1 Introduction

1.1 Uncertainty

The realm of manufacturing is replete with instances of uncertainties in job processing times, machine statuses (up or down), demand fluctuations, due dates of jobs, and job priorities. These uncertainties stem from the inability to predict with sufficient accuracy information about product demand, job processing times, and occurrences such as unexpected machine breakdowns and arrivals/cancellations of orders. Although highly efficient forecasting methods are currently available, product demand errors invariably occur in the production system. Process variability is another significant factor that introduces variations and uncertainty into the manufacturing process.

Uncertainty is inarguably an undesirable factor in the manufacturing process because it does not give production managers complete control over the manufacturing process. Some of the ill effects of these uncertainties include system instability, excess inventory, customer dissatisfaction because due dates are not met, and, more important, loss of revenue. Recent advances in manufacturing management techniques, such as agile manufacturing, have made variability an important design criterion in order to ensure predictability and dependability of production systems. *Agility* is the ability of a system to thrive and prosper in an environment of constant and unpredictable change.

1.2 Uncertainty in Scheduling

As true as it would be with any other field within manufacturing, the uncertainty factor is of considerable importance in production scheduling. Scheduling is a decision-making process that plays an important role in most manufacturing as well as in most information-processing environments. From a manufacturing perspective, a *scheduling problem* is primarily the determination of the starting times of the jobs waiting to be processed, on a single machine or multiple machines (resources) for the objective of optimizing an appropriate performance measure of interest. The randomness in the scheduling

system could be due to varying processing times, machine breakdowns, and incomplete information about customer due dates, among other things. *Deterministic scheduling* involves solving a scheduling problem with the objective of optimizing a performance measure of interest when the various parameters, *viz.,* job processing times, due dates, release dates, and so on, are known with certainty. On the other hand, *stochastic scheduling* deals with problems when at least one of these parameters is not known with certainty. Scheduling under stochasticity is relatively more complex and difficult than its deterministic counterpart.

1.3 Modeling Uncertainty in Scheduling

As stated earlier, uncertainty has a major impact on scheduling decisions. Conventionally, in stochastic scheduling, the uncertain or variable scheduling parameters are modeled as random variables, and researchers endeavor to optimize a performance measure of interest that is suitable to the problem at hand. In a majority of the work, the means and variances or the distributions of the random variables are assumed to be known *a priori*. The ultimate goal of the stochastic analysis is then to find the sequence that has the "best" statistical distribution. Knowing such a distribution will enable the management to plan for capacity and quote delivery dates in a manner that achieves set target service levels and higher customer satisfaction. However, finding the distribution of a scheduling criterion is extremely complex and, at times, practically impossible. Hence, researchers resort to more modest and practically viable criteria. The objective could then be to optimize some function of the performance measure of interest. The performance measure is also a random variable because it is a function of the input variables, which are given to be random. Predominantly, this function of the output performance measure is its expectation: that is, the goal is to optimize (minimize or maximize) the expectation of the performance measure.

The reason for such an approach can be surmised easily from the fact that computing or formulating the expectation function is relatively easier and less complex than computing or formulating any other function of the random variable, for example, its variance. In addition, optimization becomes arduous and may even become impossible with the incorporation of the variance function. Furthermore, determining the variance of a performance measure is highly complicated and laborious and is not straightforward for most of the commonly used performance measures (e.g., makespan, tardiness,) in scheduling. Hence, a preponderance of the work in stochastic scheduling has dealt with optimizing the expected value of a performance measure. To cite a simple example, while scheduling jobs with random processing times on a single machine with completion time as the performance measure, the predominant motive is to minimize the total expected completion time of all the jobs. By focusing only on the

expected value and ignoring the variance of the objective, the scheduling problem becomes purely deterministic, and the significant ramifications of schedule variability are neglected. However, in many practical cases, a scheduler would prefer to have a stable schedule with minimum variance over a schedule that has lower expected value and unknown (and possibly higher) variance.

1.4 Significance of Variance in Scheduling

As mentioned earlier, it is important to consider the issue of the variance of a performance measure in scheduling problems. To illustrate the significance of variance by means of a very simple example, consider four jobs waiting to be processed on a single machine with the objective of minimizing the total completion time (total flow time). The job processing times (in some specified units) are random with known means and variances, as given in Table 1.1.

Conventionally, the approach in tackling this problem would be to minimize the total expected completion time by sequencing the jobs using the shortest expected processing time (SEPT) policy. Hence the optimal SEPT sequence is 3-4-1-2. If we let C_j be the completion time of job j, then the resulting expectation and variance of the total completion time are $E\left[\sum C_j\right] = 256$ and $\text{Var}\left[\sum C_j\right] = 357$.

However, by scheduling the jobs alternatively, say, in the 4-3-1-2 sequence, we have $E\left[\sum C_j\right] = 258$ and $\text{Var}\left[\sum C_j\right] = 217$. Schedule 2 possesses a slightly higher expected value but has a considerably lower variance than the SEPT schedule. This is illustrated in Figure 1.1, assuming that $\sum C_j$ for the two schedules follow a normal distribution.

If the manufacturer prefers to deliver all the jobs by a particular date, say, $d = 280$, we then can analyze the probability with which the deadline will be met using the two schedules in Figure 1.1:

Schedule 1 (SEPT schedule): $\mu_1 = 256$ and $\sigma_1^2 = 357$

$$\Pr\left[\sum C_j \le 280\right] = \Pr[Z \le 1.272] = 0.898 = 89.8\%$$

Schedule 2: $\mu_2 = 258$ and $\sigma_2^2 = 217$

$$\Pr\left[\sum C_j \le 280\right] = \Pr[Z \le 1.493] = 0.9324 = 93.24\%$$

(Z is the standard normal variable with mean 0 and variance 1.)

Table 1.1. *Total Completion Time Example*

N	1	2	3	4
μ	35	40	20	22
σ^2	8	5	20	0

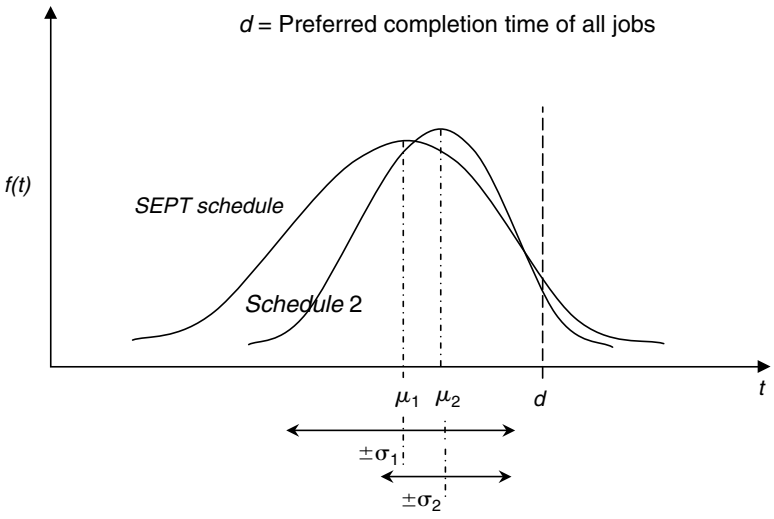

Figure 1.1. Representation of normally distributed completion time variables.

The probability of meeting the deadline is higher if the second schedule is employed. Apparently, it becomes imperative to determine a sequence that is "good" in terms of both expectation and variance. We would not have been able to identify Schedule 2, which, in fact, turned out to be practically better, had we not included the variance and, instead, had focused only on the expectation.

Soroush and Fredenhall (1984) recognized the importance of considering both mean and variance in scheduling while studying the impact of random processing times on the earliness and tardiness costs for scheduling jobs on a single machine. The significance of variance is not necessarily confined to production systems because it has been addressed in the context of other fields as well, such as telecommunication networks (Shayman and Gaucherand, 2001), financial investment decision problems (Chue and Nagasawa, 1999), and media planning/selection (de Kluyver, 1980). In communication networks, *sequential testing* is the process of identifying the defective component from a set of components that is attributed as a root cause of a failure. There is a random cost associated with the testing of each component, and traditionally, the objective has been to find a sequence that minimizes the average (or expected) sum of testing costs. Shayman and Gaucherand (2001) assert that the network scheduler should use a risk-sensitive optimality criterion to correctly model the system by taking into consideration the risk factor associated with the variance of the total cost. In optimal investment decision problems, the most desirable objective for a decision maker is to maximize the expected profit resulting from the investment as well as minimize the investment risk (variation in profit). In media scheduling problems, the objective is to select and schedule different media options that

would maximize the return (e.g., gross profit, gross audience) given a set of media options, budget, and other relevant data. In addition, it is necessary to recognize the effect of variance and control schedule variance by incorporating an effective risk-return analysis (de Kluyver, 1980; de Kluyver and Baird, 1984).

1.5 Multiobjective or Multicriteria Stochastic Scheduling

Multiobjective or multicriteria optimization, especially in the field of scheduling, has always been an interesting and challenging topic for researchers. A scheduler's endeavor, from a practical point of view, is to optimize one or more objectives of interest simultaneously and achieve a trade-off solution, which is commonly referred to as a *Pareto-optimal solution*. The solution to a multiobjective optimization problem is considered to be Pareto-optimal if there are no other solutions that are better in satisfying all the objectives simultaneously. That is, there can be other solutions that are better in satisfying one or several objectives, but they must be worse than the Pareto-optimal solution in satisfying the remaining objectives.

Deterministic multiobjective scheduling has been addressed rather extensively in the literature, and some of the work reported can be found in Wassenhove and Gelders (1980), Lin (1983), Nelson et al. (1986), Daniels and Chambers (1990), Sarin and Hariharan (2000), and Sarin and Prakash (2004), among many others. T'Kindt and Billaut (2005) have recently edited a special issue of the *European Journal of Operational Research* devoted to this topic.

On the stochastic front, Forst (1995) addressed the problem of minimizing the sum of the expected total weighted tardiness and the expected total weighted flow time for the single-machine and m-machine flow-shop scheduling problems. He proved that an optimal sequence is obtained by sequencing the jobs in increasing stochastic order of their processing times. The job processing times are assumed to be independent random variables, and the jobs have a common random due date. Lin and Lee (1995) considered a single-machine scheduling problem with known distributions of random processing times and due dates. The objective was to determine a schedule that minimized a secondary criterion subject to a primary criterion that was held at its best value. They formulated three different models with completion times and lateness-related bicriteria objectives, and they provided algorithms for obtaining optimal solutions.

Few studies have been devoted to the stochastic analysis of a schedule as compared with its deterministic counterpart. Liu et al. (1992) dealt with a discounted Markov decision model to determine a schedule with optimal expectation and variance of a criterion. They discussed the difficulties involved in minimizing variance by Markovian models. They also formulated

a multiobjective nonlinear programming problem and presented an algorithm for determining a Pareto-optimal solution.

Lu et al. (1994) addressed the problem of reducing the mean and variance of cycle times in semiconductor manufacturing environments, which feature the characteristic reentrant process flows. In reentrant flows, lots repeatedly return to the same service stations for further processing at different stages of their production. Lu et al. introduce a new class of scheduling policies, called *fluctuation smoothing policies*, that achieve the best mean and variance of the cycle time. The effectiveness of these policies was demonstrated via simulation modeling of two semiconductor manufacturing plants. Kumar and Kumar (1994), subsequently, established through their work that these policies are stable for all stochastic reentrant lines under certain conditions.

It would be germane at this juncture to mull over the fact that the multiple objectives that the researchers considered in stochastic multicriteria scheduling are related predominantly to completion time and tardiness. From a problem-modeling perspective, the different scheduling parameters, *viz.,* job processing times, due dates, and so on, were primarily modeled as random variables with known distributions. The performance measures of interest, such as the total flow time or tardiness, which are in turn random variables and a function of the random variables, were optimized. More often than not, this function is the expectation. This seemed valid enough because consideration of other functions, such as the variance of completion time or lateness, would make the problem enormously complex and difficult to solve.

1.6 Variance of the Performance Measure: Other Production Systems

As we strive to understand the importance of variance from a scheduling perspective, it is also apposite to survey the variance-related research in other production control systems. This section briefly reviews the work done in computing the variance of the output measure in serial production lines operated using CONWIP and other systems. There is an abundance of information in this domain, and only an illustrative review is provided to underline the fact that variance is indeed a parameter worthy of consideration.

1.6.1 CONWIP Systems

A CONWIP line, or *con*stant *w*ork-*i*n-*p*rocess line, is a pull-based production system proposed by Spearman et al. (1990) (Figure 1.2). The output measures in a CONWIP line are, predominantly, throughput [or time between departures (TBD)] and flow time. Considerable research has been done on CONWIP systems to study and analyze the mean and variance of these measures. Spearman

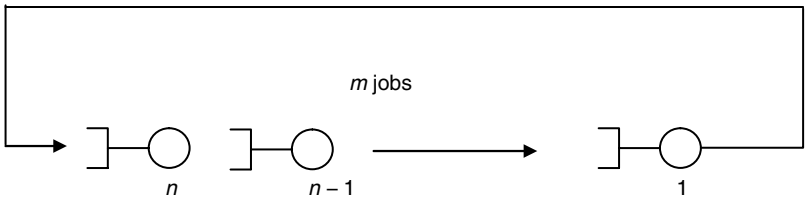

Figure 1.2. CONWIP schematic.

and Hopp (1991) developed an expression to estimate the throughput of a CONWIP manufacturing line subject to machine failures. They computed the throughput and average cycle time as a function of the work-in-progress (WIP) level.

In a subsequent paper, Duenyas et al. (1993) derived an approximation for the variance of the throughput of a CONWIP line with deterministic processing and random outages. Dar-el et al. (1998) focused on a CONWIP line to develop estimates for four important performance measures: the means and variances of the TBD and flow time. TBD is the inverse of the throughput rate.

1.6.2 Production Lines

It has been shown that the distribution of the output from a production line is asymptotically normal as a result of the central limit theorem. Hence, a majority of the work dealing with uncertain production lines, owing to random processing times or unreliable machines, had striven only to determine the expectation and variance of the output performance measure of interest. Knowing the mean and variance of the output gives the asymptotic distribution of the throughput, which could be used to derive other performance measures (e.g., meeting a customer due date) based on the probability of other events.

Hendricks (1992) analyzed the mean and variance of the output process of a serial production line of N machines with exponential processing time distributions and finite buffer capacities. Analytic expressions for the interdeparture distribution and the correlation structure of the output process were developed using a continuous-time Markov chain model.

Tan (1997) developed a closed-form expression for the variance of the throughput of an N-station production line with no intermediate buffers and time-dependent failures. Time-to-failure and time-to-repair distributions were assumed to be exponential, and the variance of the throughput was determined by modeling the process as an irreducible recurrent Markov process. In a subsequent paper, Tan (2000) determined the variance of the throughput of a production line with finite buffers by modeling the line as a discrete-time Markov chain.

1.7 Processing Time Variance in Scheduling

Variation in the processing times is a major factor or cause of uncertainty in scheduling, and the impact of variation in processing times on the efficiency of scheduling has been a subject of discussion in the literature for a long time. McKay et al. (1988) pointed out that the primary reason for poor applicability of scheduling theory in practice is its inability to properly account for extreme variations in processing times. This is primarily due to the ubiquitous attempt to use deterministic models in practical situations, which are highly stochastic. In the world of agile and lean manufacturing, effective scheduling under uncertainty has become a survival necessity for companies to meet committed shipping dates and effective utilization of available resources. Hence, it becomes imperative to devise the right scheduling strategies to employ in practice.

Dodin (1996) contends that the pseudodeterministic sequence that is obtained by sequencing tasks when all the activities are assumed to take their expected times does not fully reflect the goals of stochastic analysis of a schedule. Dodin further suggests using an alternative sequence determined using a ranking system based on *optimality indices* (OIs), defined as their respective probabilities of being the best. The OIs are computed using the strong dominance properties of distributions. Dodin conducted extensive simulations to analyze and compare the sequences obtained by the preceding two methods in order to determine which performs best. However, the results do not favor one method over the other and remain inconclusive. Portuogal and Trietsh (1998) also agreed with Dodin in stating that minimizing the expectation alone is not good enough for the scheduler. They introduced and defined two new sequences called *stochastically smallest* and *almost surely smallest sequences*. They analyzed these sequences along with Dodin's two sequences, under different scenarios, to find the sequence with the best distribution. They concluded that stochastically smallest and almost surely smallest should always be selected whenever they exist. However, stochastically smallest and almost surely smallest sequences do not always exist in practice, and even if they do exist, determining them is an arduous task. They argued that Dodin's sequence based on OIs does not take variability into account and suggested that a variance-reduction objective should be considered explicitly to attain optimal service levels while retaining the expected completion time.

Ayhan and Olsen (2000) considered scheduling on a single machine (server) that processes a number of different classes of items. They proposed two heuristic procedures for scheduling on such a multiclass single server that minimize the throughput time variance and the outer percentiles of the throughput time.

1.8 Analytic Evaluation of Expectation and Variance of a Performance Measure

In addition to these observations, a telling inference that can be made is that analytic expressions for the expectation and variance of simpler performance measures such as the total completion time on a single machine can be readily formulated and computed, but expressions for other complex measures related to tardiness are relatively complex and not as straightforward to evaluate. This task is even harder for measures related to makespan in multimachine scheduling environments such as parallel machines, flow shops, or job shops. Besides, no comprehensive work is reported in the literature that strives to address this issue. Hence, our primary motive in this book is to present a comprehensive analysis in order to devise methodologies and derive closed-form expressions (wherever possible) to determine the expectation and variance of various performance measures for different scheduling environments. The scheduling environments considered in our analysis include

1. Scheduling on a single machine
2. Permutation flow shops with unlimited intermediate storage
3. Job shops with unlimited intermediate storage
4. Scheduling on identical machines in parallel.

This analysis is contingent on the facts that the schedule is given *a priori* and that it is necessary to ascertain the expectation and variance of the given performance measure for that given schedule. The position of each job is, therefore, known with certainty from the schedule. The randomness in the scheduling process is due to variable processing times with known means and variances. All other parameters, such as the job due dates and weights, are assumed to be deterministic. In some cases, it might also be necessary to know the processing time distributions, and those instances are cited appropriately. Our interest, then, is to develop analytic expressions or methodologies to compute the parameters under consideration, *viz.,* expectation and variance of the objective function value. The analysis does not involve optimization, and it is a vital exercise in modeling the performance of the scheduling system. However, it is worth mentioning that this endeavor will only trigger and enable variance considerations in schedule optimization. This knowledge would provide valuable insights in improving the performance of a schedule. A scheduler would be in a better position to base his or her decisions knowing the variability of the schedule and appropriately striking a balance between the expected value and variance. In addition, these expressions and methodologies can be incorporated in various scheduling algorithms (and available software packages) to determine efficient schedules in terms of both the expectation and variance.

The different models considered for our analysis include

1. **Single-machine models.** The different performance measures considered for the single-machine case can be classified under two different categories, namely,
   a. Completion-time-based measures
   b. Tardiness-based measures.

   The various completion-time-based measures are
   a. Total completion time (total flow time)
   b. Total weighted completion time
   c. Total weighted discounted completion time.

   The various tardiness-based measures are
   a. Total tardiness
   b. Total weighted tardiness
   c. Total number of tardy jobs
   d. Total weighted number of tardy jobs
   e. Mean lateness
   f. Maximum lateness.
2. **Parallel-machine models.** For parallel machines, both preemption and no-preemption cases are considered. The performance measures are makespan and total completion time for the no-preemption case, and only makespan for the preemption case.
3. **Flow shops.** The objective is the makespan of a permutation flow shop with unlimited intermediate storage.
4. **Job shops.** The objective is to evaluate the makespan for a classic job shop with unlimited intermediate storage.

However, before we present our work, we would like to study and understand in detail the available literature in the field of stochastic scheduling, where researchers have attempted to consider both expectation and variance in their analyses. Our focus is only on the randomness owing to processing time uncertainty, which, as mentioned earlier, is a significant cause of stochasticity in scheduling.

1.9 Organization of the Book

The organization of this book is as follows: In this introductory chapter we have focused on the impact of uncertainty in scheduling and the need for efficient modeling of stochastic scheduling problems, as well as the need to devise effective scheduling strategies to counter the impact of uncertainty (or variability). We have specifically highlighted the prevalence of variability in job processing times and elicited the issue of neglecting variance in schedule optimization and the significance of considering variance. The need for a comprehensive