environment.

Thus, we derive the five basic decision rules printed in bold in Table 2.1, two of which apply continuous review and three of which apply periodic review.

The replenishment doctrines highlighted in bold can be described as follows. Remember that IP_t denotes the inventory position at time t.

- *(s,S)*: Order a variable amount of $Q_t = S - IP_t$ as soon as IP_t falls below s. This policy demands high standards of the supply and inventory review process. Changes in the IP must instantly be monitored, and it must be possible to order any amount in Q_t at any time. A special case of this policy, with $s := S - 1$ is discussed as base stock policy. Here, an order is placed as soon as one unit or more is taken from stock. Requirements of the supply and inventory review process are more or less equivalent to those of the (s, S) doctrine.

Table 2.1 Elementary inventory policies

No.	Monitoring t or 0	Impulse r or s	Quantity q or S	Policy
1	t	r	q	(t,r,q)
2	0	r	q	(r,q)
3	t	s	q	**(t,s,q)**
4	0	s	q	**(s,q)**
5	t	r	S	(t,r,S)
6	0	r	S	**(r,S)**
7	t	s	S	**(t,s,S)**
8	0	s	S	**(s,S)**

- *(s,q)*: Order a fixed quantity q as soon as IP_t falls below s.
  This policy reflects that the supplier may only offer certain discrete order batches of size q or an integer multiple of q. As for the (s, S) policy, instant review as well as the possibility that an order could always be placed is required for the application of this policy.
- *(t,s,S)*: Every period t, review IP_t and order a variable amount $Q_t = S - IP_t$ if IP_t is below s.
  This periodic review equivalent to the (s, S) policy requires that the supplier may deliver any amount in Q_t. Review effort is reduced to the possible order periods, where the main benefit for the process is the possibility of coordinating orders over time.
- *(t,s,q)*: Every period t, review IP_t and order the smallest multiple of the fixed quantity q that raises IP_t above s. This policy is also discussed as $(t, s, n \cdot q)$ policy to indicate that a multiple of q may be necessary to raise IP_t sufficiently.
  This periodic review equivalent to the (s, q) policy offers the possibility of coordinating order processes regarding time due to the fixed review interval t as well as regarding volume due to the fixed q.
- *(r,S)*: Every period r, order a variable amount of $Q_t = S - IP$.
  The characteristics of this policy are more or less equal to those of the (t, s, S) policy. In contrast, it lacks the possibility of skipping an order as a reaction to low demand occurrences between two order periods.
  This is the decision rule that we will examine in this study.

Comparing these policies, the (t, s, q) and (r, S) policy appear to be the most advantageous with respect to the supply and review processes. In turn, however, we are most likely observing the highest costs for operating those policies in terms of costs for inventory holding and backordering or lost sales.

2.2.4 Summary of Model Assumptions

This study considers a single-level, single-item and single-source inventory system with non-perishable goods. Replenishment orders are placed according to the periodic review order-up-to (r, S) doctrine.

From the customer side, our system observes stationary (static) stochastic or deterministic demand batches with a constant interarrival time. The distribution of the demand batches is known, and each realization is independent of previous occurrences. We assume full backordering, where two alternative delivery modes are addressed in the event of material insufficiency: we examine the option of split deliveries where a customer order may be delivered in two or more instalments as well as the restriction to the delivery of complete orders only.

From the supplier side, we observe stationary (static) stochastic or deterministic lead times that are independent of the volumes that we order. The arriving order amounts may not deviate from what has been ordered. We examine an independent lead time process as well as the case of interdependent lead time occurrences.

2.3 Performance Indicators

2.3.1 Costs

Assuming that the operating tasks of an inventory system range from material replenishment to customer order delivery, we may identify five types of relevant costs: the purchase costs of the replenishment material, the inventory holding costs, the costs for fulfilling customer demand, the costs that occur as the result of a stockout situation, and the costs of operating the inventory system itself, e.g., the effort required for data gathering and control procedures (Hadley and Whitin 1963, Chap. 1). A detailed description of the drivers of inventory system costs is given by Brooking (1987), for example.

In the more recent literature, we usually find the order fulfillment task reduced to holding the inventory available, where the actual delivery costs are not taken into account (Silver et al. 1998, Chap. 3).

For our scope, we will also neglect the costs of operating the system itself, as we will examine different parameterizations of the same basic system, where we assume identical operating costs. We thus retain the three *classical* types of costs that are considered for evaluating inventory system policies, namely purchasing costs, costs for inventory keeping, and costs that are incurred due to stockout situations.

2.3.1.1 Purchasing Costs

Purchasing costs may be incurred per unit and/or for an entire order, while both may depend on the size of the corresponding order. Dependence occurs for example, if the supplier offers volume discounts. In the following however, we will assume that costs per unit as well as costs for full orders are independent of order sizes, because this aspect is not what we want to focus our model on. In our case, acquisition costs cannot be influenced by the parameterization of an inventory system, and are thus irrelevant for the decision on order quantities. Nonetheless, we define them here because we need them for the proper calculation of inventory holding costs.

Definition 1 (Fixed acquisition price per unit). We define p as the fixed acquisition price per unit that is kept in stock.

The total order-related costs of replenishing inventory may well be influenced by the parameterization of an inventory system. These costs are obviously higher the more often replenishment orders are placed.

Definition 2 (Fixed costs per order). We define c_1 as fixed costs that are incurred for placing an order of arbitrary size.

2.3.1.2 Inventory Holding Costs

Definition 3 (Inventory holding costs). Let p be the acquisition price (Definition 1), i the interest rate per period, and h the costs of keeping one unit in stock for one period that do not include costs of tied capital. Then

$$c_2 = p \cdot i + h \tag{2.1}$$

are the costs that are incurred for keeping one unit of stock for one period.

The quantification of these costs for an inventory system requires computation of average stock levels for a certain time span that is fully representative for the system behavior. In our case of the (r, S) replenishment doctrine, this time span is one complete order cycle of length r.

From an accounting perspective, one might argue that the fixed order costs c_1 must also be considered proportionally within the expected purchase price, and thus influence c_1. Let Q be the stochastic order size that is implied by the application of a stock policy. Then we may state $c_2(.)$ depending on the order quantity as follows:

$$c_2(E[Q]) = \left(p + \frac{c_1}{E[Q]}\right) \cdot i + h. \tag{2.2}$$

Whenever the average inventory develops proportionally to the average order size, i.e., $I^{\emptyset}(E[k \cdot Q]) = k \cdot I^{\emptyset}(E[Q])$, we can easily show that the tied capital associated with the fixed order costs is independent of the order size Q:

$$\begin{aligned} c_2(E[k \cdot Q]) \cdot I^{\emptyset}(E[k \cdot Q]) &= \left(p \cdot i + \frac{c_1 \cdot i}{E[k \cdot Q]} + h\right) \cdot I^{\emptyset}(E[k \cdot q]) \\ &= \frac{c_1 \cdot I^{\emptyset}(E[k \cdot Q]) \cdot i}{E[k \cdot Q]} + \left(p \cdot i + h\right) \cdot I^{\emptyset}(E[k \cdot Q]) \\ &= \frac{c_1 \cdot k \cdot I^{\emptyset}(E[Q]) \cdot i}{k \cdot E[Q]} + \left(p \cdot i + h\right) \cdot I^{\emptyset}(E[k \cdot Q]) \\ &= \frac{c_1 \cdot I^{\emptyset}(E[Q]) \cdot i}{E[Q]} + \left(p \cdot i + h\right) \cdot I^{\emptyset}(E[k \cdot Q]) \end{aligned} \tag{2.3}$$

However, we cannot generally assume that $I^{\emptyset}(E[k \cdot Q]) = k \cdot I^{\emptyset}(E[Q])$ holds. Using a constant c_2 therefore must be considered an approximation under general conditions.

2.3.1.3 Costs of Stockout

When an inventory system fails to provide a required amount of material in time, some negative consequences have to be expected, otherwise the need for

the inventory system is called into question. In the literature, three basic types of costs are discussed to estimate the consequences of unavailability of material. See Schneider (1981) for example.

Definition 4 (Costs of being in stockout state). We define c_{31} as the costs that are incurred if an inventory system is unable to provide material in the corresponding period, regardless of the amount of material missing and the previous duration of the stockout.

Relatively irrelevant from the perspective of the inventory system's customers, costs of type c_{31} may be incurred in connection with general arrangements to overcome the stockout situation. This could be the exceptional start of a production process, for example, or the request for an expensive emergency delivery, where the actual amount of missing units is secondary for the calculation of total costs. Even if the stockout does not induce any emergency activities, costs of type c_{31} may reflect a loss of customer goodwill as a reaction to the news that the system is in trouble.

Definition 5 (Costs per missing unit). We define c_{32} as the costs that are incurred per unit of material that cannot be provided in time.

Costs of type c_{32} are typically incurred if potential sales are lost in the event that immediate delivery is not possible.

Definition 6 (Costs per missing unit and time). We define c_{33} as the costs that are incurred if the delivery of one unit of material is delayed for one (further) period.

In contrast with c_{32}, c_{33} is only observed when unsatisfied demand can be delivered with a delay (backorder case).

It will generally be difficult to give reasonable estimates for all three types of stockout costs. This problem especially emerges from the wide range of possible customer reactions when demands cannot be satisfied. At one end of the scale, customers may be willing to wait without any compensation, while at the other, they may forever be lost as business partners. To circumvent the problem of estimating stockout costs, a common approach is to define service metrics, where the inventory system must fulfill a certain prescribed level.

2.3.2 Service Metrics

The pertinent literature discusses a variety of different elementary and compound service metrics. See Zinn et al. (2002) for an illustrative overview and Boylan and Johnston (1994) for insights into service measurement when orders may include multiple items.

We focus our scope on non-compound service metrics for single items. Following the logic that underlies our definition of stockout costs in Sect. 2.3.1.3, we consider one analogous service metric for each of the three types of costs that we have defined above. See Ronen (1982), for example, for a similar scope.

We have to distinguish two perspectives for determining the values of certain service metrics. On the one hand, we can obtain these values a posteriori from the observed history of a real or simulated inventory system. In this case, we may speak of the random final state of a stochastic process that has run for a certain time. On the other hand, we can try to estimate the expected values a priori from the parametrization of a modeled system. This is the scope of Chap. 6.

In our view, it is important to understand that the metrics defined in the following sections describe either observed or expected relative frequencies. They may be *interpreted* as probabilities if it can reasonably be assumed that the system will behave in the same manner in the future as it did in the past. See Hacking (2001), p. 127–139, on the different conceptions of probability, and Hacking (2006) for an overview of the historic development that also illustrates the problems of the different conceptions.

2.3.2.1 Ready Rate

Definition 7 (Ready rate). Let IS be an inventory system that is observed in a time interval of length τ. Let A^τ be a discretely distributed random variable with two possible states, where $A^\tau = 1$ means that the system had material on hand throughout the full time interval, and $A^\tau = 0$ means that the system was out of stock for at least a fraction of the time interval regarded. Then we define

$$\alpha^\tau = E[A^\tau] \tag{2.4}$$

as the ready rate for a basic interval of length τ, $\alpha^\tau \in [0, 1]$.

As indicated by the symbol, this event-oriented service level is often referred to as α-service level, especially in the European literature.

Let IS^* be a real or simulated inventory system with an underlying and probably unknown A^τ. Let $a^{\tau,i*}$ be the recorded system behavior in one interval i of length τ, where $a^{\tau,i*} = 1$ if all demands in i could immediately be fulfilled and $a^{\tau,i*} = 0$ otherwise, and finally let $\mathcal{I}$ be a set of observed time intervals. Then

$$a^{\tau*} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} a^{\tau,i*} \tag{2.5}$$

is an unbiased estimator for α^τ. We can interpret each $a^{\tau,i*}$ as the random draw from an underlying and probably unknown A^τ, while $a^{\tau*}$ is obviously the fraction of all observed periods in which no stockout has occurred. In the literature, $a^{\tau*}$ is sometimes referred to as *realized* ready rate or *realized* α service level, respectively. See for example Suchanek (1996) in Chap. 4 and Tempelmeier (2006) in Chap. B.3.

Various values for τ may apply. Especially in practical applications, τ is often chosen as equivalent to some period length critical for the planning (i.e., 1 day, week, month, ...). In this case, α^τ is referred to as the *periodic* ready rate. Two values for τ that are more inventory system-related are the replenishment lead time and the order interarrival rate. The service level corresponding to the first value is then referred to

as *cyclic* ready rate, while the one corresponding to the latter value is equivalent to the probability that a customer order can be completely fulfilled without any delay.

If τ refers to a time unit smaller than an order cycle, we may observe different values for the *micro* periods of a cycle. For example, if replenishment orders always arrive in the second period of an order cycle, then the stockout probability for this period is most likely lower than it is for the previous period in which the replenishment material is still outstanding. It may therefore be helpful to differentiate the corresponding service levels for the periods of a subdivided order cycle.

For values of τ that are larger than the customer order interarrival rate, one or more stockout incidents might be considered as one event. Under special circumstances, the result may be influenced by the pattern we choose to jointly consider the periods of an order cycle. For example, let $SEQ_1 = 00001111000011111$ and $SEQ_2 = 10000111100001111$ be two sequences of inventory states that have been observed for two inventory systems, where 0 denotes unavailability and 1 availability of material. In the event that we consider four consecutive periods as one cycle, we observe $\alpha^{\tau=4} = 0.5$ for SEQ_1 and $\alpha^{\tau=4} = 0.0$ for SEQ_2 for the four fully observable cycles if we start with the first period and vice versa $\alpha^{\tau=4} = 0.0$ for SEQ_1 and $\alpha^{\tau=4} = 0.5$ for SEQ_2 if we start with the second period. This effect may only be observed if replenishment lead times are longer than the order cycle and demand is highly volatile. Nonetheless, it raises some doubt as to the expressiveness of the commonly used ready rate per replenishment cycle.

In the literature, the ready rate is also defined as the fraction of time for which the net inventory levels are positive, see Axsäter (2006) for example. However, we do not consider this definition appropriate for a discrete time axis. In this instance, we may consider time units τ in which more than one event can possibly alter the inventory levels, so that we may observe both positive and negative stock levels in the same underlying period.

For a discussion of fiscal-period-based versus lead-time-based measurement of the ready rate, see Haehling von Lanzenauer and Hamidi-Noori (1986). Insights on mathematical properties of a special ready rate metric are studied by Zipkin (1986a), for example.

2.3.2.2 Fill Rate

Definition 8 (Fill rate). Let IS be an inventory system that observes random demand D^τ in time intervals of length τ. Let $D^{\tau,p}$ be the random demand in τ that could be served by IS without delay. Then we define

$$\beta^\tau = \frac{E[D^{\tau,p}]}{E[D^\tau]} \tag{2.6}$$

as fill rate, $\beta^\tau \in [0, 1]$.

Let $D^{\tau,f} = D^{\tau} - D^{\tau,p}$ be the demand in periods of length τ that could not have been served in time. Then obviously

$$\beta^{\tau} = 1 - \frac{E[D^{\tau,f}]}{E[D^{\tau}]} \tag{2.7}$$

is equivalent to (2.6).

Proposition 1. *Let τ_b be a basic demand arrival interval observing i.i.d. demand occurrences of size D. Then (2.6) has an identical result for any value τ that is an integer multiple of τ_b, and $\beta = \beta^{\tau}$ for all $\tau = c \cdot \tau_b, c \in \mathbb{Z}^+$.*

Proof. Since the D are i.i.d. by definition, $E[D^{\tau}] = \tau \cdot E[D]$ holds for any $\tau = c \cdot \tau_b, c \in \mathbb{Z}^+$. The demand D^p that is satisfied in the basic period τ_b clearly depends on D. Furthermore, inventory systems in static environments typically observe a cyclic change of state, meaning that D^p also depends on the observed micro period of the characteristic cycle. Hence, the D^p are neither identically nor independently distributed. We may, however, always find a *macro* τ_m for which $D^{\tau_m,p}$ are i.i.d. due to the cyclic structure of the inventory system's behavior. Therefore, $E[D^{\tau,p}] = \tau \cdot E[D^p]$ also holds for any $\tau = c \cdot \tau_b, c \in \mathbb{Z}^+$ as long as the intervals of length τ do not systematically begin in specific micro periods, and thus equally represent all subperiods of τ_m. □

Thus, we have

$$\beta = \frac{E[D^{\tau,p}]}{E[D^{\tau}]} = \frac{\tau \cdot E[D^p]}{\tau \cdot E[D]} = \frac{E[D^p]}{E[D]}$$

for any $\tau = c \cdot \tau_b, c \in \mathbb{Z}^+$.

Note that the unit-oriented service fill rate is frequently referred to as β-service level, especially in the European literature.

Analogous to the ready rate that we discussed in the previous section, we also distinguish between the theoretical mean fill rate β described above and the *realized* fill rate β^*. For this purpose, let $IS^{\tau *}$ be a real or simulated inventory system that observes demand occurrences $d^{\tau,i*} \in D^*$ and fulfils demand $d^{\tau,p,i*} \in D^{\tau,p*}$ over time. Furthermore, let $\mathcal{I}$ be a set of observed time intervals. Then

$$\beta^* = \frac{\mu_{D^{\tau,p*}}}{\mu_{D^{\tau*}}}, \quad \mu_{D^{\tau,p*}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} d^{\tau,p,i*}, \quad \mu_{D^{\tau*}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} d^{\tau,i*} \tag{2.8}$$

is an unbiased estimator for β.

Besides the fill rate as defined here, the order fill rate is discussed in the literature as the percentage of customer orders that have been served in time. As already mentioned in Sect. 2.3.2.1, we consider this a special case of the ready rate, where τ is the replenishment order interarrival time and thus may also be a random variable.

Larsen and Thorstenson (2008) compare the latter service metric (defined as order fill rate) and the (volume) fill rate.

Just like the ready rate, the average fill rate may differ for certain subperiods of the review or replenishment cycle. In this context, Haehling von Lanzenauer and Hamidi-Noori (1986) and Haehling von Lanzenauer (1988) discuss the ready rate based on fiscal periods and propose a model that requires meeting the corresponding service level in every period of their fiscal-based scheme, not just the periods' average.

In this context, please also refer to the study of Zipkin (1986a) for a discussion of mathematical properties of the average number of backorders.

2.3.2.3 Time-Weighted Fill Rate

Definition 9 (Time-weighted fill rate). Let IS be an inventory system that observes random demand D^τ in time intervals of length τ (as in Definition 8), and let $D^{\tau,b}$ be a random variable denoting the customer demand that is outstanding due to insufficient stock (number of backorders). Then

$$\gamma = \gamma^\tau \quad 1 - \frac{E[D^{\tau,b}]}{E[D^\tau]} \tag{2.9}$$

is called the time-weighted fill rate, $\gamma \in [-\infty, 1]$.

This unit- and time-oriented service level is often referred to as γ-service level, especially in the European literature. Under the same assumptions as for β and using analogous argumentation, we can consider γ as independent from the underlying time interval τ. As for the ready rate and (classical) fill rate, we distinguish the theoretically expected time-weighted fill rate and the *realized* time-weighted fill rate. The latter is defined in the same manner as the fill rate, where we regard the number of backorders instead of the delayed or non-delayed demand. We will therefore not go into a detailed description here.

The interpretation of this metric is difficult because of two properties. First, γ is not bound to the left and may therefore not be interpreted as a fraction of any kind. Second, two performance aspects of an inventory system are inseparably mixed: the fill rate, and the customer waiting time.

As an illustration of the first problem, consider the following example. Let IS^* be an assemble-to-order system with fixed demand D, reorder cycle r, order quantity $r \cdot D$ and lead time $l = 0$ for simplicity. Then in each micro period of the order cycle, an amount D is backlogged and finally fulfilled in the order period. The expected backlog per period is thus given as:

$$E[D^b] = \frac{1}{r} \cdot \sum_{i=1}^{r} i \cdot D = \frac{D}{r} \cdot \sum_{i=1}^{r} i = \frac{D}{r} \cdot \frac{r \cdot (r+1)}{2} = \frac{D \cdot (r+1)}{2}.$$

The time-weighted fill rate for this system is given as follows:

$$\gamma = 1 - \frac{\frac{D \cdot (r+1)}{2}}{r \cdot D} = \frac{1-r}{2}$$

Thus, we have $\lim_{r \to \infty} \gamma = -\infty$, or more generally, γ converges with $-\infty$ in the (theoretical) case that an inventory system observes positive demand and does not deliver at all. Clearly, the time-weighted fill rate does not describe a relative frequency and may not be interpreted as probability. The metric must therefore be viewed critically, and we recommend considering the customer waiting times to analyze the time span that an item or order has to wait until it can be delivered.

2.3.2.4 Customer Waiting Times

Definition 10 (Customer waiting times per order). Let IS be an inventory system that completely serves a random number of orders O^τ in an interval τ, and let O_w^τ be the (complying) random number of orders that are completely served by IS after a delay of exactly w periods. Then we define

$$W^O: \quad P\{W^O = w\} = p_w^O = \frac{E[O_w^\tau]}{E[O^\tau]} \quad \forall \quad w \in W^O. \tag{2.10}$$

as the customer waiting time distribution for complete order fulfillment, where we follow the concept of frequency probability.

Let IS^* be a real or simulated inventory system with an underlying and probably unknown W^O. Let $o_w^{\tau *}$ be the number of orders in time interval τ that have been served with a delay of w periods. Then we have

$$p_w^{O*} = \frac{o_w^{\tau *}}{\sum_{w=0}^{\infty} o_w^{\tau *}}$$

as unbiased estimator for p_w^O, where p_w^{O*} describes the *realized* waiting time frequencies.

Analogously to the waiting times per customer order, we define the waiting times per order unit.

Definition 11 (Customer waiting times per part). Let IS be an inventory system that serves a random number of order units (parts) V^τ in an interval τ and let V_w^τ be the (complying) random number of parts that are completely served by IS after a delay of exactly w periods. Then we define

$$W^V: \quad P\{W^V = w\} = p_w^V = \frac{E[V_w^\tau]}{E[V^\tau]} \quad \forall \quad w \in W^V. \tag{2.11}$$

as the customer waiting time distribution per part delivered.

The corresponding *realized* waiting time frequencies p_w^{V*} are determined in the same manner as described for p_w^{O*}, where $v_w^{\tau*}$, the number of parts delivered with delay w in time interval τ, replaces the number of orders $o_w^{\tau*}$.

http://www.springer.com/978-3-642-20478-4

Periodic Review Inventory Systems
Performance Analysis and Optimization of Inventory Systems within Supply Chains
Wensing, Th.
2011, XI, 151 p. 1 illus., Softcover
ISBN: 978-3-642-20478-4