Lecture Notes 4AB00

Version: January 31, 2018

Edited by: Ivo Adan,
Erjen Lefeber,
Sasha Pogromsky

Technische Universiteit Eindhoven University of Technology
Department of Mechanical Engineering
Eindhoven University of Technology
P.O. Box 513, 5600 MB, Eindhoven, The Netherlands

Contents

Contents 3

1 Introduction 5

2 Basics 9

2.1 Permutations and combinations . . . 9

2.2 Standard series . . . 9

3 Probability Models 11

3.1 Basic ingredients . . . 11

3.2 Conditional probabilities . . . 15

3.3 Discrete random variables . . . 19

3.4 Continuous random variables . . . 29

3.5 Central limit theorem . . . 37

3.6 Joint random variables . . . 38

3.7 Conditioning . . . 41

4 Manufacturing Models 47

4.1 Terminology . . . 49

4.2 Key performance measures . . . 49

4.3 Capacity, flow time and WIP . . . 50

4.4 Little's law . . . 54

4.5 Variability . . . 57

4.6 Process time variability . . . 58

4.6.1 Natural variability . . . 59

4.6.2 Preemptive outages . . . 59

4.6.3 Non-Preemptive outages . . . 60

4.6.4 Rework . . . 61

4.7 Flow variability . . . 61

4.8 Variability interactions - Queueing . . . 64

4.9 Zero-buffer model . . . 66

4.10 Finite-buffer model . . . 67

4.11 Single machine station 69
4.12 Multi machine station 72
4.13 Serial production lines 75
4.14 Batching 79

Appendices **85**

A KIVA model **87**

B Zero-buffer model **89**

C Finite-buffer model **91**

D Single machine model **93**

E Multi machine model **95**

F Serial production line **97**

1

Introduction

In the course Analysis of Manufacturing Systems we study the behavior of manufacturing systems. Understanding its behavior and basic principles is critical to manufacturing managers and engineers trying (*i*) to develop and control new systems and processes, or (*ii*) to improve existing systems. The importance of manufacturing management should not be underestimated: the success of a company strongly depends on the effectiveness of its management. One might say that the future of the Dutch manufacturing industry, which constitutes about 17% of the Dutch gross national product and employs about 15% of the Dutch work force, depends on how well manufacturing managers and engineers are able to exploit the newest developments in information and manufacturing technology, also referred to as Smart Industry or the fourth industrial revolution.

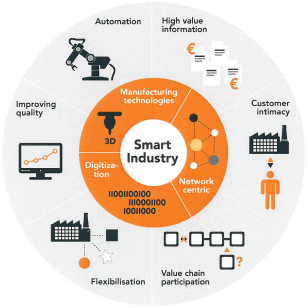


Figure 1.1: Smart Industry – Dutch industry fit for the future

The aim of this course is to develop understanding of manufacturing operations, with a focus on the material flow through the plant and key performance indicators, such as throughput, equipment efficiency, investment in equipment material, work in process and so on. Clearly, managers and engineers need to rely on their "manufacturing intuition" in order to fully understand the consequences of the decisions they make on system design, control and maintenance.

This course is based on the book Factory Physics [2]. This book provides the basis for manufacturing science, by offering (*i*) a clear description of the basic manufacturing principles (i.e., the physical laws underlying manufacturing operations), (*ii*) understanding and intuition about system behavior, and (*iii*) a unified modeling framework to facilitate synthesis of complex manufacturing systems.

As we will see, a crucial and disrupting element in manufacturing operations is variability. Theories that effectively describe the sources of variability and their interaction, are Probability Theory and Queueing Theory. So, not surprisingly, Factory Physics is firmly grounded on these theories, since a good probabilistic intuition is a powerful tool for the manager and engineer. This also explains why the present course consists of two main parts:

- **Manufacturing models.**
  This part is devoted to Manufacturing Science. It is based on the book Factory Physics [2],

Figure 1.2: Kiva warehouse system (source: http://www.kivasystems.com)

explaining the basic models and principles in manufacturing operations. The goal of this part is to develop, through analytical manufacturing models, understanding of the behavior and basic principles of manufacturing operations, with a focus on the material flow through the plant.

- **Probability models.**
  As mentioned above, a disrupting element in manufacturing operations is variability, a phenomenon that can be effectively described by Probability Theory. Hence, this part treats elementary concepts in Probability Theory, including probability models, conditional probabilities, discrete and continuous random variables, expectation, central limit theorem and so on. It is based on the book Understanding Probability [5], which puts emphasis on developing probabilistic intuition and which is full of challenging examples. The goal of this part is to develop basic skills in formulating and analyzing probability models.

Detailed models of real-life manufacturing systems are often too complicated for analytical treatment. A powerful, and in manufacturing practice, a very popular tool to analyse complicated real-life models is discrete-event computer simulation. Therefore, this course also offers an introduction to the use of simulation models, for which we will employ the simulation language χ3.0. We would like to emphasize that simulation is not only a tool to study complex systems, but it is also ideal to develop probabilistic intuition. Most of the challenging problems in [5] can be "solved" or verified by simulation and famous results such as the central limit theorem can be demonstrated "in action" by using simulation. So the third part of this course consists of:

- **Simulation models.**
  The simulation language χ3.0 is used as vehicle to demonstrate simulation modeling and analysis. We believe that simulation modeling should be learned by doing. Hence, this part is based on self-study. To support self-study, an online tutorial for χ3.0 is available, and during the lectures, many examples of χ3.0 models will be presented. The goal of this part is to obtain some hands-on experience in developing and using simulation models, through the language χ3.0.

We conclude the introduction by an illustrative example in the area of automated warehousing.

Example 1.1. (Kiva systems) A new technology in warehousing is mobile shelf-based order picking, pioneered by Kiva Systems. The items are stored on movable storage racks, see Figure 1.2. The racks, containing items ordered by a customer, are automatically retrieved from the storage area and transported to an order picker by robots. These robots are small autonomous drive units that can carry a rack (by moving under and then lifting the rack). The benefits of such automated warehousing systems are, for example, high throughput capability, flexibility and scalability. By adding more robots, the throughput capacity of handling additional customer orders can be increased in a relatively short time span. The question is: How many robots are needed to achieve a certain target throughput capacity?

To illustrate how this question can be answered, we consider an order picker, with an average pick time of 3 minutes per rack. When the items for a customer order are picked, the robot stores the rack

at some place in the storage area and then retrieves the next rack from another place and brings it to the order picker. The average time required by a robot to store and retrieve a rack from the storage area is 15 minutes. Clearly, the variability of the storage and retrieval times will be high, since the racks can be located anywhere, and intensive traffic of robots will cause unpredictable congestion in the storage area. On arrival at the pick station, the robot joins the queue of robots waiting for the order picker, and once its items are picked, the robot returns to the storage area and the cycle repeats. This storage and retrieval process is illustrated in Figure 1.3.

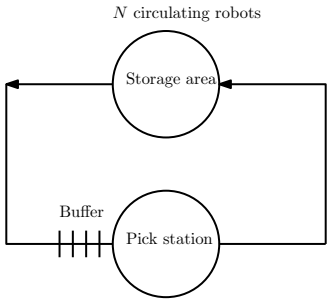


Figure 1.3: Storage and retrieval cycle of robots

Suppose that management aims at a throughput capacity of 15 racks per hour per picker, and wants to determine the minimal number of robots that should be assigned to each of the order pickers. To do so, we can use the χ3.0 simulation model, part of which is listed below (the whole listing is in Appendix A) to estimate the throughput *TH* for any given number of robots *N* assigned to the order picker, and then generate the graph in Figure 1.4. The χ3.0 simulation model is based on (simplifying) assumptions on the storage and picking process and it ignores congestion in the storage area.

```
model KIVA():
    int N = 1;
    real la = 4.0, mu = 20.0;
    chan pod a, b, c;

    run G(a, N),
        unwind j in range(N):
            S(a, b, la)
        end,
        B(b, c), P(c, a, mu, 10000)
end
```

The curve in Figure 1.4 shows the relationship between number of robots and throughput, from which we can conclude that, to achieve a throughput of picking 15 shelves per hour, at least 6 robots are required per picker, and that the benefit of adding more robots sharply decreases (and merely results in additional robots waiting at the picker).

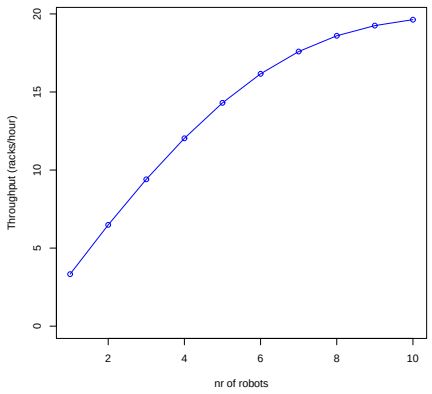


Figure 1.4: Throughput *TH* as function of the number of robots *N*

The results in Figure 1.4 are obtained by computer simulation, but under certain model assumptions, it is also possible to derive a closed formula for the throughput TH as a function of the number of robots N, where $\frac{1}{\mu}$ is the average pick time and $\frac{1}{\lambda}$ the average storage and retrieval time of a robot. This formula explicitly shows how the throughput depends on the model parameters N, μ and λ.

$$TH = \mu \left(1 - \frac{\left(\frac{\mu}{\lambda}\right)^N}{\sum_{i=1}^{N} \left(\frac{\mu}{\lambda}\right)^i} \right).$$

Of course, at this point, we do not ask you to understand the above $\chi 3.0$ simulation model nor the formula for TH, but during this course we intend to teach you how models and formulas, like the ones above, can be derived, where these models and formulas do apply and where not, and what they learn us about the behavior of manufacturing operations.

2

Basics

In this chapter we briefly summarize some basic results, useful for studying probability models (see also the Appendix in [5]).

2.1 Permutations and combinations

The total number of ways you can arrange the letters A, B, C and D is 24, since for the first position you have 4 possibilities, for the second 3 and so on, so the total number is $4 \cdot 3 \cdot 2 \cdot 1 = 24$. In general, the total number of ways n different objects can be ordered in a sequence is $n \cdot (n-1) \cdots 2 \cdot 1$. This number is abbreviated as $n!$, so

$$n! = 1 \cdot 2 \cdots (n-1) \cdot n,$$

where by convention, $0! = 1$. Another question is in how many ways can you choose three letters from A, B, C and D? Two answer this question, you first count the number of ordered sequences of three letters, which is $4 \cdot 3 \cdot 2$. Then, to obtain the number of three letter combinations, you have to divide this number by $3!$, since the order of the three letter sequence is not relevant. In general, the total number of ways you can choose k different objects (irrespective of their order) from n different objects is equal to

$$\frac{n \cdot (n-1) \cdots (n-k+1)}{k!} = \frac{n!}{k!(n-k)!}.$$

This number is abbreviated as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

which is also referred to as binomial coefficient (since it is the coefficient of x^k in $(1+x)^n$), and thus counts the number of (unordered) combinations of k objects out of a set of n objects.

2.2 Standard series

In this section we briefly mention some standard series, that are used in many probability models. The first one is the sum $1 + 2 + \cdots + n$. This sum is equal to n (the number of terms) times their average $(1+n)/2$, so for any $n \geq 0$,

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

The next sum is $1 + x + x^2 + \cdots + x^n$. To calculate this one, we set

$$s = 1 + x + x^2 + \cdots + x^n.$$

Multiplying s by x gives

$$xs = x + x^2 + \cdots + x^{n+1}$$

and then subtracting xs from s yields

$$(1-x)s = 1 - x^{n+1}.$$

So we can conclude that

$$1 + x + x^2 + \cdots + x^n = \frac{1 - x^{n+1}}{1-x}, \tag{2.1}$$

where x should be not equal to 1 (in which case the right hand side is equal to n). We now consider the infinite sum

$$1 + x + x^2 + \cdots = \sum_{i=0}^{\infty} x^i,$$

valid for $|x| < 1$ (since, otherwise, the infinite sum does not exist). Taking n to infinity in (2.1) yields

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}, \quad |x| < 1. \tag{2.2}$$

The above series is also referred to as the geometric series. Based on (2.2) several variations of the geometric series can be calculated, such as for example, when we have an extra i,

$$\sum_{i=0}^{\infty} i x^i = \frac{x}{(1-x)^2}.$$

This follows from

$$\sum_{i=0}^{\infty} i x^i = x \sum_{i=0}^{\infty} i x^{i-1} = x \sum_{i=0}^{\infty} \frac{d}{dx} x^i = x \frac{d}{dx} \sum_{i=0}^{\infty} x^i = x \frac{d}{dx} \frac{1}{1-x} = \frac{x}{(1-x)^2}.$$

Similarly we have

$$\sum_{i=0}^{\infty} i^2 x^i = \sum_{i=0}^{\infty} i(i-1)x^i + \sum_{i=0}^{\infty} i x^i = x^2 \sum_{i=0}^{\infty} i(i-1)x^{i-2} + x \sum_{i=0}^{\infty} i x^{i-1} == \frac{2x^2}{(1-x)^3} + \frac{x}{(1-x)^2} = \frac{x^2+x}{(1-x)^3}.$$

The next series is the exponential series or Taylor expansion of e^x, given by

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}.$$

Note that, for example,

$$\begin{aligned}
\sum_{i=0}^{\infty} i \frac{x^i}{i!} &= \sum_{i=1}^{\infty} \frac{x^i}{(i-1)!} = x \sum_{i=1}^{\infty} \frac{x^{i-1}}{(i-1)!} = x e^x, \\
\sum_{i=0}^{\infty} i(i-1) \frac{x^i}{i!} &= \sum_{i=2}^{\infty} \frac{x^i}{(i-2)!} = x^2 \sum_{i=2}^{\infty} \frac{x^{i-2}}{(i-2)!} = x^2 e^x, \\
\sum_{i=0}^{\infty} i^2 \frac{x^i}{i!} &= \sum_{i=0}^{\infty} i(i-1) \frac{x^i}{i!} + \sum_{i=0}^{\infty} i \frac{x^i}{i!} = (x^2 + x) e^x.
\end{aligned}$$

3

Probability Models

In this chapter we briefly review the main concepts of probability models.

3.1 Basic ingredients

Over time several definitions of probability have been proposed. Probabilities can be defined before an experiment. For example, in the experiment of throwing a die, it is reasonable to assume that all outcomes are equally likely. Alternatively, probabilities can be defined after an experiment, as the relative frequencies of the outcomes by repeating the experiment many times. Sometimes it is not possible to repeat experiments (think of predicting the weather), in which case probabilities can be estimated subjectively. Fortunately, there is a unifying and satisfying approach to define probability models, due the Russian mathematician Andrej Kolmogrov, where probabilities are defined as a function on the subsets of the space of all possible outcomes, called the sample space, which should obey some appealing rules.

The ingredients of a probability model are as follows:

- The sample space S (often denoted by Ω) which is the set of all possible outcomes;
- Events are subsets of the possible outcomes in S;
- New events can be obtained by taking the union, intersection, complement (and so on) of events.

The sample space S can be discrete or continuous, as shown in the following examples.

Example 3.1. Examples of the sample space:

- Flipping a coin, then the outcomes are Head (H) and Tail (T), so $S = \{H, T\}$.
- Rolling a die, then the outcomes are the number of points the die turns up with, so $S = \{1, 2, \dots, 6\}$.
- Rolling a die twice, in which case the outcomes are the points of both dies, $S = \{(i, j), i, j = 1, 2, \dots, 6\}$.
- Process time realizations on a machine, which can be any nonnegative (real) number, so $S = [0, \infty)$.
- Sometimes process times have a fixed off-set, say 5 (minutes), in which case $S = [5, \infty)$.
- The number of machine failures during a shift, $S = \{0, 1, 2, \dots\}$.

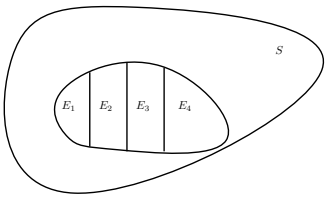


Figure 3.1: Disjoint events E_1, E_2, E_3 and E_4

- Throwing darts where the target is the unit disk, $S = \{(x, y), \sqrt{x^2 + y^2} \leq 1\}$.

The events are all subsets of the sample space S (though in case of a continuous sample space S one should be careful and exclude "weird" subsets).

Example 3.2. Examples of events of the sample spaces mentioned in the previous example are:

- Flipping a coin, $E = \emptyset$ (i.e., the empty set), $E = \{H\}$, $E = \{T\}$ and $E = \{H, T\} = S$ (these are all possible events).
- Rolling a die, $E = \{1, 2\}$, $E = \{1, 3, 5\}$ (the odd outcomes).
- Rolling a die twice, $E = \{(1, 2), (3, 4), (5, 6)\}$, $E = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)\}$ (the first die turned up with 1).
- Process times, $E = (0, 1)$, $E = (1, \infty)$
- Process times with offset, $E = (10, 15)$
- Number of failures, $E = \{3, 4, ...\}$
- Throwing darts, $E = \{(x, y), 0 \leq x \leq \frac{1}{4}, 0 \leq y \leq \frac{1}{4}\}$.

The other ingredient of a probability model are, of course, probabilities. These are defined as a function of events, and this function should obey the following elementary rules.

For each event E there is a number $\mathrm{P}(E)$ (the probability that event E occurs) such that:

- $0 \leq \mathrm{P}(E) \leq 1$ (a probability should be a number between 0 and 1).
- $\mathrm{P}(S) = 1$ (the probability that something happens is 1).
- If the events $E_1, E_2, ...$ are mutually disjoint (the sets E_i have nothing in common), then

$$\mathrm{P}(E_1 \cup E_2 \cup \cdots) = \mathrm{P}(E_1) + \mathrm{P}(E_2) + \cdots$$

(probability of the union of disjoint events is the sum of the probabilities of these events, see Figure 3.1).

Example 3.3.

- Flipping a fair coin, $\mathrm{P}(\{H\}) = \mathrm{P}(\{T\}) = \frac{1}{2}$, $\mathrm{P}(\{H, T\}) = 1$.
- Rolling a die, $\mathrm{P}(E)$ is the number of outcomes in E divided the total number of outcomes (which is 6), so $\mathrm{P}(\{1\}) = \frac{1}{6}$, $\mathrm{P}(\{1, 2\}) = \mathrm{P}(\{1\}) + \mathrm{P}(\{2\}) = \frac{1}{3}$ and so on.
- Rolling a die twice, $\mathrm{P}(i, j) = \frac{1}{36}$ for $i, j = 1, ... 6$.
- Throwing darts, $\mathrm{P}(E)$ is the area of E, divided by area of unit disk (which is π).

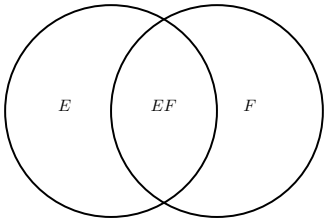


Figure 3.2: Probability of union of two events E and F

In case the sample space S is discrete, so $S = \{s_1, s_2, ...\}$, then we can assign probabilities $\mathrm{P}(s)$ to each $s \in S$, which should be between 0 and 1 and add up to 1. Then

$$\mathrm{P}(E) = \text{sum of the probabilities of the outcomes in the set } E = \sum_{s \in E} \mathrm{P}(s). \tag{3.1}$$

If S is a finite set, $S = \{s_1, s_2, \dots, S_N$, and all outcomes are equally likely, so $\mathrm{P}(s_i) = 1/N$, then (3.1) reduces to

$$\mathrm{P}(E) = \frac{N(E)}{N}$$

where $N(E)$ is the number of outcomes in the set E. An example is the experiment of rolling two dice.

Based on the ingredients of a probability model, the following properties of probabilities can be mathematically derived. Note that these properties all correspond to our intuition.

Property 3.1.

- $\mathrm{P}(\emptyset) = 0$ (probability of nothing happening is 0).
- If event F includes E, then its probability is greater,

$$\mathrm{P}(E) \leq \mathrm{P}(F) \quad \text{if } E \subset F.$$

- If finitely many $E_1, E_2, \dots, E_n$ are mutually disjoint (they have nothing in common), then

$$\mathrm{P}(E_1 \cup E_2 \cup \cdots \cup E_n) = \mathrm{P}(E_1) + \mathrm{P}(E_2) + \cdots + \mathrm{P}(E_n).$$

- If the event E^c is the complement of E (all outcomes except the ones in E), so $E^c = S \setminus E$, then

$$\mathrm{P}(E^c) = 1 - \mathrm{P}(E).$$

- For the union of two events (so event E or F occurs),

$$\mathrm{P}(E \cup F) = \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(EF),$$

  where $EF = E \cap F$ is the intersection of both events (so event E and F occur), see Figure 3.2.

It is remarkable that based on the simple ingredients of a probability model the frequency interpretation of probabilities can be derived, i.e., the probability of event E can be estimated as the fraction of times that E happened in a large number of (identical) experiments.

Property 3.2. (Law of large numbers)
If an experiment is repeated an unlimited number of times, and if the experiments are independent of each other, then the fraction of times event E occurs converges with probability 1 to $\mathrm{P}(E)$.

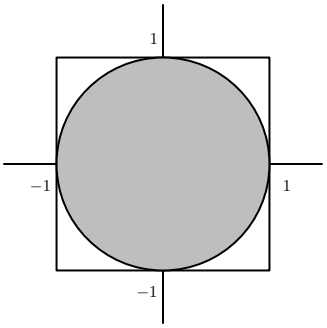


Figure 3.3: Randomly sampling points from the square $[-1,1] \times [-1,1]$

For example, if we flip a fair coin an unlimited number of times, then an outcome s of this experiment is an infinite sequence of Heads and Tails, such as

$$s = (H, T, T, H, H, H, T, ...).$$

Then, if $K_n(s)$ is the number if Heads appearing in the first n flips of realization s, we can conclude that, according to the law of large numbers,

$$\lim_{n\to\infty} \frac{K_n(s)}{n} = \frac{1}{2}$$

with probability 1.

The method of computer simulation is based on this law.

Example 3.4. The number π can be estimated by the following experiment. We randomly sample a point (x, y) from the square $[-1,1] \times [-1,1]$, see Fig. 3.3 and the experiment is successful if (x, y) falls in unit circle. Then by the law of large numbers:

$$\frac{\text{number of successful experiments}}{\text{total number of experiments}} \approx \frac{\text{area unit circle}}{\text{area square}} = \frac{\pi}{4}.$$

Exercise 1. (*Problem 7.3 [5]***)** Four black socks and five white socks lie mixed up in a drawer. You grab two socks at random from the drawer. What is the probability of having grabbed one black sock and one white sock?

Exercise 2. (*Problem 7.5 [5]***)** Two players A and B each roll one die. The absolute difference of the outcomes is computed. Player A wins if the difference is 0, 1, or 2; otherwise, player B wins. What is the probability that player A wins?

Exercise 3. (*Problem 7.7 [5]***)** You have four mathematics books, three physics books and two chemistry books. The books are put in random order on a bookshelf. What is the probability of having the books ordered per subject on the bookshelf?

Exercise 4. (*Problem 7.29 [5]***)** A small transport company has two vehicles, a truck and a van. The truck is used 75% of the time. Both vehicles are used 30% of the time and neither of the vehicles is used for 10% of the time. What is the probability that the van is used on any given day?

Exercise 5. (*Problem 7.33 [5]***)** You roll a fair die six times in a row. What is the probability that all of the six face values will appear? What is the probability that one or more sixes will appear?

3.2 Conditional probabilities

Conditional probabilities are an intuitive concept: intuitively the conditional probability $\mathrm{P}(E|F)$ is the probability that event E occurs, given that we are being told that event F occurs. Suppose we do an experiment n times, and we observe that event F occurs r times together with E, and s times without E. Then the relative frequency of occurring E and F is

$$f_n(EF) = \frac{r}{n}$$

and for F,

$$f_n(F) = \frac{r+s}{n}.$$

The frequency $f_n(E|F)$ of event E in those experiments in which also F occurred, is given by

$$f_n(E|F) = \frac{r}{r+s} = \frac{f_n(EF)}{f_n(F)}.$$

Since $f_n(EF) \approx \mathrm{P}(EF)$ and $f_n(F) \approx \mathrm{P}(F)$, this motivates the definition of the conditional probability as

$$\mathrm{P}(E|F) = \frac{\mathrm{P}(EF)}{\mathrm{P}(F)},$$

We call $\mathrm{P}(E|F)$ the conditional probability of E *given* F.

Example 3.5. Suppose we roll a die twice, so $S = \{(i,j), i,j = 1,2,\dots,6\}$ and $\mathrm{P}(\{i,j\}) = \frac{1}{36}$.

- Given that $i = 4$, what is probability that $j = 2$? Note that $F = \{(4,j), j = 1,2,\dots,6\}$ and $E = \{(i,2), i, = 1,2,\dots,6\}$. Hence $\mathrm{P}(F) = \frac{1}{6}$ and $\mathrm{P}(EF) = \mathrm{P}(\{(4,2)\}) = \frac{1}{36}$, so

  $$\mathrm{P}(E|F) = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}.$$

  This corresponds to our intuition: knowing that the first roll is 4 does not tell us anything about the outcome of the second roll.

- Given that one of the dice turned up with 6, what is the probability that the other one turned up with 6 as well? Now $F = \{(6,j), j = 1,2,\dots,5\} \cup \{(i,6), i = 1,2,\dots,5\} \cup \{(6,6)\}$ and $EF = \{(6,6)\}$. Hence

  $$\mathrm{P}(E|F) = \frac{\frac{1}{36}}{\frac{11}{36}} = \frac{1}{11}.$$

Example 3.6. Consider the experiment of throwing darts on a unit disk, so $S = \{(x,y), \sqrt{x^2+y^2} \leq 1\}$ and $\mathrm{P}(E)$ is the area of E, divided by area of unit disk, which is π . Given that the outcome is in the right half of the unit disk, what is the probability that its distance to the origin is greater that $\frac{1}{2}$? For this conditional probability we get (see Figure 3.4)

$$\mathrm{P}(\text{distance of } (x,y) \text{ to } 0 > \tfrac{1}{2} | x > 0) = \frac{\frac{1}{2}\pi - \frac{1}{8}\pi}{\frac{1}{2}\pi} = \frac{3}{4}.$$

The formula for conditional probability can be rewritten in the intuitive form

$$\mathrm{P}(EF) = \mathrm{P}(E|F)\mathrm{P}(F).$$

This form, which is also known as the product rule, is frequently used to calculate probabilities, since in many situations, it simplifies calculations or the conditional probabilities are directly given.

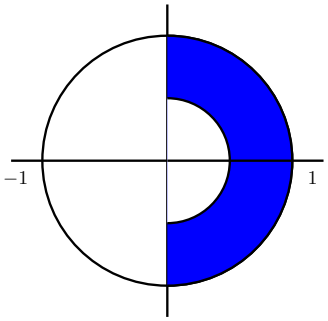


Figure 3.4: Conditional probability that distance $> \frac{1}{2}$ given that outcome is in right half of unit disk

Example 3.7. (Champions League [5]) Eight soccer teams reached the quarter finales, two teams from each of the countries England, Germany, Italy and Spain. The matches are determined by drawing lots, one by one. What is the probability that the two teams from the same country play against each other in all four matches? This probability can be calculated by counting all outcomes of the lottery. The total number of outcomes of the lottery, in which only teams from the same country play against each other, is $2^4 \cdot 4!$ (for example, $(E_1, E_2; G_2, G_1; I_2, I_1; S_1, S_2)$ is a possible outcome). The total number of outcomes is $8!$. Hence,

$$\mathrm{P}(\text{Only teams of the same country play against each other}) = \frac{2^4 \cdot 4!}{8!}. \tag{3.2}$$

It is, however, easier to use conditional probabilities. Imagine that the first lot is drawn. Then, given that the first lot is drawn, the probability that the second lot is the other team of the same country is $\frac{1}{7}$. Then the third lot is drawn. Given that first two lots are of the same country, the probability that the fourth one is of the same country as the third one is $\frac{1}{5}$, and so on. This immediately gives

$$\mathrm{P}(\text{Only teams of the same country play against each other}) = \frac{1}{7} \cdot \frac{1}{5} \cdot \frac{1}{3},$$

which is indeed the same answer as (3.2).

The special case that

$$\mathrm{P}(E|F) = \mathrm{P}(E),$$

means that knowledge about the occurrence of event F has no effect on the probability of E. In other words, event E is independent of F. Substitution of $\mathrm{P}(E|F) = \mathrm{P}(EF)/\mathrm{P}(F)$ leads to the conclusion that events E and F are independent if

$$\mathrm{P}(EF) = \mathrm{P}(E)\mathrm{P}(F).$$

Example 3.8. (Independent experiments) The outcomes of n independent experiments can be modeled by the sample space $S = S_1 \times S_2 \times \cdots \times S_n = \{(s_1, \ldots, s_n),\ s_1 \in S_1, \ldots, s_n \in S_n\}$, where S_i is the sample space of experiment i. To reflect that the experiments are independent, the probability of the event $E = E_1 \times E_2 \times \cdots \times E_n = \{(s_1, \ldots, s_n),\ s_1 \in E_1, \ldots, s_n \in E_n\}$ is defined as

$$\mathrm{P}(E_1 \times E_2 \times \cdots \times E_n) = \mathrm{P}(E_1) \cdots \mathrm{P}(E_n).$$

Example 3.9. (Coin tossing) Two fair coins are tossed. E is the event that the first coin turns up with H and F is the event that both coins turn up with the same outcome. Are the events E and F independent? Straightforward calculations yield

$$\begin{aligned}
\mathrm{P}(EF) &= \mathrm{P}(\{(H,H)\}) = \frac{1}{4}, \\
\mathrm{P}(E) &= \mathrm{P}(\{(H,H),(H,T)\}) = \mathrm{P}(\{(H,H)\}) + \mathrm{P}(\{(H,T)\}) = \frac{1}{2}, \\
\mathrm{P}(F) &= \mathrm{P}(\{(H,H),(T,T)\}) = \frac{1}{2},
\end{aligned}$$

so $\mathrm{P}(EF) = \mathrm{P}(E)\mathrm{P}(F)$, and thus E and F are independent.

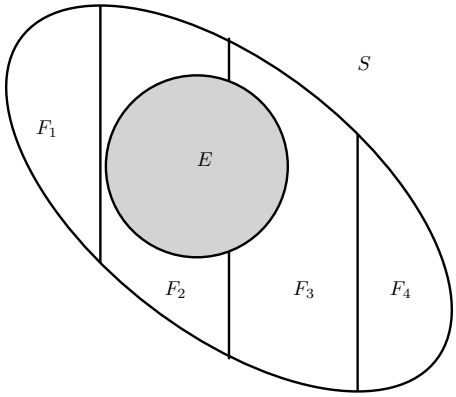


Figure 3.5: Illustration of relation (3.3)

Example 3.10. (Rolling a die twice) For the experiment of rolling a die twice, consider the events $F = \{(4, j), j = 1, 2, \dots, 6\}$ and $E = \{(i, j), i + j = 6\} = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$. Are E and F independent? We have $\mathrm{P}(E) = \frac{5}{36}$, $\mathrm{P}(F) = \frac{1}{6}$ and $\mathrm{P}(EF) = \mathrm{P}(\{(4, 2)\}) = \frac{1}{36} \neq \mathrm{P}(E)\mathrm{P}(F)$, so E and F are not independent. For $E = \{(i, j), i + j = 7\}$ it follows that E and F are independent (check!).

The calculation of unconditional probabilities can be often be simplified by using exclusive events and conditional probabilities, as expressed in the following result (also known as the law of conditional probability). Suppose that the events $F_1, \dots, F_n$ are disjoint and together form the whole sample space S, so (see also Figure 3.5)

$$S = F_1 \cup F_2 \cup \cdots \cup F_n.$$

Then for any event E we have

$$\mathrm{P}(E) = \mathrm{P}(E|F_1)\mathrm{P}(F_1) + \cdots + \mathrm{P}(E|F_n)\mathrm{P}(F_n). \tag{3.3}$$

The following examples illustrate the use of relation (3.3).

Example 3.11. First a fair die is rolled, and then a fair coin is tossed for that many times turned up on the die. What is the probability that heads will not appear? Let E be the event that no heads appears. To calculate $\mathrm{P}(E)$, we condition on the number that turns up on the die. Let F_k be the event that k turns up. Clearly $\mathrm{P}(F_k) = \frac{1}{6}$ and $\mathrm{P}(E|F_k) = \left(\frac{1}{2}\right)^k$. Hence,

$$\begin{aligned} \mathrm{P}(E) &= \mathrm{P}(E|F_1)\mathrm{P}(F_1) + \mathrm{P}(E|F_2)\mathrm{P}(F_2) + \cdots + \mathrm{P}(E|E_6)\mathrm{P}(F_6) \\ &= \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{4} \cdot \frac{1}{6} + \cdots + \frac{1}{64} \cdot \frac{1}{6} \\ &= 0.1641. \end{aligned}$$

Example 3.12. (Batch production) Items are produced in batches on a machine, the size of which ranges from 1 to n. The probability that the batch size is k is $\frac{1}{n}$ for $k = 1, \dots, n$. Immediately after production, the quality of each item in the batch is checked, and with probability p the quality of the item meets the threshold, and otherwise, it is scrapped. What is probability that all items in the batch meet the quality threshold? Let E be the event that none of the items in the batch is scrapped, and let F_k be the event that the batch size is k. Then $\mathrm{P}(F_k) = \frac{1}{n}$ and $\mathrm{P}(E|F_k) = p^k$. Hence,

$$\mathrm{P}(E) = \sum_{k=1}^{n} \mathrm{P}(E|F_k)\mathrm{P}(F_k) = \frac{1}{n} \sum_{k=1}^{n} p^k = \frac{p}{n} \frac{1 - p^n}{1 - p}.$$

Example 3.13. (Tour de France [5]) The Tour de France will take place from July 1 through July 23 with 180 cyclists participating. What is the probability that none of them will have same birthdays during the tournament? Let E be the event that none of them have the same birthday. To calculate $\mathrm{P}(E)$ we first condition on the number of cyclists having birthday during the tournament. Let F_k be

the event that exactly k of them have birthday during the tournament. Then $F_0, \dots, F_{180}$ are disjoint events. To calculate $\mathrm{P}(F_k)$, note that you can choose $\binom{180}{k}$ different groups of size k from the 180 cyclists, and the probability each one in the group has his birthday during the tournament of 23 days is $\left(\frac{23}{365}\right)^k$, while the probability that the other $180-k$ cyclists do not have their birthday during the tournament is $\left(\frac{365-23}{365}\right)^{180-k}$. Hence,

$$\mathrm{P}(F_k) = \binom{180}{k}\left(\frac{23}{365}\right)^k\left(\frac{365-23}{365}\right)^{180-k}, \quad k = 0, 1, \dots, 180,$$

Clearly $\mathrm{P}(E|F_0) = 1$, and

$$\mathrm{P}(E|F_k) = \frac{23}{23}\cdot\frac{22}{23}\cdots\frac{32-k+1}{23}, \quad k = 1, \dots, 23$$

and $\mathrm{P}(E|F_k) = 0$ for $k \geq 24$. Hence

$$\mathrm{P}(E) = \sum_{k=0}^{180}\mathrm{P}(E|F_k)\mathrm{P}(F_k) = \sum_{k=0}^{23}\mathrm{P}(E|F_k)\mathrm{P}(F_k) = 0.8841.$$

In some applications, the probabilities $\mathrm{P}(E)$, $\mathrm{P}(F)$ and $\mathrm{P}(E|F)$ are given, while we are interested in $\mathrm{P}(F|E)$. This probability can then be calculated as follows

$$\mathrm{P}((F|E) = \frac{\mathrm{P}(EF)}{\mathrm{P}(E)} = \frac{\mathrm{P}(E|F)\mathrm{P}(F)}{\mathrm{P}(E)}.$$

This is also known as Bayes' rule (see Section 8.3 in [5]).

Example 3.14. The reliability of a test for a certain disease is 99%. This means that the probability that the outcome of the test is positive for a patient suffering form the disease, is 99%, and it is negative with probability 99% for a patient free from this disease. It is known that 0.1% of the population suffers from this disease. Suppose that, for a certain patient, the test is positive. What is the probability that this patient indeed suffers form this disease? Let E be the event that the test is positive, and F is the event that the patient has the disease. Then we need to calculate $\mathrm{P}(F|E)$. It is given that $\mathrm{P}(E|F) = \mathrm{P}(E^c|F^c) = 0.99$, $\mathrm{P}(F) = 0.001$ and thus $\mathrm{P}(E) = \mathrm{P}(E|F)\mathrm{P}(F) + \mathrm{P}(E|F^c)\mathrm{P}(F^c) = 0.99\cdot 0.001 + 0.01\cdot 0.999 = 0.01098$. Hence

$$\mathrm{P}(F|E) = \frac{\mathrm{P}(E|F)\mathrm{P}(F)}{\mathrm{P}(E)} = \frac{0.99\cdot 0.001}{0.01098} = 0.09.$$

Note that this number is much smaller than might have been suggested by the reliability of the test!

Exercise 6. (*Problem 8.3 [5]*) Every evening, two weather stations issue a weather forecast next day. The weather forecasts of the two stations are independent of eachother. On average, the weather forecast of station 1 is correct in 90% of the cases, irrespective of the weather type. This percentage is 80% for station 2. On a given day, station 1 predicts sunny weather for the next day, whereas station 2 predicts rain. What is the probability that the weather forecast of station 1 will be correct?

Exercise 7. (*Problem 8.5 [5]*) You simultaneously grab two balls at random from an urn containing two red balls, one blue ball and one green ball. What is the probability that you have grabbed two non-red balls given that you have grabbed at least one non-red ball? What is the probability that you have grabbed two non-red balls given that you have grabbed the green ball? Can you give an intuitive explanation of why the second probability is larger than the first one?

Exercise 8. (*Problem 8.17 [5]*) Two fair coins are tossed. Let A be the event that heads appears on the first coin and let B be the event that the coins display the same outcome. Are the events A and B independent?

Exercise 9. (*Problem 8.18 [5]*) You have two identical boxes in front of you. One of the boxes contains 10 balls numbered 1 to 10 and the other box contains 25 balls numbered 1 to 25. You choose at random one of the boxes and pick a ball at random from the chosen box. What is the probability of picking the ball with number 7 written on it?

Exercise 10. (*Problem 8.19 [5]*) A bag contains three coins. One coin is two-headed and the other two are normal. A coin is chosen at random from the bag and is tossed twice? What is the probability of getting two heads? If two heads appear, what is the inverse probability that the two-headed coin was chosen?

3.3 Discrete random variables

A random variable is a function that assigns a numerical value to each outcome of an experiment. More precisely, it is a real-valued function on the sample space S. Clearly, it is often more convenient to do calculations with random variables than with outcomes of an experiment. It is common to use capitals for random variables, such X and Y.

Example 3.15.

- X is the sum of the outcomes i and j of rolling a die twice, so $X = i + j$.
- Y is the maximum of the outcomes i and j of rolling a die twice, so $Y = \max(i, j)$.
- N is the number of flips of a coin until the first head.
- M is the number of times a machine fails during a working day.
- Z is the number of production orders that arrives during the week.

A discrete random variable X can only take (possibly infinitely many) discrete values, say $x_1, x_2, \ldots$, and the function

$$p_j = \mathrm{P}(X = x_j), \quad j = 1, 2, \ldots,$$

is called the probability mass function or probability distribution of X.

Example 3.16.

- If $X = i + j$ where i is the first and j the second roll of a die, then

$$\mathrm{P}(X = k) = \mathrm{P}(X = 14 - k) = \frac{k-1}{36}, \quad k = 2, \ldots, 7.$$

- If $Y = \max(i, j)$ where i is the first and j the second roll of a die, then

$$\mathrm{P}(Y = k) = \frac{2k-1}{36}, \quad k = 1, \ldots, 6.$$

- If N is the number of flips until first H, with $\mathrm{P}(H) = 1 - \mathrm{P}(T) = p$,

$$\mathrm{P}(N = n) = (1-p)^{n-1}p, \quad n = 1, 2, \ldots$$

For a random variable X with probability mass function $p_j = \mathrm{P}(X = x_j), j = 1, 2, ...$, its expected value (or expectation or first moment or "centre of probability mass") is defined as

$$\mathrm{E}[X] = \sum_{j=1}^{\infty} x_j p_j,$$

where we assume that the infinite sum exists. So the expected value of X is a weighted average of possible values of X. It is not the same, however, as the most probable value, nor is it restricted to possible values of X.

Example 3.17.

- If X is the number that turns up when rolling a die, then

$$\mathrm{E}[X] = \frac{1}{6} + \frac{2}{6} + \cdots + \frac{6}{6} = 3.5.$$

- If N is the number of flips until the first Head, with $\mathrm{P}(H) = 1 - \mathrm{P}(T) = p$, then

$$\mathrm{E}[N] = \sum_{n=1}^{\infty} n(1-p)^{n-1}p = \frac{1}{p}.$$

For example, by repeatedly rolling a die, the average value of the numbers that turn up, gets closer and closer to 3.5 as the number of rolls increases. This is the law of large numbers for the expected value. More general, if X_k is the outcome of the kth repetition of an experiment, then the average $\frac{1}{n}(X_1 + \cdots + X_n)$ over the first n repetitions converges with probability 1 to $E(X)$. Hence, the expected value $E(X)$ can thus be interpreted as the long run average of X.

Example 3.18. Let Y be the total number that shows up by rolling a die twice. Then

$$\mathrm{E}[Y] = \sum_{i=1}^{6}\sum_{j=1}^{6}(i+j)\frac{1}{36} = 7 = 2 \times 3.5,$$

so $\mathrm{E}[Y]$ is two times the expected value of a single roll. This is no coincidence, since by writing $Y = X_1 + X_2$ where X_1 is first roll and X_2 is second roll, we get

$$\begin{aligned}
\mathrm{E}[Y] &= \sum_{i=1}^{6}\sum_{j=1}^{6}(i+j)\mathrm{P}(X_1 = i, X_2 = j) \\
&= \sum_{i=1}^{6}\sum_{j=1}^{6} i\mathrm{P}(X_1 = i, X_2 = j) + \sum_{i=1}^{6}\sum_{j=1}^{6} j\mathrm{P}(X_1 = i, X_2 = j) \\
&= \sum_{i=1}^{6} i\sum_{j=1}^{6}\mathrm{P}(X_1 = i, X_2 = j) + \sum_{j=1}^{6} j\sum_{i=1}^{6}\mathrm{P}(X_1 = i, X_2 = j) \\
&= \sum_{i=1}^{6} i\mathrm{P}(X_1 = i) + \sum_{j=1}^{6} j\mathrm{P}(X_2 = j) \\
&= \mathrm{E}[X_1] + \mathrm{E}[X_2] \;=\; 7.
\end{aligned}$$

Example 3.18 shows that $\mathrm{E}[X_1 + X_2] = \mathrm{E}[X_1] + \mathrm{E}[X_2]$ where X_1 is the first roll of a die and X_2 the

second one. This property is true in general. For any two random variables X and Y we have

$$\begin{aligned}
\mathrm{E}[X+Y] &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(x_i+y_j)\mathrm{P}(X=x_i, Y=y_j) \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}x_i\mathrm{P}(X=x_i, Y=y_j) + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}y_j\mathrm{P}(X=x_i, Y=y_j) \\
&= \sum_{i=1}^{\infty}x_i\sum_{j=1}^{\infty}\mathrm{P}(X=x_i, Y=y_j) + \sum_{j=1}^{\infty}y_j\sum_{i=1}^{\infty}\mathrm{P}(X=x_i, Y=y_j) \\
&= \sum_{i=1}^{\infty}x_i\mathrm{P}(X=x_i) + \sum_{j=1}^{\infty}y_j\mathrm{P}(Y=y_j),
\end{aligned}$$

so

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

Hence the expectation of the sum is the sum of the expectations. More general, for any number of random variables $X_1, X_2, \ldots, X_n$,

$$\mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n].$$

Example 3.19.

- (Tour de France [5]) What is the expected number of joint birth days during the tournament? Let X_i be 1 if there is a joint birthday on day i of the tournament, and 0 otherwise. If none or exactly one of the cyclists has its birthday on day i, then day i is not a joint birth day, so $X_i = 0$. Hence,

$$\mathrm{P}(X_i=0) = 1 - \mathrm{P}(X_i=1) = \left(\frac{364}{365}\right)^{180} + 180\cdot\frac{1}{365}\left(\frac{364}{365}\right)^{179} = 0.912,$$

  so $\mathrm{E}[X_i] = 0\cdot\mathrm{P}(X_i=0) + 1\cdot\mathrm{P}(X_i=1) = \mathrm{P}(X_i=1)$, and the expected number of joint birthdays is

$$\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_{23}] = 23(1-0.912) = 2.02.$$

- (Tall children [5]) Suppose that n children of different lengths are placed in line at random. You start with the first child and then walk till the end of the line. When you encounter a taller child than seen so far, she will join you. Let X be the number that joins you. What is $E(X)$? Let X_i be 1 if child i in line joins you, and 0 otherwise. Then $X = X_1 + \cdots + X_n$. Child i joins you if she is the tallest among the first i children in line. Since the first i are ordered at random, the probability that the tallest is at place i is $\frac{1}{i}$. Hence

$$\mathrm{P}(X_i=1) = 1 - \mathrm{P}(X_i=0) = \frac{1}{i},$$

  and $\mathrm{E}[X_i] = 0\cdot\mathrm{P}(X_i=0) + 1\cdot\mathrm{P}(X_i=1) = \frac{1}{i}$, thus

$$\mathrm{E}[X] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n] = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

A convenient property is that for any function $g(x)$, the expectation of the random variable $Y = g(X)$ can be directly calculated as

$$\mathrm{E}[Y] = \mathrm{E}[g(X)] = \sum_{x_i} g(x_i)\mathrm{P}(X=x_i) = \sum_{x_i} g(x_i)p_i.$$

Example 3.20. Let N be the number of flips of a fair coin until the first Head. Then $\mathrm{P}(N = i) = \frac{1}{2^i}$ and $E(N) = 2$ (see Example 3.17). For $g(x) = x^2$, we get

$$\mathrm{E}\left[g(N)\right] = \mathrm{E}\left[N^2\right] = \sum_{i=1}^{\infty} i^2 \frac{1}{2^i} = 6.$$

In general, $\mathrm{E}\left[g(X)\right] \neq g(\mathrm{E}\left[X\right])$ (for the example above, we have $\mathrm{E}\left[N^2\right] = 6$ and $(\mathrm{E}\left[N\right])^2 = 2^2 = 4$). However, linear functions $g(x) = ax + b$ are an exception, since for any constants a and b,

$$\begin{aligned} \mathrm{E}\left[aX + b\right] &= \sum_{x_i} (ax_i + b)\mathrm{P}(X = x_i) \\ &= a\sum_{x_i} x_i \mathrm{P}(X = x_i) + b\sum_{x_i} \mathrm{P}(X = x_i) \\ &= a\mathrm{E}\left[X\right] + b. \end{aligned} \tag{3.4}$$

An important measure for the spread of the possible values of X is the variance, defined as the expected squared deviation from the mean,

$$\mathrm{Var}\left[X\right] = \mathrm{E}\left[(X - E(X))^2\right].$$

In many situations it is, however, more convenient to consider the standard deviation of X, which is the square root of the variance,

$$\sigma[X] = \sqrt{\mathrm{var}(X)}.$$

This quantity has the same units as $E(X)$. The variance $\mathrm{Var}\left[X\right]$ can be calculated as

$$\begin{aligned} \mathrm{Var}\left[X\right] &= \mathrm{E}\left[(X - \mathrm{E}\left[X\right])^2\right] \\ &= \sum_{x_i} (x_i - \mathrm{E}\left[X\right])^2 \mathrm{P}(X = x_i) \\ &= \sum_{x_i} \left(x_i^2 - 2\mathrm{E}\left[X\right] x_i + (\mathrm{E}\left[X\right])^2\right) \mathrm{P}(X = x_i) \\ &= \sum_{x_i} x_i^2 \mathrm{P}(X = x_i) - 2\mathrm{E}\left[X\right] \sum_{x_i} x_i \mathrm{P}(X = x_i) + (\mathrm{E}\left[X\right])^2 \sum_{x_i} \mathrm{P}(X = x_i) \\ &= \mathrm{E}\left[X^2\right] - 2\mathrm{E}\left[X\right]\mathrm{E}\left[X\right] + (\mathrm{E}\left[X\right])^2 \\ &= \mathrm{E}\left[X^2\right] - (\mathrm{E}\left[X\right])^2 \end{aligned}$$

and for $Y = aX + b$ we then get

$$\begin{aligned} \mathrm{Var}\left[aX + b\right] &= \sum_{j=1}^{\infty} (ax_j + b - (a\mathrm{E}\left[X\right] + b))^2 p_j \\ &= a^2 \sum_{j=1}^{\infty} (x_j - \mathrm{E}\left[X\right])^2 p_j \\ &= a^2 \mathrm{Var}\left[X\right]. \end{aligned}$$

Example 3.21. Let N be the number of flips of a fair coin until the first head. Then $\mathrm{P}(N = i) = \frac{1}{2^i}$, $E(N) = 2$ and $E(N^2) = 6$ (see Example 3.20), so

$$\mathrm{Var}\left[N\right] = \mathrm{E}\left[N^2\right] - (\mathrm{E}\left[N\right])^2 = 6 - 4 = 2.$$

For the number of tails hat appear before the first head, which is $N - 1$, we get

$$\mathrm{Var}\left[N - 1\right] = \mathrm{Var}\left[N\right] = 2.$$

We have seen that $\mathrm{E}[X+Y] = \mathrm{E}[X]+\mathrm{E}[Y]$ for any two random variables X and Y. However, usually $\mathrm{Var}[X+Y]$ is not equal to $\mathrm{Var}[X]+\mathrm{Var}[Y]$, though it is true for *independent* random random variables X and Y. This is shown below.

The random variables X and Y are independent if for all x, y, the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent, thus

$$\mathrm{P}(X \leq x, Y \leq y) = \mathrm{P}(X \leq x)\mathrm{P}(Y \leq y)$$

or equivalently, when X and Y are discrete random variables,

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y).$$

It is clear that when X and Y are independent, then so are the random variables $f(X)$ and $g(Y)$ for any functions $f(x)$ and $g(y)$. Further, if X and Y are independent, then

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i y_j \mathrm{P}(X = x_i, Y = y_j) \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i y_j \mathrm{P}(X = x_i)\mathrm{P}(Y = y_j) \\
&= \sum_{i=1}^{\infty} x_i \mathrm{P}(X = x_i) \sum_{j=1}^{\infty} y_j \mathrm{P}(Y = y_j) \\
&= \mathrm{E}[X]\,\mathrm{E}[Y],
\end{aligned}$$

and as a consequence,

$$\begin{aligned}
\mathrm{Var}[X+Y] &= \mathrm{E}\left[(X+Y)^2\right] - (\mathrm{E}[X]+\mathrm{E}[Y])^2 \\
&= \mathrm{E}\left[X^2+2XY+Y^2\right] - ((\mathrm{E}[X])^2 + 2\mathrm{E}[X]\,\mathrm{E}[Y] + (\mathrm{E}[Y])^2) \\
&= \mathrm{E}\left[X^2\right] + 2\mathrm{E}[XY] + \mathrm{E}\left[Y^2\right] - ((\mathrm{E}[X])^2 + 2\mathrm{E}[X]\,\mathrm{E}[Y] + (\mathrm{E}[Y])^2) \\
&= \mathrm{E}\left[X^2\right] + 2\mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}\left[Y^2\right] - ((\mathrm{E}[X])^2 + 2\mathrm{E}[X]\,\mathrm{E}[Y] + (\mathrm{E}[Y])^2) \\
&= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 + \mathrm{E}\left[Y^2\right] - (\mathrm{E}[Y])^2 \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y].
\end{aligned}$$

So the variance of the sum is the sum of the variances, provided the random variable are independent!

Suppose that the random variables $X_1, \ldots, X_n$ have the same probability distribution, so in particular

$$\mathrm{E}[X_1] = \cdots = \mathrm{E}[X_n], \quad \mathrm{Var}[X_1] = \cdots = \mathrm{Var}[X_n].$$

For the mean of the sum $X_1 + \cdots + X_n$ we get

$$\mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = n\mathrm{E}[X_1],$$

and if $X_1, \ldots, X_n$ are independent, then also

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = n\mathrm{Var}[X_1].$$

Hence, for the standard deviation of the sum of $X_1, \ldots, X_n$ we obtain

$$\sigma[X_1 + \cdots + X_n] = \sqrt{n}\sigma[X_1].$$

So the mean of $X_1 + \cdots + X_n$ grows linear in n, but its standard increases more slowly at a rate of $\sqrt{n}$ as n tends to infinity. The latter is due to cancelation of random errors (i.e., some of the X_i will be greater than their mean, whereas others will be smaller). This property is also known as the square-root law or variability pooling.

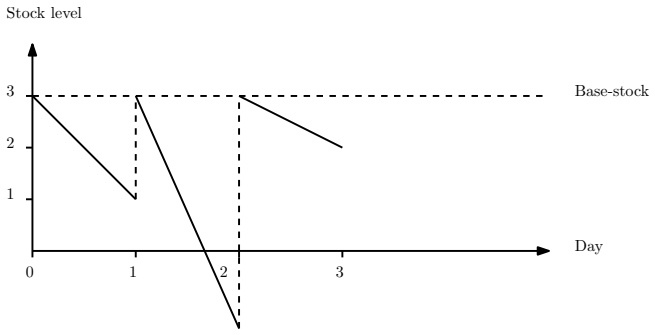


Figure 3.6: Base-stock policy with a base-stock level of 3 products

Example 3.22. (Variability pooling) A company offers 100 configurations of a product, all requiring component A. The mean demand per day for each configuration is 10 products, with a standard deviation of 3 products. To satisfy demand, the company uses a base-stock policy: all demand that occurs during the day is immediately ordered at the supplier and arrives at the beginning of the next day. Hence the stock level at the beginning of each day is always the same, and referred to as the base-stock level, see Figure 3.6.

The company sets the base-stock level equal to the mean demand plus a safety stock which is 2 times the standard deviation of the demand. The rational is that for this stock level, approximately 95% of the demand can be satisfied from stock. The company considers two options to keep products on stock: (i) Stock all product configurations, (ii) Stock only components A and assemble to order. What is the total stock level of components A for these two options? For option (i), the stock level for each configuration is $10 + 2 \cdot 3 = 16$. Hence, since we have 100 configurations, the total stock level of components A is $100 \cdot 16 = 1600$. For option (ii), note that the mean total demand for components A is $100 \cdot 10$ and the standard deviation of the total demand is $3\sqrt{100}$. Hence, the total stock level for components A is $100 \cdot 10 + 2 \cdot \sqrt{100} \cdot 3 = 1060$, which is a reduction of more than 30% in stock level compared to option (i)!

Below we list some important discrete random variables.

Bernoulli

A Bernoulli random variable X with success probability p, takes the values 0 and 1 with probability

$$\mathrm{P}(X = 1) = 1 - \mathrm{P}(X = 0) = p.$$

Then

$$\begin{aligned} \mathrm{E}[X] &= 0 \cdot \mathrm{P}(X = 0) + 1 \cdot \mathrm{P}(X = 1) = p \\ \mathrm{E}\left[X^2\right] &= 0^2 \cdot \mathrm{P}(X = 0) + 1^2 \cdot \mathrm{P}(X = 1) = p \end{aligned}$$

and

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = p - p^2 = p(1 - p).$$

Binomial

A Binomial random variable X is the number of successes in n independent Bernoulli trials $X_1, \ldots, X_n$, each with probability p of success,

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots n.$$

The Binomial distribution is shown in Figure 3.7 for $n = 20$ and $p = 0.25, 0.5, 0.65$ (which graph corresponds to which parameter p?). Since $X = X_1 + \cdots + X_n$, we get

$$\mathrm{E}[X] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = np, \quad \mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = np(1 - p).$$

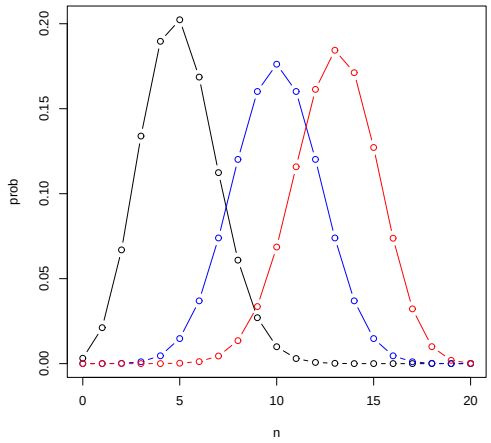


Figure 3.7: Binomial probability distribution with parameters $n = 20$ and $p = 0.25, 0.5, 0.65$

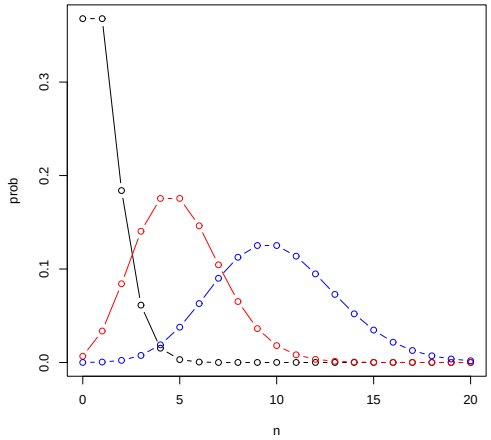


Figure 3.8: Poisson distribution with parameter $\lambda = 1, 5, 10$

Poisson

A Poisson random variable X with parameter $\lambda > 0$, has probability distribution

$$\mathrm{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, ...$$

The Poisson distribution is shown in Figure 3.8 for $\lambda = 1, 5, 10$ (which graph corresponds to which parameter λ?).

Then

$$\mathrm{E}[X] = \sum_{k=0}^{\infty} k\mathrm{P}(X = k) = \sum_{k=1}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$$

and accordingly,

$$\mathrm{E}[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)\mathrm{P}(X = k) = \sum_{k=2}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.$$

Hence, since $\mathrm{E}[X(X-1)] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]$, we get

$$\mathrm{E}\left[X^2\right] = \lambda^2 + \lambda$$

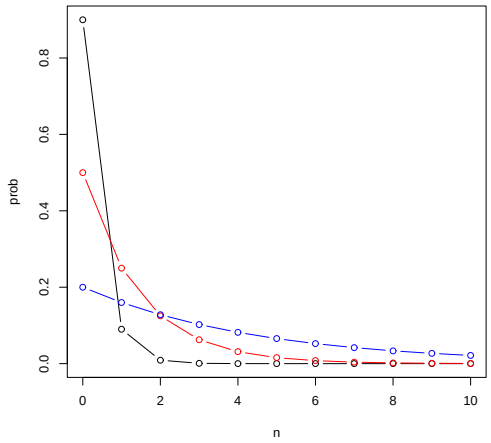


Figure 3.9: Geometric distribution with success probability $p = 0.9, 0.5, 0.2$

and thus

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = \lambda.$$

So, for a Poisson random variable, the variance is equal to the mean .

Hypergeometric

A Hypergeometric random variable X is the number of red balls in a random selection of n balls taken from an urn with R red balls and W white balls (so $n \leq R + W$),

$$\mathrm{P}(X = r) = \frac{\binom{R}{r}\binom{W}{n-r}}{\binom{R+W}{n}}, \quad r = 0, 1, \ldots, n.$$

Then, by writing $X = X_1 + \cdots + X_n$ where X_i indicates whether ball i from the selection of n balls is red or not, so $\mathrm{P}(X_i = 1) = 1 - \mathrm{P}(X_i) = \frac{R}{R+W}$, we get

$$\mathrm{E}[X] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = n\frac{R}{R+W}.$$

Geometric

A Geometric random variable X is the number Bernoulli trials till the first success, each trial with probability p of success, so

$$\mathrm{P}(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots$$

The Geometric distribution is shown in Figure 3.8 for $p = 0.9, 0.5, 0.2$ (which graph corresponds to which success probability p?).

Then

$$\mathrm{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = p\frac{1}{p^2} = \frac{1}{p},$$

and accordingly,

$$\mathrm{E}[X(X-1)] = \sum_{k=1}^{\infty} k(k-1)(1-p)^{k-1}p = p(1-p)\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} = p(1-p)\frac{2}{p^3} = \frac{2(1-p)}{p^2}.$$

Hence, since $\mathrm{E}[X(X-1)] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]$, we obtain

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

Negative binomial

A Negative binomial random variable X is the number Bernoulli trials till the rth success, each trial with probability p of success,

$$\mathrm{P}(X = k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r, \quad k = r, r+1, \ldots$$

We can write $X = X_1 + \cdots X_r$, where X_i are independent and Geometric with success probability p, so

$$\mathrm{E}[X] = \mathrm{E}[X_1] + \ldots + \mathrm{E}[X_r] = \frac{r}{p}, \quad \mathrm{Var}[X] = \mathrm{Var}[X_1] + \ldots \mathrm{Var}[X_r] = \frac{r(1-p)}{p^2}.$$

Remark 3.1. (Relation between Binomial and Poisson) The total number of successes X in n independent Bernoulli trials, each with success probability p, is Binomial distributed, so

$$\mathrm{P}(X = k) = \binom{n}{k}p^k(1-p)^{n-k}, \quad k = 0, 1, \ldots n.$$

If the number of trials n is large and the success probability p small, then

$$\mathrm{P}(X = k) = \binom{n}{k}p^k(1-p)^{n-k} \approx e^{-np}\frac{(np)^k}{k!}.$$

Hence, for large n and small p, the Binomial distribution of X is approximately equal to the Poisson distribution with parameter $\lambda = np$. This is demonstrated in Figure 3.10.

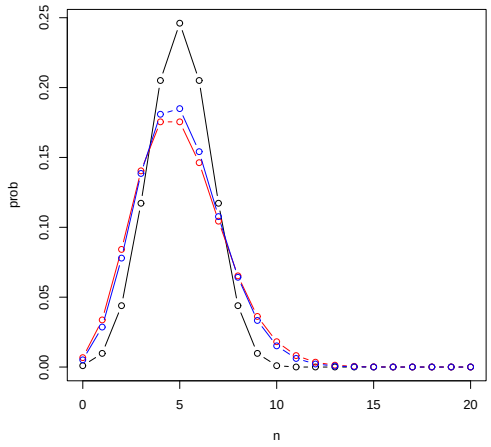


Figure 3.10: Binomial distribution with $n = 10$ and $p = \frac{1}{2}$ (black), Binomial distribution with $n = 50$ and $p = \frac{1}{10}$ and Poisson distribution with $\lambda = 5$ (red)

Example 3.23. (k out of n system) Consider a technical system composed of n identical components, and q is the probability that a component works (independent of the others), see Figure **??**.

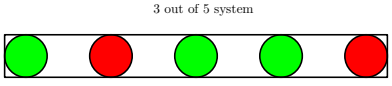


Figure 3.11: 3 out of 5 system

At least k components have to work for the system to work. What is the probability that the system works? Let X_i indicate whether component i works or not, so $\mathrm{P}(X_i = 1) = 1 - \mathrm{P}(X_i = 0) = q$, and

let X be the total number of components that work, $X = X_1 + \cdots + X_n$. Then X is Binomial with parameters n and q, and the probability Q that the system works is equal to

$$Q = \mathrm{P}(X \geq k) = \sum_{i=k}^{n} \mathrm{P}(X = i) = \sum_{i=k}^{n} \binom{n}{i} q^i (1-q)^{n-i}.$$

Example 3.24. (Coincidence) Two people, complete strangers to one another, both living in Eindhoven, meet each other in the train. Each has approximately 200 acquaintances in Eindhoven. What is the probability that these two strangers have an acquaintance in common? Eindhoven has (approximately) 200000 inhabitants. To find the desired probability, imagine that the acquaintances of the first stranger are colored red, and all other inhabitants of Eindhoven are colored white. Assuming that the acquaintances of the second strangers are randomly distributed among the population of Eindhoven, then the question can be translated into: What is the probability by randomly selecting 200 inhabitants that at least one of them is red? Let X denote the number of red inhabitants found in this random selection. Then X is Hypergeometric with parameters $n = R = 200$ and $W = 200000 - 200$. So

$$\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X = 0) = 1 - \frac{\binom{200000-200}{200}}{\binom{200000}{200}} = 1 - 0.818 = 0.182.$$

So, actually, it is quite likely to have an acquaintance in common!

Example 3.25. Consider a workstation perform tasks. The time to perform a task takes exactly 1 time unit. Each task is checked after completion. With probability p the task is done correctly. If not, the task has to be repeated till it is eventually correct. What is the distribution of the total time till the task has been done correctly? Let X denote the total time till correct completion, so it is the total *effective task time*. Clearly, X is Geometric with success probability p,

$$\mathrm{P}(X = k \text{ time units}) = (1-p)^{k-1} p, \quad k = 1, 2, \ldots$$

Hence, the mean and variance of X are given by

$$\mathrm{E}[X] = \frac{1}{p}, \quad \mathrm{Var}[X] = \frac{1-p}{p^2}.$$

Exercise 11. (*Problem 9.3 [5]***)** A bag contains three coins. One coin is two-headed and the other two are normal. A coin is chosen at random from the bag and is tossed twice. Let the random variable X denote the number of heads obtained. What is the probability mass function of X?

Exercise 12. (*Problem 9.7 [5]***)** You throw darts at a circular target on which two concentric circles of radius 1 cm and 3 cm are drawn. The target itself has a radius of 5 cm. You receive 15 points for hitting the target inside the smaller circle, 8 points for hitting the middle annular region, and 5 points for hitting the outer annular region. The probability of hitting the target at all is 0.75. If the dart hits the target, the hitting point is a completely random point on the target. Let the random variable X denote the number of points scored on a single throw of the dart. What is the expected value of X?

Exercise 13. (*Problem 9.15 [5]***)** What is the expected number of distinct birthdays within a randomly formed group of 100 persons? What is the expected number of children in a class with r children sharing a birthday with some child in another class with s children? Assume that the year has 365 days and that all possible birthdays are equally likely.

Exercise 14. (*Problem 9.17 [5]***)** What is the expected number of times that two consecutive numbers will show up in a lotto drawing of six different numbers from the numbers $1, 2, \ldots, 45$?

Exercise 15. (*Problem 9.33 [5]*) In the final of the World Series Baseball, two teams play a series consisting of at most seven games until one of the two teams has won four games. Two unevenly matched teams are pitted against each other and the probability that the weaker team will win any given game is equal to 0.45. Assuming that the results of the various games are independent of each other, calculate the probability of the weaker team winning the final. What are the expected value and the standard deviation of the number of games the final will take?

Exercise 16. (*Problem 9.42 [5]*) In European roulette the ball lands on one of the numbers $0, 1, \ldots, 36$ in every spin of the wheel. A gambler offers at even odds the bet that the house number 0 will come up once in every 25 spins of the wheel. What is the gambler's expected profit per dollar bet?

Exercise 17. (*Problem 9.44 [5]*) In the Lotto 6/45 six different numbers are drawn at random from the numbers $1, 2, \ldots, 45$. What are the probability mass functions of the largest number drawn and the smallest number drawn?

Exercise 18. (*Problem 9.46 [5]*) A fair coin is tossed until heads appears for the third time. Let the random variable X be the number of tails shown up to that point. What is the probability mass function of X? What are the expected value and standard deviation of X?

Exercise 19. (*Homework exercises 1*) Consider a bin with 10 balls: 5 are red, 3 are green and 2 are blue. You randomly pick 3 balls from the bin. What is the probability that each of these balls has a different color?

Exercise 20. (*Homework exercises 2*) Jan and Kees are in a group of 12 people that are seated randomly at a round table (with 12 seats). What is the probability that Jan and Kees are seated next to each other?

3.4 Continuous random variables

A continuous random variable X can take continuous values $x \in \mathbb{R}$. Its probability distribution function is of the form

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(y)dy, \tag{3.5}$$

where the function $f(x)$ satisfies

$$f(x) \geq 0 \text{ for all } x, \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

the function $f(x)$ is called the probability density of X. The probability that X takes a value in the interval $(a, b]$ can then be calculated as

$$\begin{aligned} \mathrm{P}(a < X \leq b) &= \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) \\ &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx \\ &= \int_{a}^{b} f(x)dx, \end{aligned}$$

and the probability that X takes a specific value b is equal to 0, since

$$\mathrm{P}(X = b) = \lim_{a \uparrow b} \mathrm{P}(a < X \leq b) = \lim_{a \uparrow b} \int_{a}^{b} f(x)dx = \int_{b}^{b} f(x)dx = 0.$$

The function $f(x)$ can be interpreted as the density of the probability mass, that is, we have for small $\Delta > 0$ that

$$\mathrm{P}(x < X \leq x + \Delta) = \int_x^{x+\Delta} f(y)dy \approx f(x)\Delta,$$

from which we can conclude (or from (3.5)) that $f(x)$ is the derivative of the distribution function $F(x)$,

$$f(x) = \frac{d}{dx}F(x).$$

Example 3.26. (Uniform)

- The Uniform random variable X on the interval $[a, b]$ has density

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{else.} \end{cases}$$

  and the distribution function is given by

$$F(x) = \begin{cases} 0 & x \leq 0, \\ x & 0 < x \leq 1, \\ 1 & x > 1. \end{cases}$$

- The Uniform random variable X on the interval $[a, b]$ has density

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{else.} \end{cases}$$

A simple recipe for finding the density is to first determine the distribution $F(x) = \mathrm{P}(X \leq x)$ (which is in many cases easier) and then differentiate $F(x)$.

Example 3.27.

- (Example 10.1 in [5]) Break a stick of length 1 at random point in two pieces and let X be the ratio of shorter piece and longer piece (so $0 < X < 1$). What is $F(x) = \mathrm{P}(X \leq x)$ for $0 \leq x \leq 1$? Let U be the point where the stick is broken. Then $X = \frac{U}{1-U}$ if $U < \frac{1}{2}$ and $X = \frac{1-U}{U}$ otherwise. Hence the event $X \leq x$ is valid for all U satisfying $0 \leq U \leq \frac{x}{1+x}$ or $\frac{1}{1+x} \leq U \leq 1$ (see Figure 3.12), so

$$F(x) = \mathrm{P}(X \leq x) = \mathrm{P}\left(U \leq \frac{x}{1+x}\right) + \mathrm{P}\left(U \geq \frac{1}{1+x}\right) = \frac{x}{1+x} + 1 - \frac{1}{1+x} = \frac{2x}{1+x}, \quad 0 < x < 1.$$

  The density of X is obtained by taking the derivative of $F(x)$, yielding

$$f(x) = \frac{d}{dx}F(x) = \frac{2}{(1+x)^2}, \quad 0 < x < 1.$$

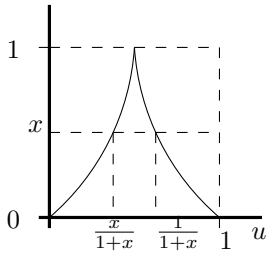
- Let $X = -\ln(U)/\lambda$ where U is random on $(0, 1)$. What is density of X? Clearly $X \geq 0$, so $F(x) = 0$ for $x < 0$. For $x \geq 0$ we get

$$F(x) = \mathrm{P}(X \leq x) = \mathrm{P}(-\ln(U)/\lambda \leq x) = \mathrm{P}(\ln(U) \geq -\lambda x) = \mathrm{P}(U \geq e^{-\lambda x}) = 1 - e^{-\lambda x}.$$

  Hence,

$$f(x) = \frac{d}{dx}F(x) = \lambda e^{-\lambda x}), \quad x \geq 0,$$

  and $f(x) = 0$ for $x < 0$. This distribution is called the Exponential distribution.



$$x = \begin{cases} \frac{u}{1-u} & u < \frac{1}{2} \\ \frac{1-u}{u} & u > \frac{1}{2} \end{cases}$$

Figure 3.12: Random variable $X = \frac{U}{1-U}$ if $U < \frac{1}{2}$ and $X = \frac{1-U}{U}$ if $U > \frac{1}{2}$ where U is random point on stick of unit length

- The random variable X is the distance to the origin 0 of a random point in a disk of radius r. What is density of X? Clearly $0 \leq X \leq r$, so $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > r$. In case $0 \leq x \leq r$ we get
$$F(x) = \mathrm{P}(X \leq x) = \frac{\pi x^2}{\pi r^2} = \frac{x^2}{r^2}.$$
Hence,
$$f(x) = \frac{d}{dx}F(x) = \frac{2x}{r^2}, \quad 0 \leq x \geq r,$$
and $f(x) = 0$ otherwise.

For a continuous random variable X with density $f(x)$, its expected value (or expectation or first moment) is defined as
$$\mathrm{E}[X] = \int_{-\infty}^{\infty} xf(x)dx$$
where we assume that the integral exists, and for any function $g(x)$, the expectation of $Y = g(X)$ can be calculated as
$$\mathrm{E}[Y] = \mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$
The variance of X is
$$\mathrm{Var}[X] = \mathrm{E}\left[(X - E(X))^2\right] = \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 f(x)dx,$$
and the standard deviation of X is the square root of the variance, so
$$\sigma(X) = \sqrt{\mathrm{Var}[X]}.$$

Example 3.28. The random variable X is the distance to the origin 0 of a random point in a disk of radius r. Then
$$\mathrm{E}[X] = \int_0^r xf(x)dx = \int_0^r x\frac{2x}{r^2}dx = \frac{2}{3}r$$
and
$$\mathrm{E}\left[X^2\right] = \int_0^r x^2 f(x)dx = \frac{1}{2}r^2.$$
Hence
$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = \frac{1}{18}r^2.$$

Below we list some important continuous random variables.

Uniform

A Uniform random variable X on $[a, b]$ has density
$$f(x) = \frac{1}{b-a}, \quad a < x < b,$$

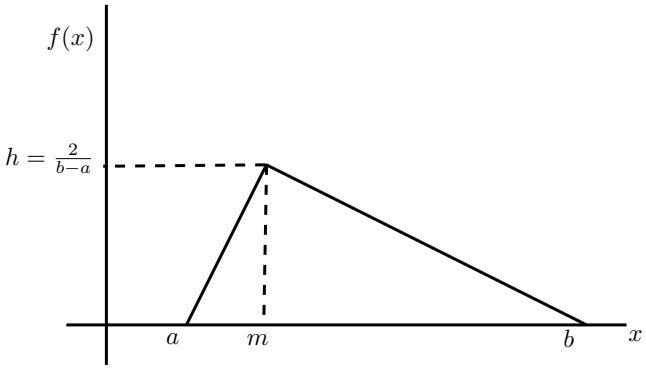


Figure 3.13: Triangular density

and $f(x) = 0$ otherwise. Then

$$\mathrm{P}(X \leq x) = \frac{x-a}{b-a}, a < x < b, \quad \mathrm{E}[X] = \frac{a+b}{2}, \quad \mathrm{Var}[X] = \frac{1}{12}(b-a)^2.$$

Triangular

A Triangular random variable X on the interval $[a, b]$ has density

$$f(x) = \begin{cases} h\frac{x-a}{m-a} & a \leq x \leq m, \\ h\frac{b-x}{b-m} & m \leq x \leq b, \end{cases}$$

and $f(x) = 0$ otherwise, see Figure 3.13. The height h follows from the requirement that $\int_a^b f(x)dx = \frac{(b-a)h}{2} = 1$, yielding $h = \frac{2}{b-a}$. Then

$$\mathrm{E}[X] = \frac{1}{3}(a+b+m), \quad \mathrm{Var}[X] = \frac{1}{18}(a^2+b^2+m^2-ab-am-bm).$$

Exponential

An Exponential random variable X with parameter (or rate) $\lambda > 0$ has density

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0,$$

and $f(x) = 0$ otherwise. Then

$$\mathrm{P}(X \leq x) = 1 - e^{-\lambda x}, x > 0, \quad \mathrm{E}[X] = \frac{1}{\lambda}, \quad \mathrm{Var}[X] = \frac{1}{\lambda^2}.$$

Erlang

An Erlang random variable X with parameters n and λ, is the sum of n independent Exponential random variables $X_1, \ldots, X_n$, each with parameter λ. Its density is given by

$$f(x) = \lambda e^{-\lambda x}\frac{(\lambda x)^{n-1}}{(n-1)!}, \quad x > 0,$$

and $f(x) = 0$ otherwise. For the distribution function we get

$$\mathrm{P}(X \leq x) = \int_0^x \lambda e^{-\lambda x}\frac{(\lambda x)^{n-1}}{(n-1)!}dx = 1 - \sum_{i=0}^{n-1} e^{-\lambda x}\frac{(\lambda x)^i}{(i)!}, \quad x > 0.$$

Then

$$\mathrm{E}[X] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = \frac{n}{\lambda}, \quad \mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = \frac{n}{\lambda^2}.$$

Gamma

A Gamma random variable X with parameters $\alpha > 0$ and $\lambda > 0$, has density

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}, \quad x > 0,$$

and $f(x) = 0$ otherwise, where $\Gamma(\alpha)$ is the gamma function,

$$\Gamma(\alpha) = \int_0^{\infty} e^{-y} y^{\alpha-1} dy, \quad \alpha > 0.$$

Then

$$\mathrm{E}[X] = \frac{\alpha}{\lambda}, \quad \mathrm{Var}[X] = \frac{\alpha}{\lambda^2}.$$

Normal

A Normal random variable X with parameters μ and $\sigma > 0$, has density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}, \quad -\infty < x < \infty.$$

Then

$$E(X) = \mu, \quad \mathrm{var}(X) = \sigma^2.$$

The density $f(x)$ is denoted as $N(\mu, \sigma^2)$ density.

Standard Normal

A Standard Normal random variable X is a Normal random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$. So it has the $N(0, 1)$ density,

$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

and

$$P(X \leq x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy.$$

The Exponential distribution has some remarkable properties.

Property 3.3. (Exponential)

- **Memoryless:** For all $t, s > 0$, we have (see Figure 3.14)

$$P(X > t + s | X > s) = P(X > t). \tag{3.6}$$

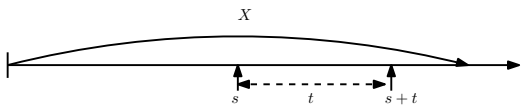


Figure 3.14: Exponential X overshooting s with more than t

Thinking of X as the lifetime of a component, then this property states that, if at time s the component is still working, then its remaining lifetime is again Exponential with exactly the same mean as a new component. In other words, used is as good as new! This memoryless property

is important in many application and can be straightforwardly derived from the definition of conditional probabilities,

$$\begin{aligned} \mathrm{P}(X > t+s | X > s) &= \frac{\mathrm{P}(X > t+s, X > s)}{\mathrm{P}(X > s)} \\ &= \frac{\mathrm{P}(X > t+s)}{\mathrm{P}(X > s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ &= e^{-\lambda t}. \end{aligned}$$

An alternative form of (3.6) is that for small $\Delta > 0$

$$\mathrm{P}(X \leq s + \Delta | X > s) = 1 - e^{-\lambda\Delta} \approx \lambda\Delta, \tag{3.7}$$

stating that, when the component is operational at time s, it will fail in the next time interval Δ with probability $\lambda\Delta$.

- **Minimum:** Let X_1 and X_2 be two independent Exponential random variables with rates λ_1 and λ_2. Then we obtain for the distribution of the minimum of the two,

$$\begin{aligned} \mathrm{P}(\min(X_1, X_2) \leq x) &= 1 - \mathrm{P}(\min(X_1, X_2) > x) \\ &= 1 - \mathrm{P}(X_1 > x, X_2 > x) \\ &= 1 - \mathrm{P}(X_1 > x)\mathrm{P}(X_2 > x) \\ &= 1 - e^{-(\lambda_1+\lambda_2)x}. \end{aligned}$$

Hence, the minimum of X_1 and X_2 is again Exponential, and the rate of the minimum is the sum of the rates λ_1 and λ_2.

Remark 3.2. By integrating the gamma function by parts we obtain

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1).$$

So, if $\alpha = n$, then

$$\Gamma(n) = (n-1)\Gamma(n-1) = \cdots = (n-1)!\Gamma(1) = (n-1)!,$$

since $\Gamma(1) = 1$. Hence, the Gamma distribution with parameters n and λ is the same as the Erlang distribution.

The Normal distribution has the following useful properties.

Property 3.4. (Normal)

- **Linearity:** If X is Normal, then $aX + b$ is Normal. This can be shown by calculating

$$\mathrm{P}(aX + b \leq x) = \mathrm{P}(X \leq \frac{x-b}{a}) = \int_{-\infty}^{\frac{x-b}{a}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2/\sigma^2} dy,$$

where we assumed that $a > 0$ (but the same calculations can be done for $a < 0$). Substituting $z = ay + b$ yields

$$\mathrm{P}(aX + b \leq x) = \int_{-\infty}^{x} \frac{1}{a\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(z-(a\mu+b))^2/(a\sigma)^2} dz,$$

from which we can conclude that $aX + b$ is Normal with parameters $a\mu + b$ and $a\sigma$.

- **Additivity:** If X and Y are independent and Normal, then $X + Y$ is Normal.

- **Excess probability:** The probability that a Normal random variable X lies more than z times the standard deviations above its mean is

$$\mathrm{P}(X \geq \mu + z\sigma) = 1 - \mathrm{P}(X \leq \mu + z\sigma) = 1 - \mathrm{P}(\frac{X-\mu}{\sigma} \leq z) = 1 - \Phi(z),$$

since $\frac{X-\mu}{\sigma}$ is Standard Normal.

- **Percentiles:** The $100p\%$ percentile z_p of the Standard Normal distribution is the unique point z_p for which

$$\Phi(z_p) = p.$$

For example, for $p = 0.95$ we have $z_{0.95} = 1.64$, and $z_{0.975} = 1.96$ for $p = 0.975$.

We conclude this section by the concept of failure rate for positive random variables. Thinking of X as the lifetime of a component, then the probability that the component will fail in the next Δ time units when it reached the age of x time units, is equal to

$$\mathrm{P}(X \leq x + \Delta | X > x) = \frac{\mathrm{P}(x < X \leq x + \Delta)}{\mathrm{P}(X > x)} \approx \frac{f(x)\Delta}{1 - F(x)}.$$

Dividing this probability by Δ yields the rate $r(x)$ at which the component will fail at time x given that it reached time x. This rate is called the failure rate or hazard rate of X, and it given by

$$r(x) = \frac{f(x)}{1 - F(x)}.$$

Example 3.29. (Exponential) If X is Exponential with rate λ, then (see also (3.7))

$$r(x) = \frac{f(x)}{1 - F(x)} = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}} = \lambda.$$

Hence, the failure rate of an Exponential random variable is constant: at any point in time the component is equally likely to fail, or in other words, the coomponents is always as good as new. In fact, this is the only distribution for which the failure rate is constant. In practice, however, many components have a bath-tube failure rate, reflecting that initially and at the end of the lifetime of a component is higher than average.

Exercise 21. (*Problem 10.1 [5]*) The life time of an appliance is a continuous random variable X and has a probability density $f(x)$ of the form $f(x) = c(1+x)^{\hat{a}\check{L}\check{S}3}$ for $x > 0$ and $f(x) = 0$ otherwise. What is the value of the constant c? Find $\mathrm{P}(X \leq 0.5)$, $\mathrm{P}(0.5 < X \leq 1.5)$ and $\mathrm{P}(0.5 < X \leq 1.5 | X > 0.5)$.

Exercise 22. (*Problem 10.3 [5]*) Sizes of insurance claims can be modeled by a continuous random variable with probability density $f(x) = \frac{1}{50}(10 - x)$ for $0 < x < 10$ and $f(x) = 0$ otherwise. What is the probability that the size of a particular claim is larger than 5 given that the size exceeds 2?

Exercise 23. (*Problem 10.4 [5]*) The lengths of phone calls (in minutes) made by a travel agent can be modeled as a continuous random variable with probability density $f(x) = 0.25e^{-0.25x}$ for $x > 0$. What is the probability that a particular phone call will take more than 7 minutes?

Exercise 24. (*Problem 10.5 [5]*) Let X be a positive random variable with probability density function $f(x)$. Define the random variable Y by $Y = X^2$. What is the probability density function of Y? Also, find the density function of the random variable $W = V^2$ if V is a number chosen at random from the interval $(-a, a)$ with $a > 0$.

Exercise 25. (*Problem 10.8 [5]*) A stick of unit length is broken at random into two pieces. Let the random variable X represent the length of the shorter piece. What is the probability density of X? Also, use the probability distribution function of X to give an alternative derivation of the probability density of the random variable $X/(1-X)$, which is the ratio of the length of the shorter piece to that of the longer piece.

Exercise 26. (*Problem 10.11 [5]*) The javelin thrower Big John throws the javelin more than x meters with probability $\mathrm{P}(x)$, where $\mathrm{P}(x) = 1$ for $0 \leq x < 50$, $\mathrm{P}(x) = \frac{1200-(x-50)^2}{1200}$ for $50 \leq x < 80$, $\mathrm{P}(x) = \frac{(90-x)^2}{400}$ for $80 \leq x < 90$, and $\mathrm{P}(x) = 0$ for $x \geq 90$. What is the expected value of the distance thrown in his next shot?

Exercise 27. (*Problem 10.19 [5]*) A point Q is chosen at random inside a sphere with radius r. What are the expected value and the standard deviation of the distance from the center of the sphere to the point Q?

Exercise 28. (*Problem 10.20 [5]*) The lifetime (in months) of a battery is a random variable X satisfying $\mathrm{P}(X \leq x) = 0$ for $x < 5$, $\mathrm{P}(X \leq x) = [(x-5)^3 + 2(x-5)]/12$ for $5 \leq x < 7$ and $\mathrm{P}(X \leq x) = 1$ for $x \geq 7$. What are the expected value and the standard deviation of X?

Exercise 29. (*Problem 10.23 [5]*) In an inventory system, a replenishment order is placed when the stock on hand of a certain product drops to the level s, where the reorder point s is a given positive number. The total demand for the product during the lead time of the replenishment order has the probability density $f(x) = \lambda e^{-\lambda x}$ for $x > 0$. What are the expected value and standard deviation of the shortage (if any) when the replenishment order arrives?

Exercise 30. (*Problem 10.27 [5]*) The lifetime of a light bulb has an uniform probability density on $(2, 12)$. The light bulb will be replaced upon failure or upon reaching the age 10, whichever occurs first. What are the expected value and the standard deviation of the age of the light bulb at the time of replacement?

Exercise 31. (*Problem 10.28 [5]*) A rolling machine produces sheets of steel of different thickness. The thickness of a sheet of steel is uniformly distributed between 120 and 150 millimeters. Any sheet having a thickness of less than 125 millimeters must be scrapped. What are the expected value and the standard deviation of a non-scrapped sheet of steel?

Exercise 32. (*Problem 10.30 [5]*) Limousines depart from the railway station to the airport from the early morning till late at night. The limousines leave from the railway station with independent inter-departure times that are exponentially distributed with an expected value of 20 minutes. Suppose you plan to arrive at the railway station at three o'clock in the afternoon. What are the expected value and the standard deviation of your waiting time at the railway station until a limousine leaves for the airport?

Exercise 33. (*Homework exercises 12*) The lifetimes of two components in an electronic system are independent random variables X_1 and X_2, where X_1 and X_2 are exponentially distributed with an expected value of 2 respectively 3 time units. Let random variable Y denote the time between the first failure and the second failure. What is the probability density distribution $f(y)$ of Y?

Exercise 34. (*Homework exercises 14*) Consider an equilateral triangle (gelijkzijdige driehoek). Denote the length of each of the edges by X, where X is a continuous variable with a uniform density over interval $(1, 3)$. What is the expected area of the equilateral triangle?

3.5 Central limit theorem

One of the most beautiful and powerful results in probability is the Central Limit Theorem. Consider independent random variables $X_1, X_2, ...$, all with the same distribution with mean μ and standard deviation σ. Clearly, for the sum $X_1 + \cdots + X_n$ we have

$$\mathrm{E}[X_1 + \cdots + X_n] = n\mu, \quad \sigma[X_1 + \cdots + X_n] = \sigma\sqrt{n}$$

But what is the distribution of $X_1 + \cdots + X_n$ when n is large? The Central Limit Theorem states that, for any $a < b$,

$$\lim_{n\to\infty} \mathrm{P}(a \leq \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq b) = \Phi(b) - \Phi(a),$$

or in words: the sum $X_1 + \cdots + X_n$ has approximately a normal distribution when n is large, no matter what form the distribution of X_i takes! This also explains why the Normal distribution often appears in practice, since many random quantities are addition of many small random effects.

Example 3.30. (Coin tossing [5]) A friend claims to have tossed 5227 heads in 10000 tosses. Do you believe your friend? Let X_i indicate whether head turns up in toss i or not, so $\mathrm{P}(X_i = 1) = P(X_i = 0) = \frac{1}{2}$ (assuming your friend has a fair coin). Then the total number of heads is

$$X = X_1 + \cdots + X_{10.000}.$$

Clearly, $\mathrm{E}[X] = 10000 \cdot \frac{1}{2} = 5000$ and $\mathrm{Var}[X] = 10000 \cdot \frac{1}{4} = 2500$, so $\sigma[X] = 50$. According to the Central Limit Theorem, the distribution of X is approximately Normal with $\mu = 5000$ and $\sigma = 50$, so the probability of a realization greater or equal to $5227 = \mu + 4.54\sigma$ is $1 - \Phi(4.54) \approx 3.5 \cdot 10^{-6}$. So do you believe your friend?

An important application of the Central Limit Theorem is the construction of confidence intervals. Consider the problem of estimating the unknown expectation $\mu = E(X)$ of a random variable X. Suppose n independent samples $X_1, \ldots, X_n$ are generated, then by the law of large numbers, the sample mean is an estimator for the unknown $\mu = E(X)$,

$$\overline{X}(n) = \frac{1}{n}\sum_{k=1}^{n} X_k$$

The Central Limit Theorem tells us that

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}(n) - \mu}{\sigma/\sqrt{n}}$$

has approximately a Standard Normal distribution, where σ is the standard deviation of X. Hence

$$\mathrm{P}\left(-z_{1-\frac{1}{2}\alpha} \leq \frac{\overline{X}(n) - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{1}{2}\alpha}\right) \approx 1 - \alpha, \tag{3.8}$$

where the percentile $z_{1-\frac{1}{2}\alpha}$ is the point for which the area under the Standard Normal curve between points $-z_{1-\frac{1}{2}\alpha}$ and $z_{1-\frac{1}{2}\alpha}$ equals $100(1-\alpha)\%$. Rewriting (3.8) leads to the following interval containing μ with probability $1 - \alpha$,

$$\mathrm{P}\left(\overline{X}(n) - z_{1-\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}(n) + z_{1-\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha. \tag{3.9}$$

Thus, for large n, an approximate $100(1 - \alpha)\%$ confidence interval for μ is $\overline{X}(n) \pm z_{1-\frac{1}{2}\alpha}\frac{S(n)}{\sqrt{n}}$, that is, this interval covers the mean μ with probability $1 - \alpha$. Note that the confidence interval is random, not the mean μ, see Figure 3.15.

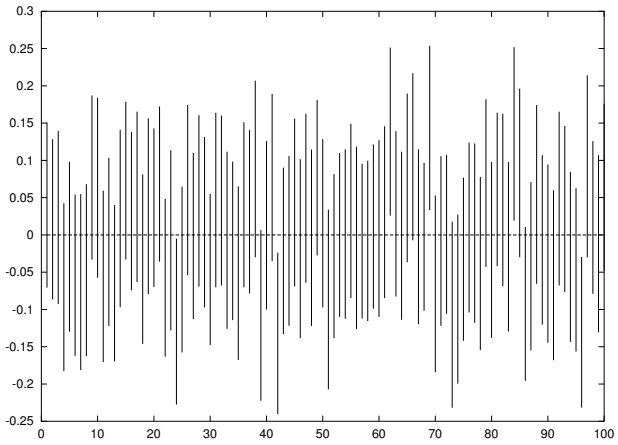


Figure 3.15: 100 confidence intervals to estimate the mean 0 of the Uniform random variable on $(-1, 1)$, where each interval is based on 100 samples

Remark 3.3.

- If the standard deviation σ is unknown, it can be estimated by square root of the sample variance

$$S^2(n) = \frac{1}{n}\sum_{k=1}^{n}\left[X_k - \overline{X}(n)\right]^2$$

  and then this estimate $S(n)$ can be used in (3.9) instead of σ.

- About 100 times as many samples are needed to reduce the width of a confidence interval by a factor $\frac{1}{10}$!

3.6 Joint random variables

If X and Y are discrete random variables, then

$$p(x, y) = \mathrm{P}(X = x, Y = y)$$

is the joint probability mass function of X and Y. The marginal probability mass functions of X and Y follow from

$$\begin{aligned} P_X(x) &= P(X = x) = \sum_y P(X = x, Y = y), \\ P_Y(y) &= P(Y = y) = \sum_x P(X = x, Y = y). \end{aligned}$$

Example 3.31. We repeatedly draw a random number from $1, \ldots, 10$. Let X be number of draws until the first 1 appears and let Y be the number until the first 10 appears. What is joint probability mass of X and Y? Let us calculate the probability $\mathrm{P}(X = n, Y = n + k)$ for $n, k = 1, 2, \ldots$ (the probabilities $\mathrm{P}(X = n + k, Y = n$ can be calculated along the same lines, and in fact, $\mathrm{P}(X = n, Y = n + k) = \mathrm{P}(X = n + k, Y = n)$ by symmetry). This means that we first draw $n - 1$ numbers different from 1 and 10, then we draw 1, and subsequently we draw $k - 1$ numbers different from 10 (so there could also be some 1s) and finally, the number 10 appears. Hence

$$\mathrm{P}(X = n, Y = n + k) = \left(\frac{8}{10}\right)^{n-1}\frac{1}{10}\left(\frac{9}{10}\right)^{k-1}\frac{1}{10}.$$

If X and Y are continuous random variables, then

$$\mathrm{P}(X \leq a, Y \leq b) = \int_{x=-\infty}^{a} \int_{y=-\infty}^{b} f(x,y)dxdy$$

is the joint probability distribution function of X and Y, where $f(x,y)$ is the joint density, satisfying

$$f(x,y) \geq 0, \quad \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f(x,y)dxdy = 1.$$

The function $f(x,y)$ can be interpreted as the density of the joint probability mass, that is, for small $\Delta > 0$

$$\mathrm{P}(x < X \leq x+\Delta, y < Y \leq y+\Delta) \approx f(x,y)\Delta^2$$

The joint density $f(x,y)$ can be obtained from the joint distribution $\mathrm{P}(X \leq x, Y \leq y)$ by taking partial derivatives,

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} P(X \leq x, Y \leq y).$$

The marginal densities of X and Y follow from

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy, \\ f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx. \end{aligned}$$

The random variables X and Y are independent if

$$f(x,y) = f_X(x)f_Y(y) \quad \text{for all } x, y.$$

Example 3.32.

- The random variable X is the distance to the origin 0 and Y is angle (in radians) of a random point in a disk of radius r. What is the joint distribution and joint density of X and Y?

  $$P(X \leq x, Y \leq y) = \frac{\pi x^2 \frac{y}{2\pi}}{\pi r^2} = \frac{x^2}{r^2}\,\frac{y}{2\pi}, \quad 0 \leq x \leq r, 0 \leq y \leq 2\pi$$

  and by taking the partial derivatives, the joint density is found to be

  $$f(x,y) = \frac{2x}{r^2}\,\frac{1}{2\pi}, \quad 0 \leq x \leq r, 0 \leq y \leq 2\pi.$$

  Hence the marginal densities are

  $$f_X(x) = \int_{y=0}^{2\pi} f(x,y)dy = \frac{2x}{r^2}, \quad 0 \leq x \leq r, \quad f_Y(y) = \int_{x=0}^{r} f(x,y)dx = \frac{1}{2\pi}, 0 \leq y \leq 2\pi,$$

  so $f(x,y) = f_X(x)f_Y(y)$, and thus X and Y are independent.

- Let X_1 and X_2 be two independent Exponential random variables, with rates λ_1 and λ_2. What is the probability that X_1 is the smallest one? We need to calculate

  $$\mathrm{P}(X_1 \leq X_2) = \int_{x_1=0}^{\infty} \int_{x_2=x_1}^{\infty} f(x_1,x_2)dx_2dx_1.$$

  Since X_1 and X_2 are independent, the joint density $f(x_1,x_2)$ is equal to

  $$f(x_1,x_2) = \lambda_1 e^{-\lambda_1 x_1} \lambda_2 e^{-\lambda_2 x_2},$$

so

$$\begin{aligned}
\mathrm{P}(X_1 \le X_2) &= \int_{x_1=0}^{\infty}\int_{x_2=x_1}^{\infty} \lambda_1 e^{-\lambda_1 x_1}\lambda_2 e^{-\lambda_2 x_2} dx_2 dx_1 \\
&= \int_{x_1=0}^{\infty} \lambda_1 e^{-\lambda_1 x_1}\int_{x_2=x_1}^{\infty} \lambda_2 e^{-\lambda_2 x_2} dx_2 dx_1 \\
&= \int_{x_1=0}^{\infty} \lambda_1 e^{-\lambda_1 x_1} e^{-\lambda_2 x_1} dx_1 \\
&= \frac{\lambda_1}{\lambda_1+\lambda_2}.
\end{aligned}$$

Thus the probability that X_1 is the smallest one is proportional to the rates.

- Let X be random on $(0,1)$ and Y be random on $(0,X)$. What is the density of the area of the rectangle with sides X and Y? So we need to calculate the density of $Z = XY$. First note that the joint density of X and Y is equal to

$$f(x,y) = \frac{1}{x}, \quad 0 \le y \le x \le 1,$$

and $f(x,y) = 0$ elsewhere. Hence, for $0 \le z \le 1$,

$$\begin{aligned}
\mathrm{P}(Z \le z) &= \mathrm{P}(XY \le z) \\
&= \int_{x=0}^{\sqrt{z}}\int_{y=0}^{x} f(x,y)dydx + \int_{x=\sqrt{z}}^{1}\int_{y=0}^{\frac{z}{x}} f(x,y)dydx \\
&= \int_{x=0}^{\sqrt{z}}\int_{y=0}^{x} \frac{1}{x}dydx + \int_{x=\sqrt{z}}^{1}\int_{y=0}^{\frac{z}{x}} \frac{1}{x}dydx \\
&= \sqrt{z} + \int_{x=\sqrt{z}}^{1} \frac{1}{x^2}dx \\
&= \sqrt{z} + \sqrt{z} - z \\
&= 2\sqrt{z} - z.
\end{aligned}$$

Taking the derivative yields the density

$$f_Z(z) = \frac{d}{dz}\mathrm{P}(Z \le z) = \frac{1}{\sqrt{z}} - 1, \quad 0 \le z \le 1.$$

The expectation of the random variable $g(X, Y)$, where $g(x,y)$ is any function, can be calculated as

$$\mathrm{E}\left[g(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy.$$

For example, taking $g(x,y) = ax + by$, where a and b are constants,

$$\begin{aligned}
\mathrm{E}\left[aX+bY\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax+by)f(x,y)dxdy \\
&= a\int_{-\infty}^{\infty} x\int_{-\infty}^{\infty} f(x,y)dydx + b\int_{-\infty}^{\infty} y\int_{-\infty}^{\infty} f(x,y)dxdy \\
&= a\int_{-\infty}^{\infty} x f_X(x)dx + b\int_{-\infty}^{\infty} y f_Y(y)dy \\
&= a\mathrm{E}\left[X\right] + b\mathrm{E}\left[Y\right].
\end{aligned}$$

and similarly, if X and Y are independent (so $f(x,y) = f_X(x)f_Y(y)$),

$$\mathrm{E}\left[XY\right] = \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]. \tag{3.10}$$

Example 3.33. Pick two points X and Y at random from the interval $(0, 1)$. What is the mean distance between these two points? We need to calculate $\mathrm{E}[|X - Y|]$. The joint density of X and Y is $f(x, y) = 1$ for $0 \leq x, y \leq 1$, so

$$\mathrm{E}[|X - Y|] = 2\int_{x=0}^{1}\int_{y=0}^{x}(x - y)dydx = 2\int_{x=0}^{1}\frac{1}{2}x^2dx = \frac{1}{3},$$

as expected.

The covariance of two random variables X and Y is defined as

$$\begin{aligned}\mathrm{cov}[X, Y] &= \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]\end{aligned}$$

and the correlation coefficient of X and Y is a normalized version of the covariance,

$$\rho[X, Y] = \frac{\mathrm{cov}[X, Y]}{\sigma[X]\sigma[Y]}.$$

It can be shown (check!) that $-1 \leq \rho[X, Y] \leq 1$. The correlation coefficient is often used as a measure for dependence between the random variables X and Y. If X and Y are independent, then it follows from (3.10) that $\mathrm{cov}[X, Y] = 0$, and thus that $\rho[X, Y] = 0$. If $X = aY$, then $\rho[X, Y] = 1$ if $a > 0$ and $\rho[X, Y] = -1$ if $a < 0$, which express maximal positive and negative correlation.

Exercise 35. (*Problem 11.8 [5]*) Let the joint probability density function of the random variables X and Y be given by $f(x, y) = ce^{-2x}$ for $0 < y \leq x < \infty$ and $f(x, y) = 0$ otherwise. Determine the constant c. What is the probability density of $X - Y$?

Exercise 36. (*Problem 11.12 [5]*) The joint density of the random variables X and Y is given by $f(x, y) = 4xe^{-2x(1+y)}$ for $x, y > 0$. What are the marginal densities of X and Y?

Exercise 37. (*Homework exercises 17*) We select a point (x, y) at random from a square with sides 1. Let X be the random variable for the x-coordinate and Y the random variable for the y-coordinate of that point. What is the probability that $X > 1.5 - Y$?

3.7 Conditioning

In this section we summarize the concepts of conditional probabilities and conditional expectations for random variables. Conditioning is a fruitful technique to calculate probabilities and expectations.

Let X and Y be discrete random variables. The conditional probability mass function of X given $Y = y$ is defined as

$$\mathrm{P}(X = x|Y = y) = \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(Y = y)}.$$

Then the unconditional probabilities $\mathrm{P}(X = x)$ can be calculated by conditioning on the possible values of Y,

$$P(X = x) = \sum_y P(X = x|Y = y)P(Y = y).$$

Example 3.34. Simultaneously roll 24 dice and next roll with those that showed 6. Let X be number of sixes in the first roll and let Y be those in the second roll.

- What is $\mathrm{P}(Y = y|X = x)$? This means that in the second roll, x dice are used. Hence, the probability that y dice out of x show six is

$$\mathrm{P}(Y = y|X = x) = \binom{x}{y}\left(\frac{1}{6}\right)^y\left(\frac{5}{6}\right)^{x-y}. \tag{3.11}$$

- What is $P(Y = y)$? To calculate this probability we condition on the outcome of X and then use the above result, so

$$\begin{aligned} P(Y = y) &= \sum_{x=y}^{24} P(Y = y|X = x)P(X = x) \\ &= \sum_{x=y}^{24} P(Y = y|X = x)\binom{24}{x}\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{24-x} \\ &= \sum_{x=y}^{24}\binom{x}{y}\left(\frac{1}{6}\right)^y\left(\frac{5}{6}\right)^{x-y}\binom{24}{x}\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{24-x}. \end{aligned} \tag{3.12}$$

- And what is $P(X = x|Y = y)$? Using the definition of conditional probability,

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)},$$

  so this probability can be calculated by substituting (3.11) and (3.12) and

$$P(X = x) = \binom{24}{x}\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{24-x}.$$

Now let X and Y be continuous random variables with joint density $f(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$. The conditional density of X given $Y = y$ is defined analogously to the discrete case,

$$\begin{aligned} f_X(x|y)dx &= P(x < X \le x + dx|y < Y \le y + dy) \\ &= \frac{P(x < X \le x + dx, y < Y \le y + dy)}{P(y < Y \le y + dy)} \\ &= \frac{f(x, y)dxdy}{f_Y(y)dy} \\ &= \frac{f(x, y)}{f_Y(y)}dx, \end{aligned}$$

so the conditional density of X given $Y = y$ is

$$f_X(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

Then the conditional probability $P(X \le x|Y = y)$ follows from

$$P(X \le x|Y = y) = \int_{z=-\infty}^{x} f_X(z|y)dz$$

and the unconditional probability $P(X \le x)$ can be calculated by conditioning on the possible values of Y,

$$P(X \le x) = \int_{-\infty}^{\infty} P(X \le x|Y = y)f_Y(y)dy.$$

Example 3.35.

- Point (X, Y) is randomly chosen in the unit circle. What is the conditional density of X given $Y = y$? We have

$$f(x, y) = \frac{1}{\pi}, \quad x^2 + y^2 \le 1,$$

and $f(x,y) = 0$ otherwise. So for $-1 \leq y \leq 1$,

$$f_Y(y) = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{1}{\pi} dx = \frac{2}{\pi}\sqrt{1-y^2}.$$

Hence, for $-\sqrt{1-y^2} \leq x \leq \sqrt{1-y^2}$,

$$f_X(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{\frac{1}{\pi}}{\frac{2}{\pi}\sqrt{1-y^2}} = \frac{1}{2\sqrt{1-y^2}},$$

which is the Uniform density on the interval $(-\sqrt{1-y^2}, \sqrt{1-y^2})$, as expected.

- You are waiting for the metro. Once the metro has stopped, the distance to the nearest metro door is Uniform between 0 and 2 metres. If the distance to the nearest door is y metres, then you are able to find a place to sit with probability

$$1 - \sqrt{\frac{1}{2}y}$$

What is the probability Mr Johnson finds a place to sit? Let the random variable Y be the distance to the nearest door and X indicate whether you find a place to sit (so $X = 1$ if you can sit and $X = 0$ otherwise). To calculate the probability $\mathrm{P}(X = 1)$, note that Y is Uniform on $(0, 2)$, so

$$f_Y(y) = \frac{1}{2}, \quad 0 < y < 2,$$

and $f_Y(y) = 0$ elsewhere. Also,

$$\mathrm{P}(X = 1|Y = y) = 1 - \sqrt{\frac{1}{2}y}.$$

Hence

$$\mathrm{P}(X = 1) = \int_{y=0}^{2} \mathrm{P}(X = 1|Y = y) f_Y(y) dy = \int_{y=0}^{2} (1 - \sqrt{\frac{1}{2}y})\frac{1}{2} dy = \frac{1}{3}.$$

Note that this probability is not equal to $1 - \sqrt{\frac{1}{2}\mathrm{E}[Y]} = 1 - \sqrt{\frac{1}{2}}$!

Conditioning is a fruitful technique, not only to calculate probabilities, but also to calculate expectations. For discrete random variables X and Y, the conditional expectation of X given $Y = y$ is defined as

$$\mathrm{E}[X|Y = y] = \sum_x x\mathrm{P}(X = x|Y = y),$$

and then the (unconditional) expectation of X can be calculated by conditioning on the possible values of Y,

$$\mathrm{E}[X] = \sum_y \mathrm{E}[X|Y = y]\,\mathrm{P}(Y = y).$$

Analogously, for continuous random variables X and Y, we have

$$\mathrm{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_X(x|y) dx,$$

and

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} \mathrm{E}[X|Y = y]\, f_Y(y) dy.$$

Example 3.36.

- Generate two random numbers X_1 and X_2 from $(0,1)$. Let X be the smallest of X_1 and X_2 and Y the largest. What are $\mathsf{E}[X|Y=y]$ and $\mathsf{E}[X]$? Note that

$$f(x,y) = f_{X_1}(x)f_{X_2}(y) + f_{X_1}(y)f_{X_2}(x) = 1+1 = 2, \quad 0 < x < y < 1,$$

and

$$f_Y(y) = \mathsf{P}(X_1 < y)f_{X_2}(y) + f_{X_1}(y)\mathsf{P}(X_2 < y) = 2y, \quad 0 < y < 1.$$

So

$$f_X(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{1}{y}, \quad 0 < x < y,$$

which means that X is uniform on $(0,y)$ if $Y=y$ (as expected). Hence,

$$\mathsf{E}[X|Y=y] = \int_{x=0}^{y} x\frac{1}{y}dx = \frac{1}{2}y,$$

and

$$\mathsf{E}[X] = \int_{y=0}^{1} \mathsf{E}[X|Y=y]\, f_Y(y)dy = \int_{y=0}^{1} y^2 dy = \frac{1}{3}.$$

Of course, since $f_X(x) = 2(1-x)$ for $0 < x < 1$, the expectation of X can also be calculated directly,

$$\mathsf{E}[X] = \int_{x=0}^{1} x f_X(x)dx = \int_{x=0}^{1} 2x(1-x)dx = \frac{1}{3}.$$

- A batch consists of a random number of items N, where $\mathsf{P}(N=n) = (1-p)p^{n-1}, n \geq 1$. The production time of a single item is Uniform between 4 and 10 minutes. What is the mean production time of a batch? Let B denote the production time of the batch, and X_i the production time of a single item (so $\mathsf{E}[X_i] = 7$ mins). Then we can write

$$B = \sum_{i=1}^{N} X_i.$$

To calculate $\mathsf{E}[B]$ we condition on $N=n$, yielding

$$\mathsf{E}[B|N=n] = \mathsf{E}[X_1 + \cdots + X_n] = \mathsf{E}[X_1] + \cdots + \mathsf{E}[X_n] = 7n.$$

Hence

$$\mathsf{E}[B] = \sum_{n=1}^{\infty} \mathsf{E}[B|N=n]\,\mathsf{P}(N=n) = \sum_{n=1}^{\infty} 7n(1-p)p^{n-1} = \frac{7}{1-p} \text{ (mins)}.$$

- Process time X is Exponential with rate Λ, where Λ itself is also random with density

$$f_\Lambda(\lambda) = \lambda e^{-\frac{1}{2}\lambda^2}, \quad \lambda > 0.$$

What is the mean process time? To calculate $\mathsf{E}[X]$ we first condition on the rate $\Lambda = \lambda$, which leads to

$$\mathsf{E}[X|\Lambda=\lambda] = \frac{1}{\lambda},$$

and then we get

$$\mathsf{E}[X] = \int_{\lambda=0}^{\infty} \mathsf{E}[X|\Lambda=\lambda]\, f_\Lambda(\lambda)d\lambda = \int_{\lambda=0}^{\infty} \frac{1}{\lambda}\lambda e^{-\frac{1}{2}\lambda^2}d\lambda = \sqrt{2\pi}\frac{1}{\sqrt{2\pi}}\int_{\lambda=0}^{\infty} e^{-\frac{1}{2}\lambda^2}d\lambda = \sqrt{2\pi}\Phi(0) = \sqrt{\frac{\pi}{2}}.$$

Example 3.37. (Random setups) Suppose that a machine needs a setup after having produced *on average* k jobs. This means that, after having produced a job, the machine needs a setup Y with probability $p = \frac{1}{k}$. Let X denote the natural process of a job, Y the setup time, and Z denotes the *effective process time*, i.e., Z is X plus the possible setup Y,

$$Z = \begin{cases} X + Y & \text{with probability } p, \\ X & \text{with probability } 1 - p. \end{cases}$$

What is the mean and variance of the effective process time Z? To calcute $\mathrm{E}[Z]$ and $\mathrm{Var}[Z]$ we condition on having a setup or not,

$$\begin{aligned} \mathrm{E}[Z] &= \mathrm{E}[X + Y]\, p + \mathrm{E}[X]\,(1 - p) \\ &= (\mathrm{E}[X] + \mathrm{E}[Y])p + \mathrm{E}[X]\,(1 - p) \\ &= \mathrm{E}[X] + p\mathrm{E}[Y], \end{aligned}$$

and similarly

$$\begin{aligned} \mathrm{E}\left[Z^2\right] &= \mathrm{E}\left[(X + Y)^2\right] p + \mathrm{E}\left[X^2\right](1 - p) \\ &= (\mathrm{E}\left[X^2\right] + 2\mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}\left[Y^2\right])p + \mathrm{E}\left[X^2\right](1 - p) \\ &= \mathrm{E}\left[X^2\right] + 2p\mathrm{E}[X]\,\mathrm{E}[Y] + p\mathrm{E}\left[Y^2\right]. \end{aligned}$$

So

$$\mathrm{Var}[Z] = \mathrm{E}\left[Z^2\right] - (\mathrm{E}[Z])^2 = \mathrm{Var}[X] + p\mathrm{Var}[Y] + p(1 - p)(\mathrm{E}[Y])^2.$$

Clearly, $\mathrm{Var}[Z]$ is greater than $\mathrm{Var}[X]$ (even when $\mathrm{Var}[Y] = 0$, i.e. Y is constant).

Exercise 38. (*Problem 13.4 [5]*) Two dice are rolled. Let the random variable X be the smallest of the two outcomes and let Y be the largest of the two outcomes. What are the conditional mass functions $\mathrm{P}(X = x|Y = y)$ and $\mathrm{P}(Y = y|X = x)$?

Exercise 39. (*Problem 13.7 [5]*) Let X and Y be two continuous random variables with the joint probability density $f(x, y) = xe^{-x(y+1)}$ for $x, y > 0$ and $f(x, y) = 0$ otherwise. What are the conditional probability densities $f_X(x|y)$ and $f_Y(y|x)$? What is the probability that Y is larger than 1 given that $X = 1$?

Exercise 40. (*Problem 13.8 [5]*) Let X and Y be two continuous random variables with the joint probability density $f(x, y) = 27(2x + 5y)$ for $0 < x, y < 1$ and $f(x, y) = 0$ otherwise. What are the conditional probability densities $f_X(x|y)$ and $f_Y Y(y|x)$? What is the probability that X is larger than 0.5 given that $Y = 0.2$?

Exercise 41. (*Problem 13.23 [5]*) You generate three random numbers from (0,1). Let X be the smallest of these three numbers and Y the largest. What are the conditional expected values $\mathrm{E}(X|Y = y)$ and $E(Y|X = x)$?

Exercise 42. (*Homework exercises 15*) Let X be a positive random variable with cumulative density function

$$P(X \leq x) = F_X(x) = 1 - \frac{c}{2 + x}.$$

(a) Determine the constant c.

(b) Calculate the conditional probability $\mathrm{P}(X > 8|X \leq 10)$.

4

Manufacturing Models

The essence of any manufacturing system is to transform raw material into physical products to meet demand. This transformation (or manufacturing) is done by a system, which is a collection of elements. This system can be viewed at various levels, see Figure 4.1:

- **Factory level.** This is the whole factory (also referred to as plant or fab), and may consists of several (functional) areas.
- **Area level.** In an area we find several machines or groups of machines.
- **Equipment level.** At this level we have machines or groups of machines (also referred to as manufacturing cells or workstations).

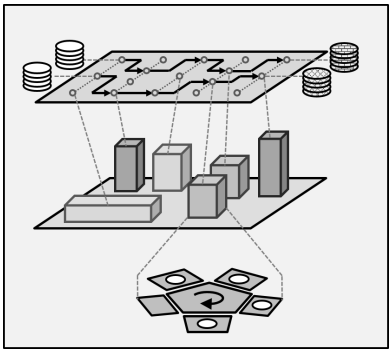


Figure 4.1: Manufacturing systems at different levels

For example, in a semi-conductor (or wafer) fab, you can distinguish the lithographic area, where you find lithographic machines or so-called wafer steppers, see Figure 4.2.

The focus in this course is on manufacturing systems at area level and equipment level. Manufacturing environments vary with respect to their process structure, which can be divided into:

- **Job shops.** High mix of small production batches are produced with a high variety of routings through the plant.
- **Disconnected flow lines.** Production batches are produced on a limited number of distinct routings through the plant. Groups of machines are not connected by a paced material handling system, so inventory can build up between groups of machines in the line.
- **Connected flow lines.** Dedicated production lines producing big production batches along a fixed routing through the line, where machines are connected by a paced material handling system.

Figure 4.2: Lithographic area of the NXP fab in Nijmegen

Figure 4.3: BIM (Breakthrough in Manufacturing) lines of NXP for the assembly of ICs

- **Continuous flow processes.** Continuous products (such as food and chemicals) literally flow through the production line along a fixed routing.

We will mainly focus our attention on manufacturing systems producing discrete parts on disconnected flow lines. Disconnected flow lines can be found in, for example, wafer fabs, producing integrated circuits (ICs) on silicon wafers, whereas the assembly of these ICs in their supporting packages is done on high volume dedicated assembly lines, see Figure 4.3.

The ideal situation is that demand is perfectly aligned with the manufacturing process. However, in practice, this is almost never the case (since machines may break down, raw material is missing, demand suddenly increases, and so on). To align these two processes, buffers are needed. So a buffer is an excess resource to correct for the misalignment between demand and manufacturing processes. It can appear in the following three forms:

- **Inventory buffer.** Extra material in the manufacturing process.
- **Time buffer.** Time delay between demand and satisfaction.
- **Capacity buffer.** Extra production capacity.

4.1 Terminology

In this section we introduce some useful terminology.

- **Workstation** (or simply station) or workcenter is group of machines or workers performing identical tasks.
- **Part** is piece of raw material, component, subassembly or assembly, worked on at a workstation.
- **Raw material** refers to parts purchased outside the plant.
- **Components** are pieces assembled into more complex products.
- **Subassemblies** are assembled products that are further assembled.
- **Assemblies** or final assemblies are end items (sold to customers).
- **Order** or customer order is a request from a customer for a particular product, in a particular quantity to be delivered on a particular (due) date.
- **Routing** is a sequence of workstations passed through by a part or job.
- **Job or lot** is set of materials (and information) traversing a routing in the plant.

4.2 Key performance measures

Below we define some key performance measures relevant in manufacturing environments.

- **Throughput** or throughput rate is the number of good (i.e., non-defective) parts or jobs produced per unit time.
- **Capacity** (or maximal throughput) is an upper limit on the throughput of a production process.
- **Work-In-Process (WIP)** is all the products from the start to the end point of a product routing.
- **Cycle time** or flow time, throughput time or sojourn time is the time it takes from the release of a job at the beginning of its routing to go through the system and to reach the end of its routing. In other words, it is the time a job spends as WIP.
- **Utilization** of a machine is the fraction of time it is not idle for lack of parts. This is not necessarily the fraction of time the machine is processing jobs, but typically also includes failures, setups, and so on. It can be calculated as

  $$u = \frac{T_{\text{nonidle}}}{T_{\text{total}}},$$

  where T_{nonidle} is the time the machine is not idle during the total time frame T_{total}, or alternatively,

  $$u = \frac{\text{realized production rate}}{\text{effective production rate}},$$

  where the effective production rate is the maximum rate at which the machine can process jobs, including effects of failures, setups and so on.

- **Cycle time factor** is the ratio of the cycle time and the **raw process time** t_0, so

$$\text{Cycle Time Factor} = \frac{\text{Cycle Time}}{t_0},$$

where t_0 is the sum of the average process times of each workstation in the line, or in other words, it is the average time it takes a single job to go through the empty line.

Example 4.1. (Capacity) A workstation consists of 2 identical machines. The raw process time of a machine is $t_0 = 0.2$ hours. Then the capacity (or maximal throughput) of each machine is $\frac{1}{t_0} = 5$ jobs per hour, and the capacity of the workstation is 10 jobs per hour.

Example 4.2. (Utilization) The raw process time of a machine is $t_0 = 0.15$ hours and its (realized) production rate is 5 lots per hour. Then the utilization u of the machine is $u = 5 \cdot 0.15 = 0.75$.

4.3 Capacity, flow time and WIP

Let us consider the manufacturing flow line in Figure 4.4. We have a two machine line. Machine M_1 has a process time of 2 hours, machine M_2 has a process time of 3 hours.

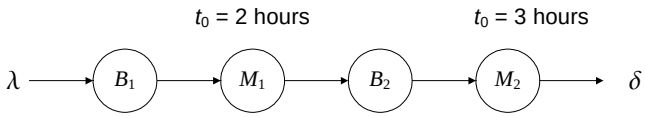


Figure 4.4: Manufacturing flow line

We assume that the buffers have ample capacity. The line in Figure 4.4 has a maximal throughput of 1/3 lots/hour (machine M_2 is the bottleneck). A basic approach for determining the flow time of a lot is with a lot-time diagram. Herein, we plot the state of individual lots against time. Figure 4.5(a) and (b) show the lot-time diagram for the line in which lots are released every 3 hours, and every 2 hours, respectively.

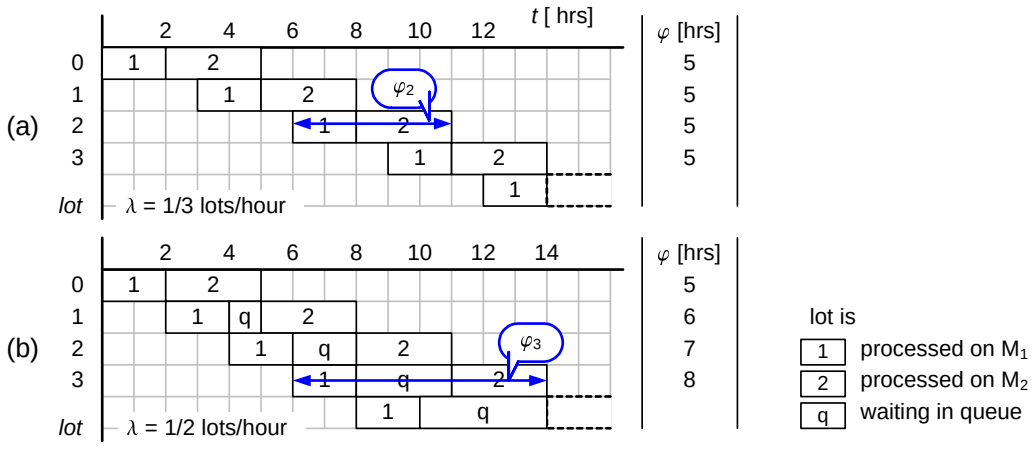


Figure 4.5: Lot-time diagram for release rate of (a) 1/3 lots/hour and (b) 1/2 lots/hour

For a release rate of 1/3 lots/hour we determine from Figure 4.5(a) that the flow time is 5 hours. For a release rate of 1/2 lots/hour we see in Figure 4.5(b) that the flow time keeps increasing. For any release rate under the maximal throughput, the flow time is 5 hours, for any release rate above the maximal throughput, the flow time grows unlimited. We can also use the lot-time diagram to determine the mean WIP level. In Figure 4.6 we derive the WIP level over time from the lot-time diagram. For instance, for a release rate of 1/3 lots/hour, at $t = 7$ there are two lots in the system (lot 1 and 2).

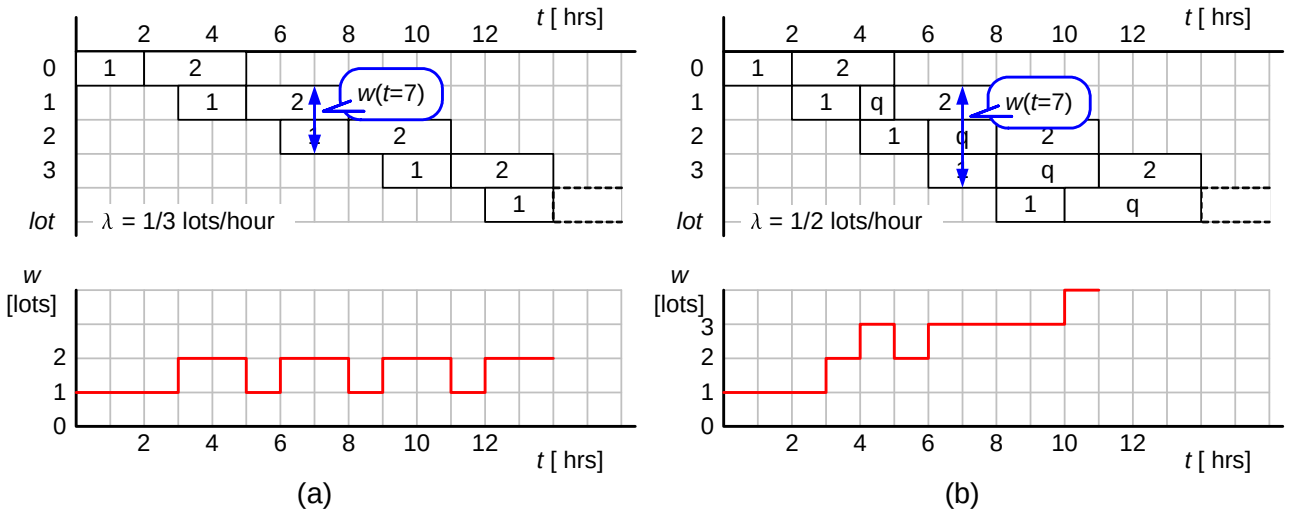


Figure 4.6: Lot-time diagram and w-t-diagram for release rate of (a) 1/3 lots/hour and (b) 1/2 lots/hour

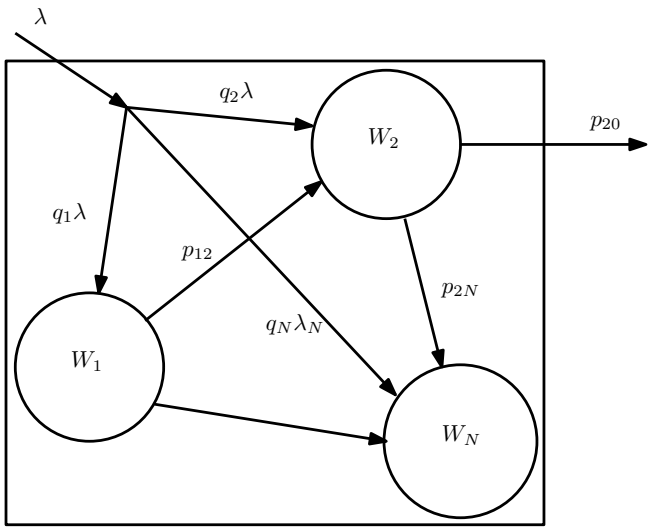


Figure 4.7: Manufacturing network

For a release rate of 1/3 lots/hour, the behaviour becomes periodic for $t > 3$ hours with a period of 3 hours, see Figure 4.6(a). The mean WIP level is $\frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 2 = \frac{5}{3}$ lots. For Figure 4.6(b) the WIP level keeps increasing. For a release rate higher than the maximal throughput, the WIP level grows to infinity.

For the flow line in Figure 4.4 it is easy to determine the maximal throughput, and for constant release rate and constant process times, lot-time diagrams can be used to determine the flow time and mean WIP level. Below we describe how to determine the maximal throughput for an arbitrary configuration of workstations. Treatment of the more complicated problem of estimating mean flow time and mean WIP level in an arbitrary configuration with stochastic arrivals and process times starts in Section 4.8.

Consider a manufacturing system with N workstations, labeled $W_1, \ldots, W_N$, see Figure 4.7. Workstation W_i consists of m_i parallel identical machines. The raw process time of a machine in workstation W_i is t_{0i}. A fraction p_{ij} of the throughput of workstation W_i is diverted to W_j, and a fraction p_{i0} of the throughput is finished and leaves the system. The arrival rate to the manufacturing system is λ jobs per unit time, and a fraction q_i is immediately diverted to workstation W_i.

The inflow or throughput of workstation W_i consists of both internal jobs (coming from other workstations) and external jobs. What is the throughput of workstation W_i?

Let δ_i denote the throughput (or total outflow) of workstation W_i. The job flow in the manufacturing system obeys the principle of conservation of flow. This means that the total flow of jobs out of W_i is equal to the total flow into W_i. Hence, for each workstation we obtain

$$\delta_i = \lambda q_i + \sum_{j=1}^{N} \delta_j p_{ji}, \quad i = 1, \ldots, N,$$

where the left hand side is the flow out, the first term at the right hand side is the fresh inflow in W_i and the second term is the total internal inflow. These linear equations have a unique solution for the throughput δ_i, provided the routing is such that every job eventually leaves the manufacturing system. The capacity of workstation W_i is $\frac{m_i}{t_{0i}}$, so W_i is able to handle all work if the *throughput is less than the capacity*,

$$\delta_i < \frac{m_i}{t_{0i}},$$

and then the utilization u_i of a machine in workstation W_i is

$$u_i = \frac{\delta_i t_{0i}}{m_i} < 1.$$

Note that "=" is only feasible in the ideal deterministic world (but as soon as there is variability, and "=" holds, the WIP in workstation W_i will grow without bounds). The **bottleneck** workstation W_b is the one with the highest utilization u_b. This station also dictates the maximal inflow $\lambda_{\max}$ for which the system is stable. That is, $\lambda_{\max}$ is the rate for which u_b becomes equal to 1.

Example 4.3. Consider the manufacturing system consisting of 4 workstations, listed in Figure 4.8. The inflow in stations W_1 and W_4 is λ jobs per hour. The throughput of the system is δ.

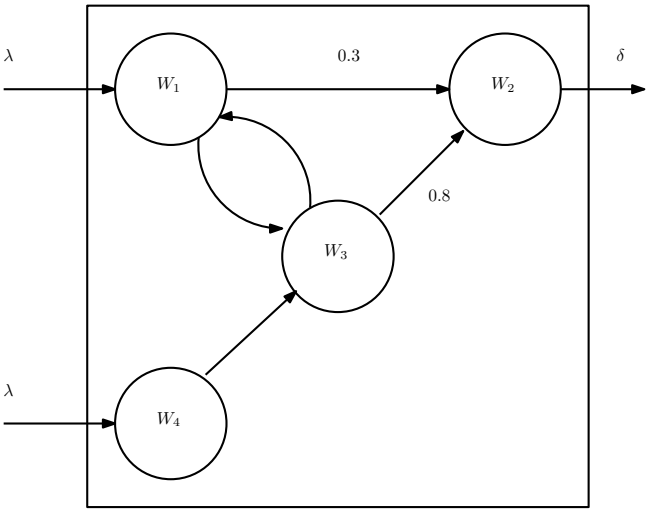


Figure 4.8: Manufacturing system with 4 workstations

The flow equations for this system are given by

$$\begin{aligned}
\delta_1 &= \lambda + 0.2\delta_3, \\
\delta_2 &= 0.3\delta_1 + 0.88\delta_3, \\
\delta_3 &= 0.7\delta_1 + \delta_4, \\
\delta_4 &= \lambda.
\end{aligned}$$

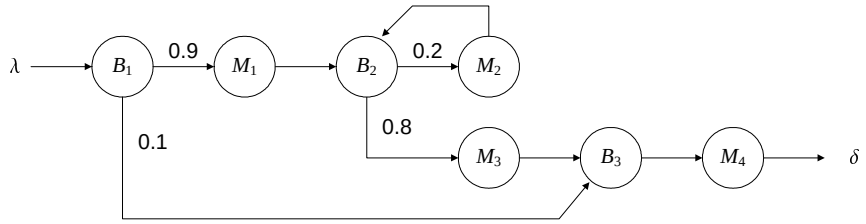
Solution of these equations yields (check!)

$$\delta_1 = \frac{1.2}{0.86}\lambda, \quad \delta_2 = 2\lambda, \quad \delta_3 = \frac{1.7}{0.86}\lambda, \quad \delta_4 = \lambda.$$

Note that, by applying balance of flow to the *whole system*, we immediately get $\delta = \delta_2 = 2\lambda$. Now assume that each workstation has a single machine, and the raw process times are equal to $t_{01} = 3$, $t_{02} = 2$, $t_{03} = 1.8$ and $t_{04} = 5.4$ hours. Then the utilization of each workstation is

$$u_1 = 4.2\lambda, \quad u_2 = 4\lambda, \quad u_3 = \frac{3.06}{0.86}\lambda, \quad u_4 = 5.4\lambda.$$

Hence, workstation W_4 is the bottleneck, and the maximal inflow rate is $\lambda_{\max} = \frac{1}{5.4} = 0.185$ jobs per hour, and the corresponding maximal throughput is $\delta_{\max} = 2\lambda_{\max} = 0.37$ jobs per hour. Note that if we increase λ beyond $\lambda_{\max}$, then the WIP in W_4 will increase without bound, but the throughput δ will stay at level $\delta_{\max}$.



machine	t_0 [hrs]
M_1	2.0
M_2	6.0
M_3	1.8
M_4	1.6

Figure 4.9: Manufacturing system with rework and bypassing

Exercise 43. (*Exercise 2.3.1 [1]*) Consider the manufacturing system with rework and bypassing in Figure 4.9. The manufacturing system consists of three buffers and four machines. Lots are released at a rate of λ lots/hour. The numbers near the arrows indicate the fraction of the lots that follow that route. For instance, of the lots leaving buffer B_1 90% goes to machine M_1 and 10% goes to buffer B_3. The process time of each machine is listed in the table in Figure 4.9.

1. Express the throughput of machine M_1, M_3, and M_4 in terms of λ.
2. Express the throughput of machine M_2 in terms of λ.
3. Express the utilisation of each machine in terms of λ.
4. What machine is the bottleneck? Determine the maximum throughput of this system.

Exercise 44. (*Exercise 2.3.2 [1]*) Consider a three-workstation flowline. The workstations each have an infinite buffer and contain 1, 3, and 2 machines respectively. The process time of the machines in workstation 1,2 and 3 is 0.9, 3.0, and 1.9 hours respectively.

1. Calculate the maximum throughput of the line.
2. Use a lot-time-diagram to determine the flowtime of a lot for release rates under the maximum throughput.
3. Determine the mean WIP level in the line for the maximum throughput.

Exercise 45. (*Exercise 2.3.3 [1]*) We have a three-workstation flowline. Each workstation consists of an infinite buffer and a number of machines. The number of machines has still to be determined. The machines in workstation 1,2, and 3 have a process time of 0.10, 0.15, and 0.06 hours respectively. We want to establish a throughput of 60 lots/hour.

1. Determine the flowtime of a lot for release rates under the maximum throughput.

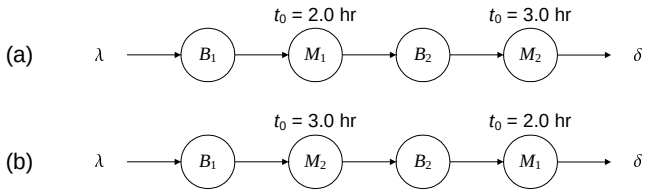


Figure 4.10: Two flowline configurations

2. Determine the number of machines in workstation 1,2, and 3 required to attain the desired throughput of 60 lots/hour.
3. Which workstation (what workstations) is (are) the bottleneck?

Exercise 46. (*Exercise 2.3.4 [1]***)** Consider a two-workstation flowline with two machines. The two machines have a process time of 2.0 and 3.0 hours respectively. The machines may be placed in any order. The two alternatives are shown in Figure 4.10.

1. What is $\delta_{\max}$ for the line in Figure 4.10(a)?
2. What happens when we set $\lambda > \delta_{\max}$?
3. What is $\delta_{\max}$ for the line in Figure 4.10(b)?
4. What happens when we now set $\lambda > \delta_{\max}$?

4.4 Little's law

Little's law is a fundamental relationship among WIP, cycle time and throughput. Define (see Figure 4.11)

- w is the average WIP level in the system,
- δ is the throughput of the system,
- φ is the average flow time in the system.

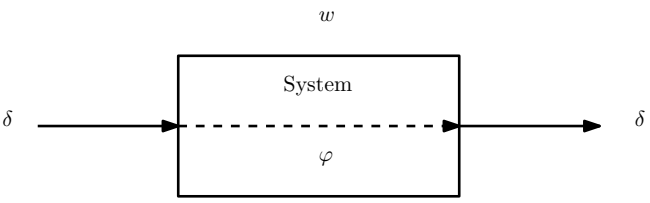


Figure 4.11: System with WIP w, throughput δ and average flow time φ

Then Little's law states that

$$w = \delta\varphi,$$

and this holds for *any system*, as long as it is *stable*, i.e., the inflow rate is equal to the throughput rate (since otherwise, if it exceeds the throughput rate, then the WIP will grow without bound).

Little's law is most easily seen from Figure 4.12.

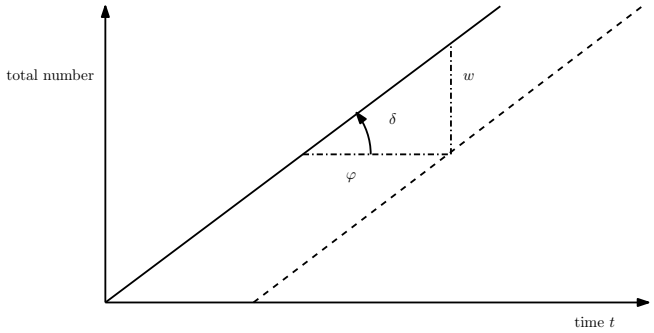


Figure 4.12: Illustration of Little's law, where the solid line is the total input to the system in $(0, t)$, and the dashed line the total output in $(0, t)$

The power of Little's law is its generality (no specific assumptions for the system are required) and flexibility (as it applies to any system). For example, it can be applied to the buffer of a specific workstation W, in which case it gives the relation

$$\text{WIP in the buffer of workstation } W = \text{Throughput of } W \times \text{Time spent in buffer},$$

and when it is applied to the whole manufacturing system, we get

$$\text{WIP in the whole manufacturing system} = \text{Throughput of the system} \times \text{Cycle time}.$$

According to Little's law, the same throughput δ can be achieved with

- large WIP w and long flow times φ, but also with
- small WIP w and short flow times φ.

What causes the difference? In most cases the answer is: Variability! The concept of variability will be studied in the following sections, and in particular, its corrupting effect on system performance.

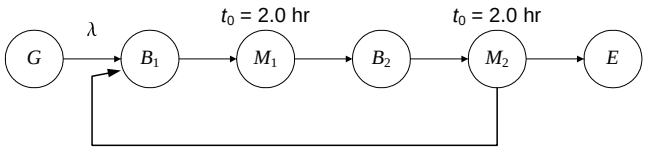


Figure 4.13: Re-entrant flowline

Exercise 47. (*Exercise 2.3.5 [1]*) Consider the re-entrant flowline in Figure 4.13. Lots are released in the line by generator G at a rate of λ lots/hour. Each lot passes through the flowline twice and has the same fixed route, namely M_1, M_2, M_1, M_2. If a newly released lot and a re-entrant lot arrive at a machine at the same time, the re-entrant lot is processed first.

1. Express the utilisation of machine 1 in terms of release rate λ and determine the maximum throughput.
2. Construct a lot-time diagram (in which the state of individual lots is shown over time) for $\lambda = \frac{1}{5}$ lots/hour and determine the (mean) flowtime of a lot.
3. From the lot-time diagram, derive a WIP-time diagram (in which the WIP is shown over time), and determine the mean WIP-level in steady state.
4. Verify your result using Little's law.

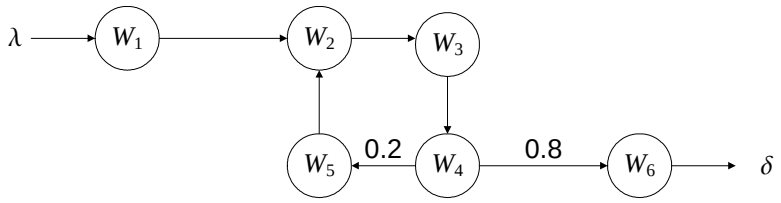


Figure 4.14: Manufacturing line

Exercise 48. (*Exercise 2.4.1 [1]*) The manufacturing system in Figure 4.14 consists of 6 workstations. 20% of the lots processed by workstation W_4 need rework. The lot is first stripped in workstation W_5 and then reprocessed by workstation W_2 and W_3. The numbers at the arrows indicate the fraction of the lots that follow that route.

1. Calculate the total throughput of workstation 2,3, and 5 in terms of release rate λ.
2. Calculate the throughput $\delta_{W_4W_5}$ and $\delta_{W_4W_6}$ for workstation 4.
3. Verify that conservation of mass holds for system $W_2W_3W_4W_5$.

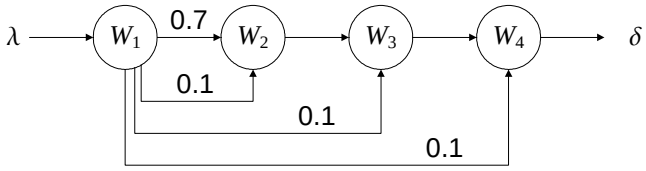


Figure 4.15: Manufacturing line with bypassing

Exercise 49. (*Exercise 2.4.2 [1]*) Figure 4.15 shows a flow line with bypassing loops going from workstation 1 to workstation 2, 3, and 4. The numbers at the arrows indicate the fraction of the lots that follow that route.

1. Use mass conservation to determine the throughput of workstations 1,2,3, and 4.
2. Verify that conservation of mass holds for system $W_1W_2W_3W_4$.

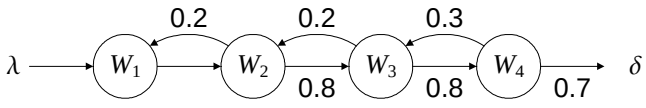


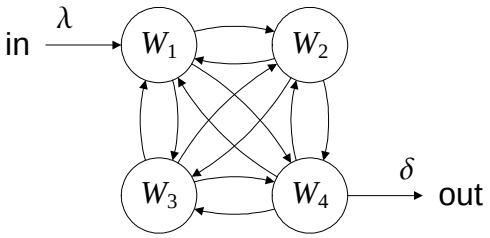
Figure 4.16: Manufacturing line with rework

Exercise 50. (*Exercise 2.4.3 [1]*) Figure 4.16 shows a flowline of workstations with rework loops. 20% of the lots processed by workstation 2 need rework on workstation 1 and 2. 20% of the processed by workstation 3 need rework on workstation 2 and 3. Finally, 30% of the lots processed by workstation 4 need rework on 3 and 4.

1. Intuitively, what workstation will have to process the most lots?
2. Write down the mass conservation equations for workstation 1 through 4.

Mass conservation on system $W_1W_2W_3W_4$ yields that $\delta = \lambda$. For the total throughput of workstation 4 we have $\delta_{W_4} = \delta + \delta_{W_4W_3}$.

1. Calculate δ_{W_4} in terms of λ.
2. Calculate the throughput of each workstation in terms of λ.
3. Assuming that the raw process time t_0 for each workstation is 1.0 hours, what workstation is the bottleneck? (Does this correspond to your initial guess in a)?)
4. What is the maximum throughput of this line?



		to				
		1	2	3	4	out
from	in	1.0				
	1		0.1	0.8	0.1	
	2	0.2		0.7	0.1	
	3	0.1	0.2		0.7	
	4	0.0	0.2	0.3		0.5

Figure 4.17: (a)Job shop, (b) From-to-matrix

Exercise 51. (*Exercise 2.4.4 [1]***)** Figure 4.17 shows a so-called job shop. In a job shop lots can each have very different routings. The from-to-matrix is used to indicate the fraction of the lots that go from a specific workstation to another workstation. For example, 10% of lots that finish processing on workstation 1 go to workstation 2, 80% goes to workstation 3, and 10% goes to workstation 4. All lots enter the job shop via workstation 1 and leave the job shop via workstation 4.

1. Intuitively, what workstation processes the most lots in this job shop?
2. Write down the mass conservation equations for each workstation.
3. Show that we can write these equations as the following matrix equation.

$$\begin{bmatrix} -1 & 0.2 & 0.1 & 0 \\ 0.1 & -1 & 0.2 & 0.2 \\ 0.8 & 0.7 & -1 & 0.3 \\ 0.1 & 0.1 & 0.7 & -1 \end{bmatrix} \begin{bmatrix} \delta_{W_1} \\ \delta_{W_2} \\ \delta_{W_3} \\ \delta_{W_4} \end{bmatrix} = \begin{bmatrix} -\lambda \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

4. Solve the matrix equation (for example using Matlab).
5. How can you easily verify that the throughput δ_{W_4} is indeed 2λ? Do the results correspond to your intuition built in part a)?
6. The process time t_0 for workstation 1,2,3, and 4 is 1.0, 1.3, 0.7, and 0.8 hours respectively. Which workstation is the bottleneck?

4.5 Variability

Variability may be formally defined as the *quality of non-uniformity of entities* [2], and it is closely related to randomness. Therefore, to understand the effect of variability, we must understand the concept of randomness which is studied in Probability Theory, and its basic results are presented in Chapter 3 of these lecture notes.

Distinction can be made between two two types of variability:

Controllable variation. This is the result of (bad) decisions. For example, variability is introduced by the mix of products produced in the plant, where each type of product can have its own processing characteristics and routing through the plant. Another example is use of batch transportation of material, where the first finished part in the batch has to wait longer for transportation than the last one in the batch.

Random variation. This is the result of events beyond our control. For example, the time elapsing between customer demands, or machine failures.

Our intuition is quite good with respect to first-moment effects (i.e., the mean): we understand that we get more products out by speeding up the bottleneck machine, or by adding more product carriers. This type of intuition is based on a deterministic world. However, most of us have a much less developed intuition for second-moment effects (i.e., the variance). For example:

- Which is more variable: the time to process an individual part or the time to produce a whole batch of those parts?
- Which results in greater improvement of line performance: Reduce variability of process times closer to raw materials (upstream the line), or closer to the customers (downstream the line)?
- Which are more disruptive to line performance: Short frequent machine failures, or long infrequent machine failures?

In the next section we will first pay attention to understanding the mean and variance in process times and flows, and then to the interaction of these two sources of randomness.

4.6 Process time variability

Instead of looking at the "clean or natural" process time, it makes sense to consider the *effective process time*, which is the total time seen by a job at a workstation. This includes the natural process time, setups, rework, operator unavailability, and all sorts of other shop floor realities. The reason is that from a logistical point of view, it only matters that the job is at the machine, not so much the reason why (for example, because it is actually being processed or because it is waiting for material).

A standard measure for variability is the standard deviation σ. This absolute measure does not immediately tell you whether variability is low or high, unless you relate it to, for example, the mean. For example, $\sigma = 1$ hour is big for a mean process time $t = 10$ minutes, but it is small for $t = 10$ hours. A useful relative measure for variability is the coefficient of variation c, defined as

$$c = \frac{\sigma}{t},$$

where t is the mean and σ the standard deviation. We say that variability is *low* if $c < 0.75$, *moderate* if $0.75 \leq c < 1.33$ and *high* if $c \geq 1.33$. There are various sources of process time variability, such as:

- "Natural" variability;
- Random outages;
- Setups;
- Operator (un)availability;
- Rework.

4.6.1 Natural variability

Natural variability is the variability inherent in the natural process time, and can be caused by differences in skills and experience of operators, machines, composition of material that is processed and so on. There is usually more natural variability in manual processes than in automated ones. The natural coefficient of variation c_0 is defined as

$$c_0 = \frac{\sigma_0}{t_0},$$

where t_0 and σ_0 are mean and standard deviation of natural process time. Natural process times typically have low variability: $c_0 < 0.75$.

4.6.2 Preemptive outages

Preemptive outages are uncontrollable down-times, such as machine breakdowns, power downs, operators being called away, running out of raw material, and so on. They can occur at random during processing of jobs. Its effect on the machine capacity can be evaluated by calculating the availability, which is the long-run fraction of time the machine is available,

$$A = \frac{m_f}{m_f + m_r},$$

where m_f is the *mean time to failure* and m_r is the *mean time to repair*. The mean natural process time t_0 can now be adjusted to account for the availability. This results in the *effective mean process time* t_e,

$$t_e = \frac{t_0}{A} \tag{4.1}$$

and the *effective capacity* of a workstation with m machines is

$$r_e = \frac{m}{t_e} = A\frac{m}{t_0} = Ar_0,$$

where $r_0 = \frac{m}{t_0}$. The effect of random breakdowns on the variability of the effective process time is more subtle. To quantify this effect, we assume the following:

- The time to failure is *Exponential*. This means that failures are truly unpredictable, in the sense that in every small time interval $(t, t + \Delta)$

  $$\mathrm{P}(\text{failure in } (t, t + \Delta)) = r_f \Delta$$

  where $r_f = \frac{1}{m_f}$ is the failure rate.
- After repair, there are several possibilities to resume processing. For example, processing of the job may need to start all over again, but here we assume that processing resumes at the point where it was interrupted by the failure.

The following example illustrates the calculation of the effective process time.

Example 4.4. (Calculation of effective process time under preemptive outages)
Suppose that the natural process time of a job takes $X_0 = 10$ minutes, and that from the start of the job, the first failure occurs after $F_1 = 3.5$ minutes, which takes a repair time of $R_1 = 6$ minutes. After repair, processing resumes and the residual process time is 6.5 minutes. Then after $F_2 = 5$ minutes the second failure occurs which requires a repair time of $R_2 = 2.5$ minutes. Then processing resumes again and, since the time till the next failure $F_3 = 3$ minutes, which is longer than 1.5 minutes, the job completes processing without any further delay. Hence, the effective process time for this job is $X_e = X_0 + R_1 + R_2 = 18.5$ minutes, see Figure 4.18

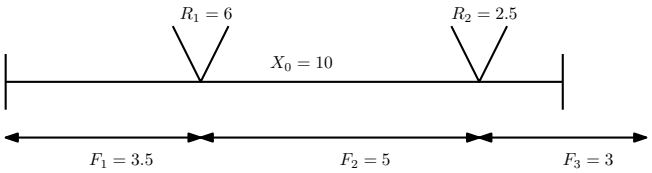


Figure 4.18: Calculation of effective process time $X_e = X_0 + R_1 + R_2 = 18.5$ minutes

Based on the above assumptions, the following result can be derived for the variance of the effective process time,

$$\sigma_e^2 = \left(\frac{\sigma_0}{A}\right)^2 + \frac{(m_r^2 + \sigma_r^2)(1-A)t_0}{Am_r}, \tag{4.2}$$

where σ_r^2 is the variance of the repair time. Hence, for the squared coefficient of variation we get

$$\begin{aligned} c_e^2 = \frac{\sigma_e^2}{t_e^2} &= c_0^2 + (1 + c_r^2)A(1-A)\frac{m_r}{t_0} \\ &= c_0^2 + A(1-A)\frac{m_r}{t_0} + c_r^2 A(1-A)\frac{m_r}{t_0}, \end{aligned} \tag{4.3}$$

where $c_r = \frac{\sigma_r}{m_r}$ is the coefficient of variation of the repair time. The first term in (4.3) is due to the natural variability, the second one is due to the occurrence of random breakdowns, and the third one accounts for variability in the repair times. Note that c_e^2 increases in m_r, so long repair times induce more variability than short ones. The following example looks into the question: What is better from the viewpoint of variability, short frequent machine failures, or long infrequent ones?

Example 4.5. Consider two machines, M_1 and M_2. For machine M_1 we have $t_0 = 15$, $\sigma_0 = 3.35$, $c_0 = 0.223$, $m_f = 744$, $m_r = 248$ minutes and $c_r = 1$. For M_2 the parameters are $t_0 = 15$, $\sigma_0 = 3.35$, $c_0 = 0.223$, $m_f = 114$, $m_r = 38$ minutes and $c_r = 1$. So M_1 has infrequent long stops, M_2 has frequent short ones. For both machines, the availability is

$$A = 0.75,$$

so $t_e = 20$ minutes. Hence, both machines have the same effective capacity $r_e = \frac{1}{t_e} = \frac{1}{20}$ jobs per minute (or 3 jobs per hour). From (4.3) we obtain that for machine M_1,

$$c_e^2 = 6.25$$

and for M_2,

$$c_e^2 = 1.$$

So machine M_1 exhibits much more variability than M_2!

4.6.3 Non-Preemptive outages

Non-preemptive outages do not randomly occur during the processing of jobs, but rather before or after jobs and we typically have some control as to exactly when. Examples of non-preemptive outages are tool changes, setups, preventive maintenance (as opposed to corrective maintenance, which is repair), shift changes, and so on. The effect of non-preemptive outages on capacity and variability can be evaluated along the same lines as for preemptive outages. For example, suppose that a machine needs a setup with mean t_s and coefficient of variation c_s after having produced *on average* N_s jobs. This means that, after having produced a job, the machine needs a setup (with mean t_s and coefficient of variation c_s) *with probability* $\frac{1}{N_s}$. It then follows for the mean and variance of the effective process time that (see Example 3.37)

$$t_e = t_0 + \frac{t_s}{N_s}, \tag{4.4}$$

$$\sigma_e^2 = \sigma_0^2 + \frac{\sigma_s^2}{N_s} + \frac{N_s - 1}{N_s^2} t_s^2. \tag{4.5}$$

Example 4.6. Suppose we have two machines, M_1 and M_2. Machine M_1 is flexible and requires no setups: $t_0 = 1.2$ hours, $c_0 = \frac{1}{2}$. Machine M_2 is fast, but it needs a setups, on average after 10 jobs: $t_0 = 1$ hour, $c_0 = \frac{1}{4}$, $N_s = 10$, $t_s = F2$ hours, $c_s = 0.25$. Note that the flexible machine exhibits a higher natural variability than the fast one. For M_1 we have $t_e = 1.2$ hours, and for M_2 we obtain $t_e = 1 + \frac{1}{10} \cdots 2 = 1.2$ hours from (4.4). Hence both machines have the same effective capacity. But which one is less variability? For M_1,

$$c_e^2 = c_0^2 = 0.25$$

and for M_2 we find

$$c_e^2 = \frac{\sigma_e^2}{t_e r} = \frac{\left(\frac{1}{4}\right)^2 + \frac{\left(\frac{1}{8}\right)^2}{10} + \frac{9}{100} \cdot 2^2}{\left(1 + \frac{1}{10} \cdot 2\right)^2} = 0.29.$$

So the flexible machine M_1 exhibits less variability than M_2.

4.6.4 Rework

Another source of variability is quality problems, and its effect can be quantified in the same way as in the previous sections. Suppose that a workstation performs a task, and then checks whether it has been done correctly. If not, the task is repeated until it is eventually correct. The probability that it is correct is q, so on average, every taks has to be repeated $N_r = \frac{1}{q}$ times. If the mean natural task time is t_0 with standard deviation σ_0, then we obtain for the effective task time,

$$t_e = N_r t_0 \tag{4.6}$$

$$\sigma_e^2 = N_r \sigma_0^2 + N_r(N_r - 1)t_0^2 \tag{4.7}$$

$$c_e^2 = \frac{c_0^2}{N_r} + \frac{N_r - 1}{N_r}$$

4.7 Flow variability

Another source of variability, besides process time variability, is the variability in job flows through the system. A job flow refers to the transfer of jobs from one station to another. Considering a single workstation, we can distinguish the arrival flow of jobs to this workstation and the departure flow from this station, see Figure 4.19.

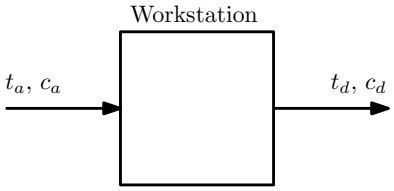


Figure 4.19: Arrival and departure flows at workstation

We start with characterizing the arrival flow. The first descriptor of job arrivals is the *arrival rate* r_a and the second one is the coefficient of variation c_a, defined as

$$r_a = \frac{1}{t_a}, \quad c_a = \frac{\sigma_a}{t_a},$$

where t_a is the mean time between arrivals or *mean inter-arrival time*, and σ_a is the standard deviation of the time between arrivals. In Figure 4.20 we display arrivals two streams, one with a low variability and the other with a high variability. The latter one clearly has a "bursty" character.

An arrival process that deserves special attention because of its practical importance is the **Poisson process**, for which the times between arrivals are independent and *Exponential* with rate λ. This proces has the following properties:

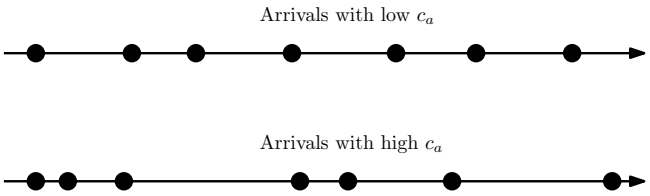


Figure 4.20: Arrival flows with low and high variability

- **Memoryless property.** Since the inter-arrival times are Exponential, and thus memoryless (see Property 3.3), we have for small $\Delta > 0$,

$$\mathrm{P}(\text{arrival in } (t, t+\Delta)) = 1 - e^{-\lambda\Delta} \approx \lambda\Delta$$

  So, in each small interval of length Δ, there is an arrival with probability $\lambda\Delta$ (and none otherwise). This means that a Poisson process is a "truly random" arrival process.

- **Binomial distribution.** By dividing the interval $(0, t)$ into many small intervals of length Δ, then we will observe in each interval 0 or 1 arrivals. Hence, the total number of arrivals in $(0, t)$ is *binomial* with parameters $n = t/\Delta$ and $p = \lambda\Delta$.

- **Poisson distribution.** Since n is large and p is small, this number is *Poisson distributed* with parameter $np = \lambda t$, see Remark 3.1. Hence, as Δ tends to 0),

$$\mathrm{P}(k \text{ arrivals in } (0, t)) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \ldots$$

  This explains the name "Poisson process".

- **Clustered arrivals.** Since the density $f(x) = \lambda e^{-\lambda x}$ is maximal for $x = 0$, short inter-arrival times occur more frequently than long ones. So arrivals tend to *cluster*, as seen in Figure **??**.

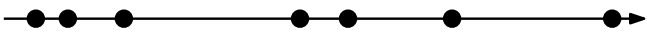


Figure 4.21: Clustering of Poisson arrivals

- **Many rare arrival flows.** The superposition of many independent rarely occurring arrival flows is close to Poisson (and the more flows, the more it will look like Poisson). This is why Poisson flows so often occur in practice!

- **Merging.** By merging two independent Poisson flows, say red arrivals with rate λ_1 and blue arrivals with rate λ_2 (see Figure 4.22), we again obtain a Poisson flow with rate $\lambda_1 + \lambda_2$, since

$$\mathrm{P}(\text{arrival in } (t, t+\Delta)) \approx \lambda_1\Delta + \lambda_2\Delta = (\lambda_1 + \lambda_2)\Delta.$$

Figure 4.22: Merging of a red and blue Poisson stream

Further, given that there is an arrival in $(t, t+\Delta)$, it is red with probability

$$\begin{aligned}
\mathrm{P}(\text{red arrival in } (t, t+\Delta) | \text{arrival in } (t, t+\Delta)) &= \frac{\mathrm{P}(\text{red arrival in } (t, t+\Delta))}{\mathrm{P}(\text{arrival in } (t, t+\Delta))} \\
&= \frac{\lambda_1\Delta}{(\lambda_1 + \lambda_2)\Delta} \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2}.
\end{aligned}$$

- **Random splitting.** Randomly splitting (or thinning) a Poisson flow with rate λ, which means that with probability p an arrival is colored red and otherwise ignored (or colored blue), yields again a red Poisson flow with rate $p\lambda$, since

$$\begin{aligned} \mathrm{P}(\text{red arrival in } (t, t+\Delta)) &= \mathrm{P}(\text{arrival is colored red}|\text{ arrival in } (t, t+\Delta))\mathrm{P}(\text{arrival in } (t, t+\Delta)) \\ &\approx p\lambda\Delta. \end{aligned}$$

Example 4.7. Suppose type A and B jobs arrive at machine M for processing. Both job types arrive according to Poisson flows, type A with rate 2 and type B with rate 3 jobs per hour.

- What is the probability that during 1 hour no jobs arrive? The flow of A and B together is again a Poisson flow with rate $2+3=5$ jobs per hour. Since the number of arrivals in an interval $(0, t)$ is Poisson distributed with parameter $5t$, we have

$$\mathrm{P}(\text{no arrival in } (0,1)) = e^{-5}.$$

- What is the probability that the next job to arrive is type A? Given that a job arrives in $(t, t+\Delta)$ it is of type A with probability $\frac{2}{2+3} = \frac{2}{5}$.
- What is the probability that during 2 hours at least 2 type B jobs arrive? The number of type B arrivals in the interval $(0, 2)$ is Poisson distributed with parameter $3 \cdot 2 = 6$. Hence,

$$\begin{aligned} \mathrm{P}(\text{at least 2 arrivals in } (0,2)) &= 1 - \mathrm{P}(0 \text{ or } 1 \text{ arrivals in } (0,2)) \\ &= 1 - e^{-6}(1+6) \;=\; 1 - 7e^{-6}. \end{aligned}$$

Remark 4.1. (General arrival flows)

- **Merging.** By merging two independent arrival flows, say red arrivals with mean inter-arrival time $t_a(1)$ and coefficient of variation $c_a(1)$, and blue arrivals with mean inter-arrival time $t_a(2)$ and coefficient of variation $c_a(2)$, we obtain an arrival flow with

$$r_a(\text{merged}) = r_a(1) + r_a(2) = \frac{1}{t_a(1)} + \frac{1}{t_a(2)}.$$

  The coefficient of variation of the time between arrivals of the merged flow is hard to estimate. A simple (though rough) approximation for the squared coefficient of variation is

$$c_a^2(\text{merged}) = \frac{r_a(1)}{r_a(1)+r_a(2)} c_a^2(1) + \frac{r_a(2)}{r_a(1)+r_a(2)} c_a^2(2).$$

  It should be noted, however, that the inter-arrival times of the merged flow are, in general, no longer independent.

- **Random splitting.** Randomly splitting (or thinning) an arrival flow with mean inter-arrival time t_a and coefficient of variation c_a, which means that with probability p an arrival is colored red and otherwise ignored (or colored blue), yields a new red arrival flow with

$$r_a(\text{red}) = pr_a = \frac{p}{t_a}, \quad t_a(\text{red}) = \frac{t_a}{p}, \quad c_a^2(\text{red}) = pc_a^2 + 1 - p.$$

We now look at the departure flow. The same measures can be used to describe departures, namely the *departure rate* r_d and the coefficient of variation c_d of the time between departures, defined as

$$r_d = \frac{1}{t_d}, \quad c_d = \frac{\sigma_d}{t_d},$$

where t_d is the mean time between departures or *mean inter-departure time*, and σ_d is the standard deviation of the time between departures, see Figure 4.19.

Clearly, $r_d = r_a$ by conservation of flow. Variability in departures from a workstation depends on both variability in arrivals and process times at that workstation. The relative contribution of these two sources of variability depends on the utilization of the workstation, defined as

$$u = \frac{r_a t_e}{m},$$

where m is the number of machines in the workstation. A simple approximation for c_d^2 in a single-machine workstation ($m = 1$) is

$$c_d^2 = (1 - u^2)c_a^2 + u^2 c_e^2,$$

and for multi-machine stations ($m \geq 1$),

$$c_d^2 = 1 + (1 - u^2)(c_a^2 - 1) + \frac{u^2}{\sqrt{m}}(c_e^2 - 1). \tag{4.8}$$

This approximation for c_d makes sense, since, if $m = 1$ and u is close to 1, then the machine nearly always busy, so

$$c_d \approx c_e.$$

On the other hand, if u is close to 0, then t_e is very small compared to t_a, so

$$c_d \approx c_a.$$

In serial production lines, all departures from workstation i are arrivals to the next workstation $i + 1$, so

$$t_a(i + 1) = t_d(i), \quad c_a(i + 1) = c_d(i),$$

where $t_a(j)$, $c_a(j)$, $t_d(j)$, $c_d(j)$ are the descriptors of arrivals and departures at workstation j, see Figure 4.23.

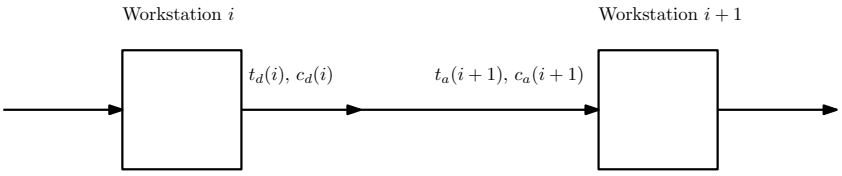


Figure 4.23: Departures and arrival flows in serial production lines

Remark 4.2.

- Successive times between departures are typically *not independent*. For example, an inter-departure time following a very long one will most likely correspond to a process time (since it is likely that during the long inter-departure time at least one job arrived, which can immediately enter service upon the departure instant). However, the assumption of independence is usually a reasonable approximation.
- If both inter-arrival times and process times are independent and *Exponential* (which implies that $c_a = c_e = 1$), then it can be shown that inter-departure times are also independent and Exponential (so $c_d = 1$). The above simple approximation agrees with this property, i.e., if $c_a = c_e = 1$, then the approximation indeed yields $c_d = 1$.

4.8 Variability interactions - Queueing

Process time variability and flow variability are the building blocks for describing effects of variability in manufacturing systems. The question is: How to use these building blocks to evaluate the impact of

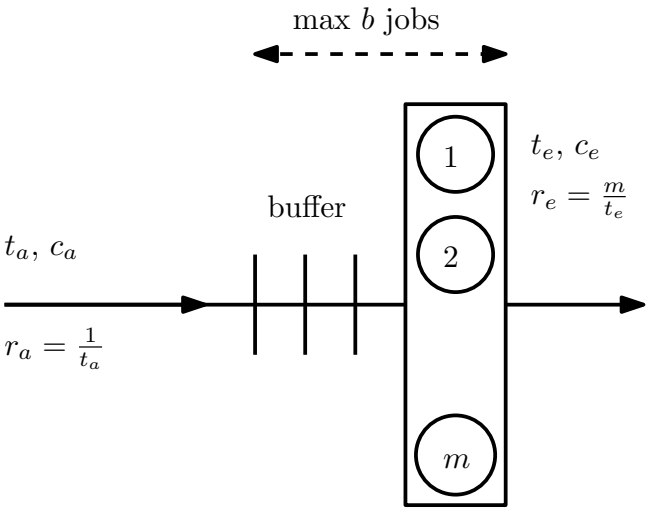


Figure 4.24: Queueing system with parameters

variability on key performance indicators such as WIP, cycle time and throughput? The total process time is often a small part of the cycle time, where the extra major part of the cycle time is due to *waiting*. So the fundamental issue of Factory Physics is to understand the underlying causes of waiting. The science of waiting is Queueing Theory. This theory studies the interactions and consequences of sources of variability through mathematical modeling. Most queueing systems consists of three components:

- **Arrival process:** Arrivals can consist of single jobs or batches, and there may a single flow or multiple flows of different jobs.
- **Service (or production) process:** The workstation can have a single machine, or multiple (non-) identical ones. Processing discipline can be first-come first-served (FCFS), last-come first-served (LCFS), shortest process time first (SPTF), random, and so on.
- **Queue (or buffer):** There may be ample (unlimited) queue space for jobs, or limited (or even no) queue space.

Clearly there is a large variety of queueing systems. A convenient notation to characterize queueing systems is the $A/B/c$ notation, due to Kendall, where the letter A specifies the distribution of the inter-arrival times, B the distribution of the service times and c the number of servers. Typical values for A and B are: D (Deterministic), M (Memoryless or Exponential) and G (General, i.e., this may be any distribution). So an $M/M/1$ is a single Exponential server serving a Poisson stream of jobs. If only three letters are used, the implicit assumption is that service is FCFS and queueing space is infinite. Otherwise, letters are added. For example, in an $M/M/1/b$ there is room for at most b jobs in the system. For queueing systems specified with parameters (see Figure 4.24)

- t_a, mean time between arrivals,
- c_a, coefficient of variation of inter-arrival times,
- $r_a = \frac{1}{t_a}$, rate of arrivals,
- m, number of parallel identical machines in station,
- b, buffer size (i.e., maximum number in station),
- t_e, mean effective process time,
- c_e, coefficient of variation of the effective process time,

- $r_e = \frac{m}{t_e}$, capacity of station,

we aim at estimating system performance in terms of

- p_n, probability (or long-run fraction of time) of having n jobs at station,
- φ_B, expected waiting time (or flow time) spent in buffer,
- φ, expected cycle time or flow time in station,
- w_B, average WIP in buffer,
- w, average WIP in station,
- δ, throughput of station.

Note that these measures are related as follows,

$$w = \sum_{n=0}^{\infty} n p_n, \quad w = w_B + mu, \quad \varphi = \varphi_B + t_e, \quad w = \delta\varphi, \quad w_B = \delta\varphi_B,$$

where u is the machine utilization. In case queueing space is unlimited ($b = \infty$), we have $\delta = r_a$ (since outflow is then equal to inflow) and the utilization u is given by

$$u = \frac{r_a}{r_e} = \frac{r_a t_e}{m} = \frac{\delta t_e}{m}.$$

In the next section we start with the zero-buffer model.

4.9 Zero-buffer model

We consider a special production line, with two machines and no buffer in between, see Figure 4.25.

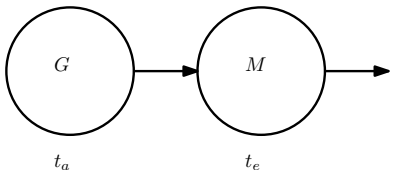


Figure 4.25: Zero-buffer model with machines G and M

The first machine is G (i.e., the generator of arrivals), and the second one M. Machine G is never starved (i.e., there is always raw material available). If machine G completes a job, and M is still busy, then G blocks and has to wait till M finishes the job, before it can move the job to M and start processing the next one. Machine M never blocks (i.e., it can always get rid of the job), but may have to wait for input from G. The mean process time of G is t_a, and the mean process time of M is t_e. What is the throughput δ? The throughput can be easily estimated the following $\chi 3.0$ model (the complete code is in Appendix B).

```
1  model real GME():
2      real ta = 1.0, te = 1.0;
3      int n = 1000;
4      chan job a, b;
5
6      run G(a, constant(ta)), M(a, b, constant(te)), E(b, n)
7  end
```

In this model we specified constant process times with mean 1. Process E is only counting the number of completed jobs (and once $n = 1000$ jobs are completed, the program stops and prints the throughput). Obviously, for constant $t_a = t_e = 1$, we get $\delta = 1$. However, if we now change to Exponential process times with mean 1, then the χ3.0 model produces the output $\delta = \frac{2}{3}$. So the throughput of the production line drops with 33%. Why? The reason is variability!

Let the random variable A be the process time of G and B the process time of M. Then the time between two departures is $\max\{A, B\}$ and

$$\delta = \frac{1}{\mathsf{E}[\max\{A, B\}]} \leq \frac{1}{\max\{\mathsf{E}[A], \mathsf{E}[B])\}},$$

with equality only for constant A and B. It is the randomness in A and B that leads to *starvation* of M ($A > B$) and *blocking* of G($A < B$), and thus to capacity loss. To deal with variations in these process times, we should use buffers! Indeed, if the buffer between G and M is sufficiently large, then

$$\delta = \frac{1}{\max\{\mathsf{E}[A], \mathsf{E}[B]\}}.$$

This can be readily verified by the following χ3.0 model (see Appendix C), to which a buffer between G and M has been added.

```
1  model real GBME():
2      real ta = 1.0, te = 1.0;
3      int n = 1000, N = 10;
4      chan job a, b, c;
5
6      run G(a, exponential(ta)), B(a, b, N), M(b, c, exponential(te)), E(c, n)
7  end
```

For a buffer of size $N = 10$, the estimated throughput is already around 0.93. In the next section we study the finite buffer model with Exponential process times, and derive an exact formula for the throughput as a function of the buffer size.

Remark 4.3. For zero buffer and Exponential process times the throughput can be exactly determined. Suppose A and B are Exponential with rate λ_1 and λ_2. Then

$$\mathsf{P}(\max\{A, B\} \leq t) = \mathsf{P}((A \leq t)\mathsf{P}(B \leq t) = (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t}) = 1 - e^{-\lambda_1 t} - e^{-\lambda_2 t} + e^{-(\lambda_1+\lambda_2)t}.$$

Hence, the density of $\max\{A, B\}$ is

$$f(t) = \frac{d}{dt}F(t) = \lambda_1 e^{-\lambda_1 t} + \lambda_2 e^{-\lambda_2 t} - (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)t},$$

so

$$\mathsf{E}[\max\{A, B\}] = \int_{t=0}^{\infty} tf(t)dt = \frac{\lambda_1^2 + \lambda_1\lambda_2 + \lambda_2^2}{\lambda_1\lambda_2(\lambda_1 + \lambda_2)},$$

and thus we get for the throughput

$$\delta = \frac{1}{\mathsf{E}[\max\{A, B\}]} = \frac{\lambda_1\lambda_2(\lambda_1 + \lambda_2)}{\lambda_1^2 + \lambda_1\lambda_2 + \lambda_2^2}.$$

4.10 Finite-buffer model

We now consider the production line with machines G and M, with a finite buffer of size $b - 2$ in between. When the first machine G completes a job, and the buffer is full, then G blocks and has to wait till M finishes the job, before it can move the job into the buffer and start processing the next one. The process times of G and M are Exponential with rates $r_a = \lambda$ and $r_e = \mu$ (so $t_a = \frac{1}{\lambda}$ and $t_e = \frac{1}{\mu}$.

The number of jobs in the system can be any number from 0 up to b, where we only count the ones that have completed processing on G. When the buffer and machine M are both empty, the number in the system is 0, and M is starved. If the number in the system is b, then the buffer is full and machine G is blocked. Let p_n be the probability (or long-run fraction of time) of having n jobs in the system. These probabilities can be determined through balance equations stating that, in equilibrium, the number of transitions per unit time from state $n-1$ to n is equal to the number from n to $n-1$, i.e.,

$$\text{Flow from state } n-1 \text{ to } n = \text{Flow from state } n \text{ to } n-1.$$

To derive the flows it is convenient to use the *flow diagram* in Figure 4.26.

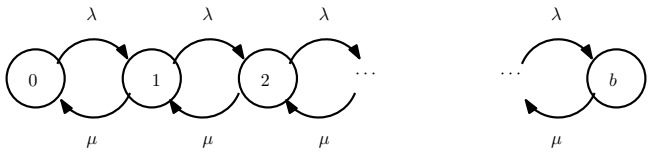


Figure 4.26: Flow diagram for $M/M/1/b$ system

To calculate the number of transitions per unit time from $n-1$ to n, we note that, by the memoryless property (3.7), machine G generates jobs at rate λ while the system is in state $n-1$. Hence, the flow from $n-1$ to n is equal to the fraction of time the system is in state $n-1$ times the rate at which machine G generates jobs (in which case the system jumps from $n-1$ to n), so $p_{n-1}\lambda$. Similarly, the flow from n to $n-1$ is $p_n\mu$. Equating these flows yields

$$p_{n-1}\lambda = p_n\mu, \quad n = 1, \dots, b.$$

These equations can be readily solved, yielding

$$p_n = p_{n-1}\frac{\lambda}{\mu} = p_{n-2}\left(\frac{\lambda}{\mu}\right)^2 = \cdots = p_0\left(\frac{\lambda}{\mu}\right)^n, \quad n = 1, \dots, b.$$

If we set $u = \frac{\lambda}{\mu}$, then these equations can be written as

$$p_n = p_0 u^n, \quad n = 1, \dots, b$$

and p_0 follows from the requirement that the probabilities $p_0, \dots, p_b$ add up to 1,

$$1 = \sum_{n=0}^{b} p_n = \sum_{n=0}^{b} p_0 u^n = p_0 \frac{1-u^{b+1}}{1-u},$$

so

$$p_0 = \frac{1-u}{1-u^{b+1}}. \tag{4.9}$$

Note that if $u = 1$, this equation simplifies to

$$p_0 = \frac{1}{b+1}.$$

The utilization of machine M is $1-p_0$ and the throughput follows from

$$\delta = \mu(1-p_0) = \mu\left(1 - \frac{1-u}{1-u^{b+1}}\right).$$

The WIP is given by

$$w = \sum_{n=0}^{b} np_n = \frac{u}{1-u} - \frac{(b+1)u^{b+1}}{1-u^{b+1}}.$$

and the average flow time φ can then be calculated by Little's law, $\varphi = \frac{w}{\delta}$. Note that the flow time starts when the job completes processing on G (and then attempts to enter the buffer) and it ends when the job leaves machine M. To investigate the effect of blocking and starvation, we also calculate the performance in case buffer space between G and M is unlimited, assuming $u < 1$ (since otherwise, the number in the system will grow to infinity). For $b = \infty$ we immediately obtain

$$\delta = \mu u, \quad w = \frac{u}{1-u}, \quad \varphi = \frac{\frac{1}{\mu}}{1-u}. \tag{4.10}$$

By comparing the finite and infinite buffer systems, we can conclude that:

- Finite buffers always force stability, regardless the rates λ and μ.
- WIP is always less than in the infinite buffer system,

$$w < \frac{u}{1-u}.$$

- Throughput is always less than in the infinite buffer system,

$$\delta < \mu u.$$

So the lesson is that the only way to reduce WIP without sacrificing too much throughput is: variability reduction!

Example 4.8. Consider the two machine line with $1/\lambda = 21$ minutes and $1/\mu = 20$ minutes (see Section 8.7.1. in [2]). Then $u = \frac{20}{21} = 0.9524$. For $b = \infty$, we get

$$w = 20 \text{ jobs}, \quad \delta = 0.0476 \text{ jobs per minute}, \quad \varphi = 420.14 \text{ minutes}$$

and for $b = 4$ (so two buffer places in between G and M),

$$w = 1.894 \text{ jobs}, \quad \delta = 0.039 \text{ jobs per minute}, \quad \varphi = 48.57 \text{ minutes}$$

This shows that limiting the buffer space in between the two machines greatly reduces the WIP and flow time, but at the price of also reducing the throughput, the cost of which may not be covered by the savings in inventory cost.

4.11 Single machine station

We start with the simplest interesting system, which is the $M/M/1$. So we assume a single machine with unlimited buffer space. Jobs arrive one at a time. The inter-arrival times are Exponential with rate $r_a = \lambda$. Processing is in order of arrival and jobs are processed one at a time. The process times are Exponential with rate $r_e = \mu$. Hence the utilization of the machine is

$$u = \frac{r_a}{r_e} = \frac{\lambda}{\mu},$$

which is assumed to be less than 1 (since otherwise the machine can not handle the work).

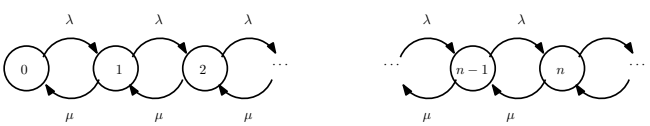


Figure 4.27: Flow diagram for $M/M/1$ system

Let p_n be the probability (or long-run fraction of time) of finding n jobs in the system, and in the same way as in the previous section, these probabilities can be determined through balance equations stating that, in equilibrium, the number of transitions per unit time from state n to $n-1$ is equal to the number from $n-1$ to n. So (see Figure 4.27)

$$p_n\mu = p_{n-1}\lambda, \quad n = 1, 2, ...$$

whence

$$p_n = p_{n-1}u = \cdots = p_0u^n = (1-u)u^n, \quad n = 0, 1, 2, ...$$

since $p_0 = 1-u$. Hence, the number of jobs in system is Geometric with parameter u. For the average WIP we get (see also (4.10))

$$w = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n(1-u)u^n = \frac{u}{1-u},$$

which, by Little's law, yields

$$\varphi = \frac{w}{\lambda} = \frac{1/\mu}{1-u} = \frac{t_e}{1-u}. \tag{4.11}$$

This shows that w and φ grow without bound as the utilization u tends to 1, which is typical for nearly every system operating close to its maximum capacity. Since $\varphi = \varphi_B + t_e$, we get from (4.11) that the mean waiting time spent in the buffer satisfies

$$\varphi_B = \frac{t_e}{1-u} - t_e = \frac{u}{1-u}t_e,$$

and thus for the average WIP in the buffer,

$$w_B = \lambda\varphi_B = \frac{u^2}{1-u}.$$

So far, we assumed Exponential inter-arrival times and process times. Now suppose that the inter-arrival times and process times are General (i.e., they can have any distribution). Then we do no longer have an exact expression for the mean waiting time in the buffer, but the following *approximation* applies,

$$\varphi_B(G/G/1) = \gamma\varphi_B(M/M/1) = \gamma \times \frac{u}{1-u} \times t_e, \tag{4.12}$$

where

$$\gamma = \frac{1}{2}(c_a^2 + c_e^2).$$

The above expression (4.12) separates φ_B into three terms $V \times U \times T$: dimensionless Variability term γ, dimensionless Utilization term $\frac{u}{1-u}$, and the process Time term t_e. Clearly, the mean waiting time decreases as the variability in arrivals and process times decreases, and eventually vanishes if there is no variability. In Figure 4.28 we show the mean waiting time as function of u for $c_a = 1$ and $c_e = 0, 1, 2$.

Below we list some useful observations:

- **M/G/1.** Expression (4.12) is exact for Poisson arrivals (i.e. Exponential inter-arrival times).
- **Insensitivity.** Mean waiting time only depends on mean and standard deviation of inter-arrival times and process times. There is no need to know the distributions of inter-arrival and process times in order to estimate the *mean* waiting time spent in the buffer.
- **Heavy load.** As the utilization u tends to 1, then:
  - φ_B tends to ∞.

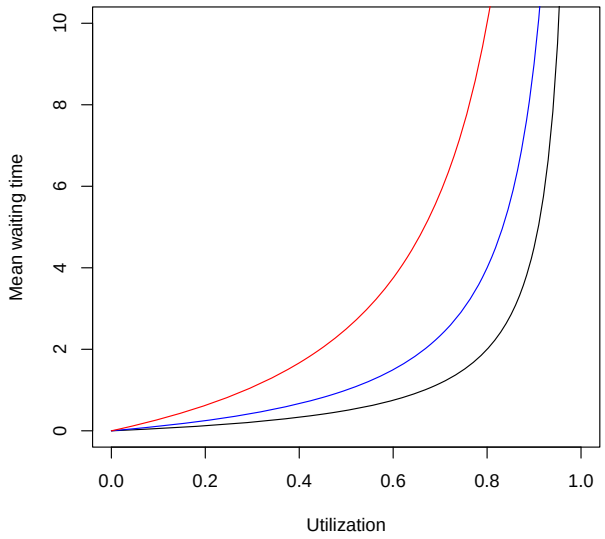


Figure 4.28: Mean waiting time φ_B as function of utilization u for $c_a = 1$ and $c_e = 0$ (black), 1 (blue), 2 (red)

- Relative error of approximation (4.12) tends to 0.
- Distribution of waiting time converges to Exponential.

Summarizing, as the system operates close to its maximum capacity, then the waiting times are long and Exponential, the mean of which is given by (4.12).

Approximation (4.12) can be readily validated by the following χ3.0 model (see Appendix D).

```
1  model GBME():
2      chan job a, b, c;
3
4      run G(a, uniform(0.0, 2.0)),
5          B(a, b), M(b, c, uniform(0.0, 1.0)),
6          E(c, 100000)
7  end
```

In this model we take Uniform inter-arrival times on $(0, 2)$ and Uniform process times on $(0, 1)$. Based on 10^5 departures, simulation produces an average waiting time of approximately 0.15. Since $t_a = 1$, $c_a^2 = \frac{1}{3}$, $t_e = \frac{1}{2}$, $c_e^2 = \frac{1}{3}$, $\gamma = \frac{1}{3}$ and $u = \frac{1}{2}$, approximation (4.12) yields $\varphi_B = \frac{1}{6}$. Another example is Constant inter-arrival times of $\frac{3}{5}$, and Exponential process times with mean $\frac{1}{2}$. Then we get $t_a = \frac{3}{5}$, $c_a^2 = 0$, $t_e = \frac{1}{2}$, $c_e^2 = 1$, $\gamma = \frac{1}{2}$ and $u = \frac{5}{6}$, so $\varphi_B = \frac{5}{4}$ where, based on 10^5 departures, the simulation estimate is 1.11. Hence, both examples show that approximation (4.12) is reasonably accurate.

By applying Little's law we can also obtain an approximation for the mean WIP in the buffer,

$$w_B(G/G/1) = \gamma \frac{u^2}{1-u},$$

and, by adding the mean number of jobs in process,

$$w(G/G/1) = \gamma \frac{u^2}{1-u} + u.$$

The above expression for w provides an estimate for the long-run average WIP in the station, but it does not tell anything about how the WIP behaves over time. In Figure 4.29 we show realizations of the WIP over time for an Exponential single machine system.

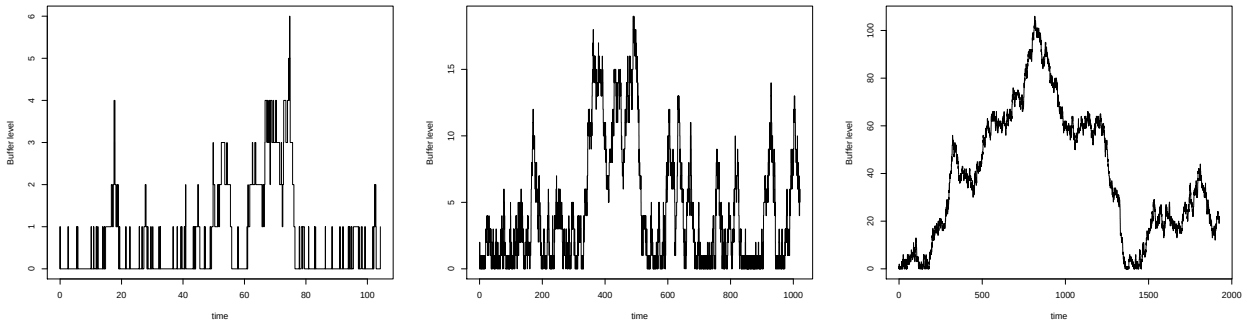


Figure 4.29: Behavior of WIP over time of Exponential single machine system, $t_a = 1.0$, $t_e = 0.5$ (left), 0.9 (middle), 0.95 (right)

The realization for $u = 0.95$ shows that, although the long-run average WIP is 19 and there is surplus capacity of 5%, the WIP is very large for very long periods! In practise, such "disasters" will not be observed, since in situations with extremely high WIP levels, one will typically try to get additional processing capacity to clear the high WIP.

4.12 Multi machine station

In this section we consider multi machine stations, and we start with the simplest one, which is the $M/M/c$, see Figure 4.31.

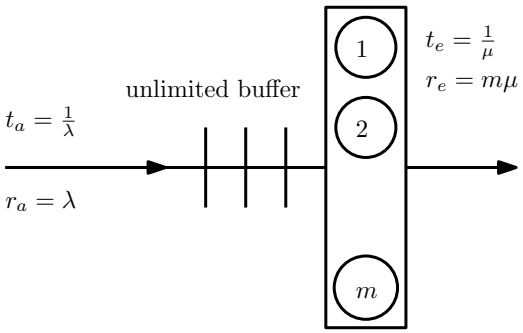


Figure 4.30: Exponential multi machine station

So we consider a workstation with unlimited buffer space and m parallel identical machines. Jobs arrive one at a time. The inter-arrival times are Exponential with rate $r_a = \lambda$. Processing is in order of arrival and jobs are processed one at a time. The process times are Exponential with rate μ. Hence the capacity of the station is $r_e = m\mu$ and the utilization of each machine is

$$u = \frac{r_a}{r_e} = \frac{\lambda}{\mu},$$

which is assumed to be less than 1.

Figure 4.31: Flow diagram for $M/M/m$ system

Let p_n be the probability (or long-run fraction of time) of finding n jobs in the system. As we have seen before, these probabilities can be determined through balance equations stating that, in equilibrium,

the number of transitions per unit time from state n to $n-1$ is equal to the number from $n-1$ to n, yielding (see Figure 4.27)

$$p_n n\mu = p_{n-1}\lambda \quad n \leq m,$$

and

$$p_n m\mu = p_{n-1}\lambda \quad n > m.$$

Hence, for $n \leq m$ we obtain

$$p_n = p_{n-1}\frac{\lambda}{n\mu} = p_{n-2}\frac{1}{n(n-1)}\left(\frac{\lambda}{\mu}\right)^2 = \cdots = p_0\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n$$

or with $u = \frac{\lambda}{m\mu}$,

$$p_n = p_0\frac{(mu)^n}{n!}, \quad n \leq m. \tag{4.13}$$

For $n > m$ we get

$$p_n = p_{n-1}\frac{\lambda}{m\mu} = p_{n-2}\left(\frac{\lambda}{m\mu}\right)^2 = \cdots = p_m\left(\frac{\lambda}{m\mu}\right)^{n-m}$$

or

$$p_n = p_m u^{n-m} = p_0\frac{(mu)^m}{m!}u^{n-m}, \quad n > m. \tag{4.14}$$

Finally, the probability p_0 follows from the requirement that the probabilities p_n have to add up to 1,

$$\sum_{n=0}^{\infty} p_n = 1$$

so

$$\frac{1}{p_0} = \sum_{n=0}^{m-1}\frac{(mu)^n}{n!} + \frac{(mu)^m}{m!}\frac{1}{1-u}.$$

From (4.13) we obtain for $n = m$ that

$$p_m = p_0\frac{(mu)^m}{m!} = \frac{\frac{(mu)^m}{m!}}{\sum_{n=0}^{m-1}\frac{(mu)^n}{n!} + \frac{(mu)^m}{m!}\frac{1}{1-u}}. \tag{4.15}$$

An important quantity is the probability Q that all machines are busy. In case of a single machine ($m = 1$) we immediately have $Q = u$, but for multiple machines this is more involved. From (4.14) and (4.15) we get

$$Q = \sum_{n=m}^{\infty} p_n = p_m\sum_{n=m}^{\infty} u^{n-m} = \frac{p_m}{1-u} = \frac{\frac{(mu)^m}{m!}}{(1-u)\sum_{n=0}^{m-1}\frac{(mu)^n}{n!} + \frac{(mu)^m}{m!}}. \tag{4.16}$$

For the calculation of Q there is a simple and reasonable *approximation* available, namely

$$Q = u^{\sqrt{2(m+1)}-1}. \tag{4.17}$$

Based on the expressions for the probabilities p_n we can now determine the average WIP in the buffer,

$$w_B = \sum_{n=m}^{\infty}(n-m)p_n = p_m\frac{u}{(1-u)^2} = \frac{Qu}{1-u}$$

and by Little's law,

$$\varphi_B = \frac{w_B}{\lambda} = \frac{Q}{1-u}\frac{t_e}{m},$$

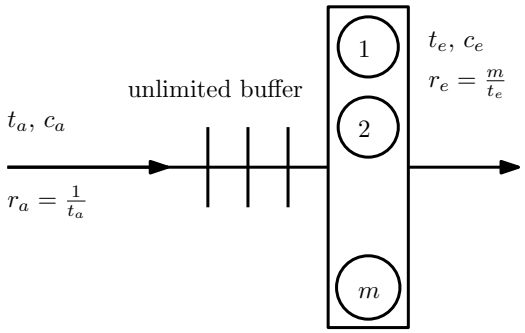


Figure 4.32: General multi machine station

from which the mean flow time and average total WIP directly follow,

$$\begin{aligned} \varphi &= \varphi_B + t_e = \frac{Q}{1-u}\,\frac{t_e}{m} + t_e, \\ w &= \lambda\varphi = \frac{Qu}{1-u} + mu. \end{aligned}$$

This concludes the treatment of the Exponential model and we now proceed by considering General inter-arrival times and process times, i.e., they can have any distribution, see Figure 4.32. So the inter-arrival times are General with mean t_a and coefficient of variation c_a, the process times are General with mean t_e and coefficient of variation c_e, and the machine utilization is given by

$$u = \frac{r_a}{r_e} = \frac{r_a t_e}{m} < 1.$$

We do no longer have an exact expression for the mean waiting time in the buffer, but we adopt, as before, the following *approximation*,

$$\varphi_B = \gamma \times \frac{Q}{1-u} \times \frac{t_e}{m}, \tag{4.18}$$

where

$$\gamma = \frac{1}{2}(c_a^2 + c_e^2).$$

The above expression (4.12) separates φ_B into three terms $V \times U \times T$: dimensionless Variability term γ, dimensionless Utilization term $\frac{Q}{1-u}$, and the process Time term $\frac{t_e}{m}$. In (4.18) we can use for Q either (4.16) (which is exact for the Exponential model) or the simple approximation (4.17).

Example 4.9. We consider a workstation with $m = 2$ machines. Inter-arrival times are Exponential with rate $r_a = 9$ (and $c_a^2 = 1$) and the process times are Uniform on $(0, \frac{2}{5})$, so $t_e = \frac{1}{5}$ and $c_e^2 = \frac{1}{3}$. Then $u = \frac{r_a t_e}{m} = \frac{9}{10}$.. Hence $\gamma = \frac{1}{2}(1 + \frac{1}{3}) = \frac{2}{3}$, $Q \approx 0.9^{\sqrt{6}-1} = 0.86$ (the exact value of $Q = \frac{81}{95} = 0.85$), so the mean waiting time can be approximated by

$$\varphi_B = \gamma\,\frac{Q}{1-u}\,\frac{t_e}{2} = \frac{2}{3} \cdot 0.86 = 0.58.$$

The quality of this approximation can be evaluated by the following χ3.0 model (see Appendix E).

```
1  model GBMmE():
2      int m = 2;
3      chan job a, b, c;
4
5      run G(a, exponential(0.111)),
6          B(a, b),
7          unwind j in range(m):
8              M(b, c, uniform(0.0, 0.4))
9          end
10         E(c, 1000000)
11 end
```

Based on 10^6 departures, the simulation estimate is 0.58. Hence, in this example, approximation (4.12) is quite accurate.

From the approximation (4.18)) for the mean waiting time in the buffer we get, by Little's law, the following approximation for the average WIP in the buffer,

$$w_B(G/G/m) = \delta\varphi_B = \gamma \frac{Qu}{1-u},$$

where $\delta = r_a$ is the throughput of the workstation. For the mean flow time and the average total WIP we obtain the approximations

$$\begin{aligned} \varphi &= \varphi_B + t_e = \gamma \frac{Q}{1-u} \frac{t_e}{m} + t_e, \\ w &= \delta\varphi = \gamma \frac{Qu}{1-u} + mu. \end{aligned}$$

Note that for $m = 1$ these approximations coincide with the approximations for the single machine model, presented in the previous section.

4.13 Serial production lines

We now have all ingredients to estimate the performance of serial production lines consisting of n workstations, see Figure 4.33 where $m(i)$, $t_a(i)$, $c_a(i)$, $t_d(i)$, $c_d(i)$, $t_e(i)$ and $c_e(i)$ are the descriptors of workstation i, $i = 1, \ldots, n$.

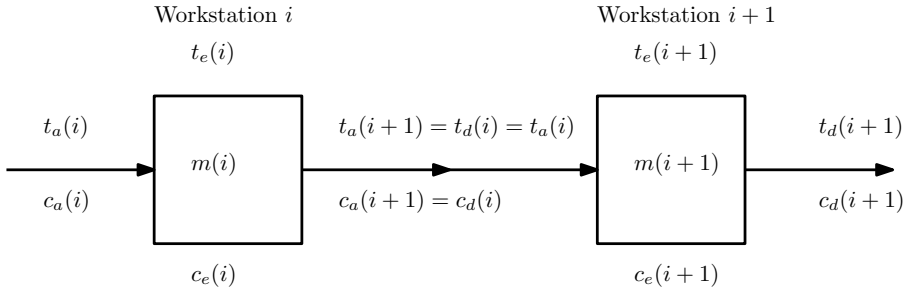


Figure 4.33: Serial production line consisting of n workstations

For stability we assume that each workstation can handle all work that is offered to that stations, so

$$u(i) = \frac{t_e(i)}{m(i)t_a(i)} < 1, \quad i = 1, \ldots, n.$$

Conservation of flow implies that the flow out of station i is equal to the flow into station i,

$$t_d(i) = t_a(i)$$

and also that the flow into station $i+1$ is equal to the flow out of station i,

$$t_a(i+1) = t_d(i).$$

The above relations link the flow rates in the serial production lines. The following relations link the variability of the flows through the production line and, in particular, the relations describe how variability *propagates through the production line*. For the variability of the output of station i we have as approximation (see (4.8))

$$c_d^2(i) = 1 + (1-u^2(i))(c_a^2(i) - 1) + \frac{u^2(i)}{\sqrt{m(i)}}(c_e^2(i) - 1)$$

and since departures from workstation i are arrivals to workstation $i+1$,

$$c_a(i+1) = c_d(i).$$

The mean waiting time in each workstation i can be estimated by (see (4.18))

$$\varphi_B(i) = \gamma(i)\,\frac{Q(i)}{1-u(i)}\,\frac{t_e(i)}{m(i)}.$$

Example 4.10. We consider a production line with two machines, see Figure 4.34.

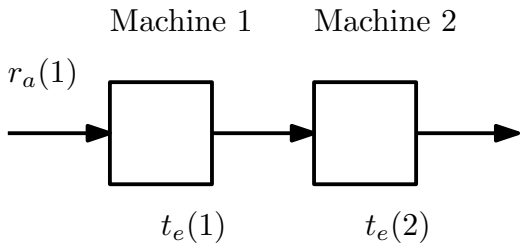


Figure 4.34: Production line with two machines

Machine 1 has Poisson inflow with rate $r_a(1) = 2$ (and $c_a(1) = 1$). The process times on Machine 1 are Constant with $t_e(1) = \frac{1}{3}$ and on Machine 2 the processs times are Uniform with $t_e(2) = \frac{2}{5}$. So $c_e^2(1) = 0$ and $c_e^2(2) = \frac{1}{3}$. What is the mean total flow time of jobs? For Machine 1 we have $u(1) = r_a(1)t_e(1) = \frac{2}{3}$, $\gamma(1) = \frac{1}{2}(1+0) = \frac{1}{2}$, and thus

$$\varphi(1) = \gamma(1)\,\frac{u(1)}{1-u(1)}\,t_e(1) + t_e(1) = \frac{2}{3}.$$

For Machine 2 we get $r_a(2) = r_a(1) = 2$ and

$$c_a^2(2) = c_d^2(1) = u^2(1)c_e^2(1) + (1-u^2(1))c_a^2(1) = \frac{5}{9}.$$

Furthermore, $c_e^2(2) = \frac{1}{3}$, so $\gamma(2) = \frac{1}{2}(\frac{5}{9} + \frac{1}{3}) = \frac{4}{9}$, $u(2) = r_a(2)t_e(2) = \frac{4}{5}$ and

$$\varphi(2) = \gamma(2)\,\frac{u(2)}{1-u(2)}\,t_e(2) + t_e(2) = 1\frac{1}{9}.$$

So the mean total flow time is

$$\varphi = \varphi(1) + \varphi(2) = 1\frac{7}{9} = 1.77.$$

This estimate for the mean total flow time is based on approximations. How good is this estimate? This can be investigated by the following χ3.0 model (see Appendix F).

```
1  model real Mline():
2      list(3) chan lot a, b;
3
4      run G(a[0], exponential(0.5)),
5          B(a[0], b[0]), M(b[0], a[1], constant(0.33)),
6          B(a[1], b[1]), M(b[1], a[2], uniform(0.0,0.8)),
7          E(a[2], 100000)
8  end
```

The χ3.0 model estimates $\varphi = 1.86$, so we can conclude that the analytical approximation is reasonably accurate.

Example 4.11. Let us consider the previous example and reverse the two machines: So first Machine 2 and then Machine 1, see Figure 4.35. Does mean total flow time increase or decrease by reversing the two machines?

The mean total flow time for the reversed system is equal to $\varphi = 1.99$, which is greater than for original configuration. Why? The reason is that variability propagates through the line! The process variability $c_e^2(2)$ on Machine 2 is higher than $c_e^2(1)$ on Machine 1, so it is better to locate Machine 2 toward the end of the production line.

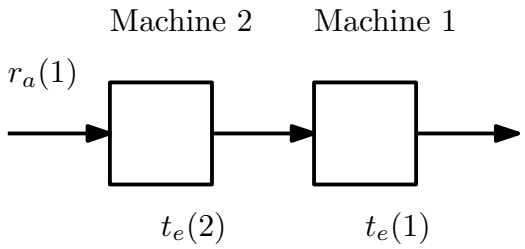


Figure 4.35: Production line with Machine 2 first

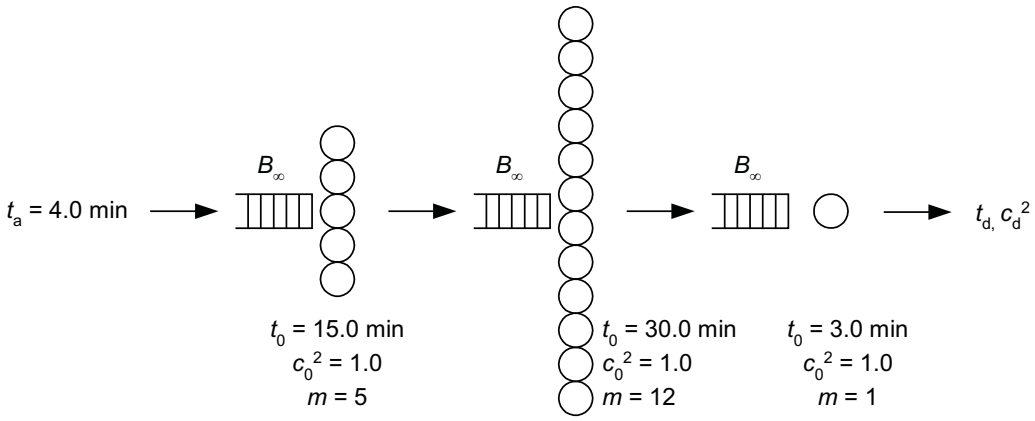


Figure 4.36: Manufacturing line with multiple-machine workstations

Exercise 52. (*Exercise 3.7.1 [1]*) We have a manufacturing line with three workstations, see Figure 4.36. From measurements done, we have the following data available for each workstation:

- Workstation 1 consists of an infinite buffer and 5 identical machines with t_0 = 15.0 min/lot and $c_0^2 = 1.0$.
- Workstation 2 consists of an infinite buffer and 12 identical machines with t_0 = 30.0 min/lot and $c_0^2 = 1.0$.
- Workstation 3 consists of an infinite buffer and one machine with $t_0 = 3.0$ min/lot and $c_0^2 = 1.0$.

We assume that lots arrive at workstation 1 with an inter-arrival time that is distributed according to an exponential distribution with mean 4.0 minutes.

1. Calculate the utilisation for workstation 1,2, and 3.
2. Calculate the mean waiting time in the buffer for workstation 1.
3. Calculate the mean inter-arrival time $t_{a,2}$ and squared coefficient of variation $c_{a,2}^2$ for lots arriving at workstation 2.
4. Calculate the mean waiting time in the buffer for workstation 2.
5. Calculate the mean waiting time in the buffer for workstation 3.
6. Calculate the mean number of products in buffer 1,2, and 3.
7. Explain the low WIP-level in buffer 2.
8. Calculate the mean flow time of a lot for the entire manufacturing line.
9. What is the mean number of products in the line (the WIP-level)?

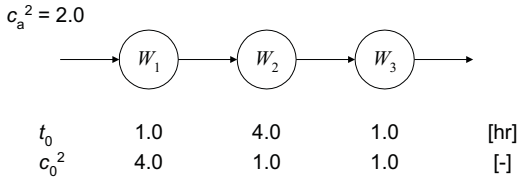


Figure 4.37: Three-workstation manufacturing line

Exercise 53. (*Exercise 3.7.2 [1]***)** Consider a manufacturing line consisting of three workstations. Each workstation consists of an infinite buffer and a single machine. In addition we have the following data:

- Lots arrive according to a distribution with mean t_a and squared coefficient of variation $c_a^2 = 2.0$.
- Workstations 1,2, and 3 have a mean process time of $t_{0,1} = 1.0$ hr, $t_{0,2} = 4.0$ hr, and $t_{0,3} = 1.0$ hr respectively and a squared coefficient of variation $c_{0,1}^2 = 4.0$, $c_{0,2}^2 = 1.0$, and $c_{0,3}^2 = 1.0$.

Figure 4.37 shows a graphical representation of the system.

1. Calculate the maximum throughput $\delta_{\max}$.
2. What is the mean flow time of a product if we require a throughput $\delta = \delta_{\max}$?

We wish to obtain a throughput $\delta = 0.9\delta_{\max}$.

1. Calculate the utilisation of machine 1,2, and 3.
2. Calculate the mean flow time of the line.
3. Calculate the mean number of products in buffer 1,2, and 3.
4. If you were to improve the flow time and/or throughput of this system, with which machine would you start?

The management decides to invest in workstation 2. We have two alternatives. In alternative one we invest in speed. We speed up the machine in workstation 2 so that $t_{0,2} = 1.0$ hr, however this induces an increase in variability: the squared coefficient of variation $c_{0,2}^2$ becomes 4.0. In alternative two we simply add 3 similar machines to the line: workstation 2 consists of 4 identical machines with $t_{0,2} = 4.0$ hr and $c_{0,2}^2 = 1.0$. Both alternatives are graphically represented in Figure 4.38.

1. What is the maximum throughput for both alternatives?
2. Calculate the mean flow time for alternative 1 for $\delta = 0.9\delta_{\max}$.
3. Calculate the mean flow time for alternative 2 for $\delta = 0.9\delta_{\max}$.
4. Calculate the WIP-level for both alternatives.
5. Which alternative do you prefer?
6. At first sight, the alternatives yield only a small decrease (or even an increase) in flow time, despite the investments done. Are these investments futile?

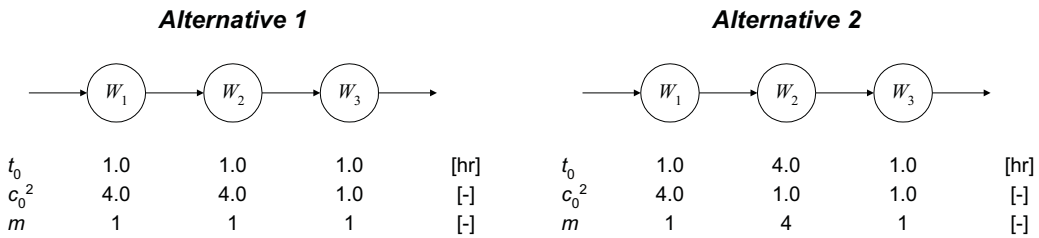


Figure 4.38: Two alternatives

4.14 Batching

In manufacturing systems we can distinguish between two kinds of batching.

- **Process batch:** Parts are processed together on a workstation. There are two types of process batches:
  - **Simultaneous:**
    Parts are produced simultaneously in a (true) batch workstation (e.g., in a furnace).
  - **Sequential:**
    Parts from a common part family are produced sequentially before the workstation is changed-over to another part family.
- **Transfer batch:** Parts are moved together from one station to another. Note that:
  - The smaller the transfer batch, the less time parts have to wait to form the batch.
  - The smaller the transfer batch, the more material handling is needed.

Below we investigate through examples the effect of batching on the manufacturing system performance.

Example 4.12. (Simultaneous batching interactions)

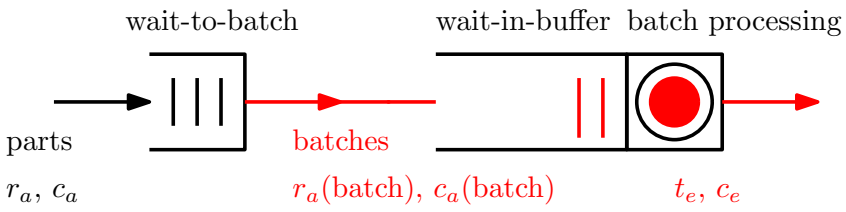


Figure 4.39: Simulataneous batching machine

Parts arrive one by one at a machine with rate r_a, see Figure 4.39. The coefficient of variation of the inter-arrival times is c_a. Parts are processed by the machine in batches of size k. The mean batch process time on the machine is t_e, and c_e is the coefficient of variation. What is mean flow time of a part? The part flow time can be decomposed in the *wait-to-batch time* plus the *batch flow time*,

$$\varphi(\text{part}) = \varphi(\text{w2b}) + \varphi(\text{batch}).$$

The wait-to-batch time is the waiting time of a part for the process batch to form. So the first part has to wait for the remaining $k-1$ parts to arrive, the second one for the remaining $k-2$ parts, and so on (see Figure 4.40). The wait-to-batch time for the last, kth part, is equal to 0. Hence, the average

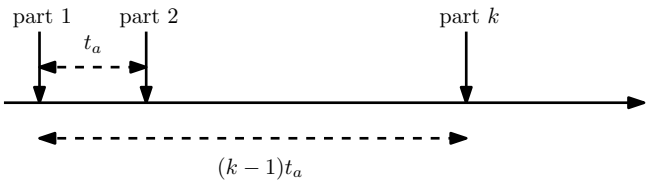


Figure 4.40: Wait-to-batch time

wait-to-batch time is equal to

$$\begin{aligned}\varphi(\text{w2b}) &= \frac{1}{k}\left((k-1)t_a + (k-2)t_a + \cdots + t_a + 0\right)\\ &= \frac{t_a}{k}\, k\frac{k-1+0}{2}\\ &= \frac{(k-1)t_a}{2}\\ &= \frac{k-1}{2r_a}.\end{aligned}$$

The batch arrival rate is $r_a(\text{batch}) = \frac{r_a}{k}$ and

$$c_a^2(\text{batch}) = \frac{\sigma_a^2(\text{batch})}{t_a^2(\text{batch})} = \frac{k\sigma_a^2}{(kt_a)^2} = \frac{c_a^2}{k}.$$

So

$$u = r_a(\text{batch})t_e = \frac{r_a t_e}{k}$$

and

$$\gamma(\text{batch}) = \frac{1}{2}\left(c_a^2(\text{batch}) + c_e^2\right) = \frac{1}{2}\left(\frac{c_a^2}{k} + c_e^2\right).$$

Hence, for the mean batch flow time, we obtain

$$\begin{aligned}\varphi(\text{batch}) &= \gamma(\text{batch})\frac{u}{1-u}t_e + t_e\\ &= \frac{1}{2}\left(\frac{c_a^2}{k} + c_e^2\right)\frac{r_a t_e}{k - r_a t_e}t_e + t_e\end{aligned}$$

and putting the above expressions together, we finally arrive at

$$\varphi(\text{part}) = \varphi(\text{w2b}) + \varphi(\text{batch}) = \frac{k-1}{2r_a} + \frac{1}{2}\left(\frac{c_a^2}{k} + c_e^2\right)\frac{r_a t_e}{k - r_a t_e}t_e + t_e.$$

Example 4.13. (Sequential batching interactions)
We consider the workstation in Example 4.15, but now assume that parts in a batch are produced sequentially, i.e., they belong to the same family and are processed one at a time. But before processing a batch, the machine needs a change-over or setup time. After the setup time, the machine processes the parts one by one, and the batch is sent downstream as soon as all parts have been processed. So the only aspect that is different from Example 4.15 is the batch process time. Let t_s denote the mean setup time, and σ_s^2 is the variance of the setup time. The mean process time of a single part is t_p and the variance is σ_p^2. The process batch size is k. For the mean and variance of the batch process time we obtain

$$t_e = t_s + kt_p, \quad \sigma_e^2 = \sigma_s^2 + k\sigma_p^2.$$

So the squared coefficient of variation is

$$\begin{aligned}c_e^2 &= \frac{\sigma_e^2}{t_e^2}\\ &= \frac{\sigma_s^2 + k\sigma_p^2}{(t_s + kt_p)^2}\end{aligned}$$

and the machine utilization,

$$\begin{aligned} u &= r_a(\text{batch})t_e \\ &= \frac{r_a}{k}t_e \\ &= r_a\left(\frac{t_s}{k}+t_p\right). \end{aligned}$$

Note that the requirement $u<1$ implies that there is *minimal feasible* batch size: $k > \frac{r_a t_s}{1-r_a t_p}$. Furthermore, from the expression for u, we can conclude that the batch size k affects the machine utilization: A larger batch size k leads to a lower utilization. The mean part flow time is the average wait-to-batch time plus the mean batch flow time, so

$$\begin{aligned} \varphi(\text{part}) &= \varphi(\text{w2b})+\varphi(\text{batch}) \\ &= \frac{k-1}{2r_a}+\frac{1}{2}\left(\frac{c_a^2}{k}+\frac{\sigma_s^2+k\sigma_p^2}{(t_s+kt_p)^2}\right)\frac{r_a(\frac{t_s}{k}+t_p)}{1-r_a(\frac{t_s}{k}+t_p)}(t_s+kt_p)+t_s+kt_p. \end{aligned}$$

Apparently, there is a *batch size trade-off*: A larger batch size k leads to a lower machine utilization, but also to a higher wait-to-batch and wait-in-batch time. This trade-off is illustrated in Figure 4.41 for the parameter setting $r_a=0.4$, $c_a=1$, $t_s=5$, $\sigma_s^2=6\frac{1}{4}$, $t_p=1$ and $\sigma_p^2=\frac{1}{4}$. In this example the optimal batch size is $k=5$.

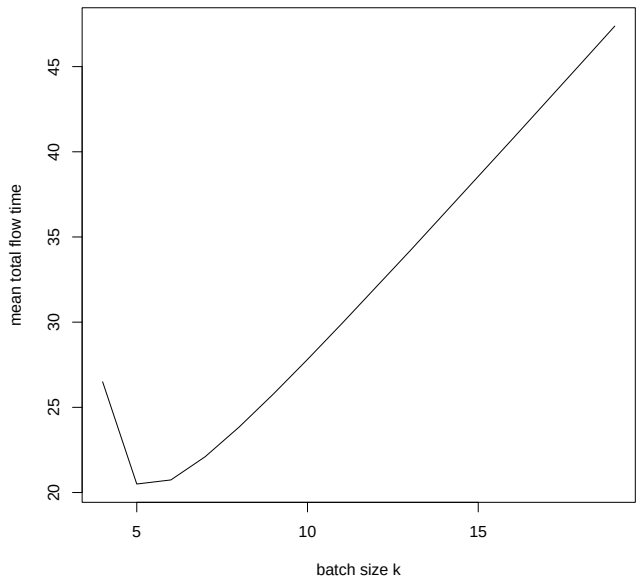


Figure 4.41: Mean part flow time as a function of the batch size k for the parameter setting $r_a=0.4$, $c_a=1$, $t_s=5$, $\sigma_s^2=6\frac{1}{4}$, $t_p=1$ and $\sigma_p^2=\frac{1}{4}$

Example 4.14. (Sequential batching interactions and job splitting)
In the previous example we assumed that the batch is sent downstream as soon as all parts have been processed. In this example we consider a different situation: Individual parts are sent downstream as soon as being processed. In other words, the batch is split and parts proceed individually

after processing. Then the mean effective process time of a part is

$$\begin{aligned} t_e(\text{part}) &= t_s + \frac{1}{k}(t_p + 2t_p + \cdots + kt_p) \\ &= t_s + \frac{t_p}{k}(1 + 2 + \cdots + k) \\ &= t_s + \frac{t_p}{k}\frac{k(1+k)}{2} \\ &= t_s + \frac{(k+1)t_p}{2}. \end{aligned}$$

Hence, the mean flow time of part with job splitting is

$$\varphi(\text{part}) = \frac{k-1}{2r_a} + \frac{1}{2}\left(\frac{c_a^2}{k} + \frac{\sigma_s^2 + k\sigma_p^2}{(t_s + kt_p)^2}\right)\frac{r_a(\frac{t_s}{k} + t_p)}{1 - r_a(\frac{t_s}{k} + t_p)}(t_s + kt_p) + t_s + \frac{(k+1)t_p}{2}.$$

Example 4.15. (Transfer batching interactions)

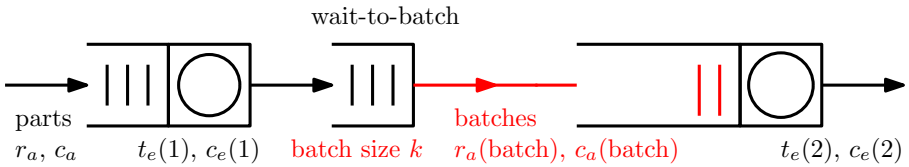


Figure 4.42: Transfer batching in a production line with two machines

We consider a production line with two machines. Single parts arrive with rate r_a at Machine 1. After processing on Machine 1, parts are transferred in batches of size k to Machine 2, the transfer time is 0 (negligible). Machine i ($i = 1, 2$) processes parts one by one, with mean $t_e(i)$ and coefficient of variation $c_e(i)$. How does mean flow time of part depend on batch size k? The mean flow time at Machine 1 is equal to

$$\varphi(1) = \frac{1}{2}(c_a^2(1) + c_e^2(1))\frac{u(1)}{1 - u(1)}\ t_e(1) + t_e(1).$$

After processing on Machine 1, parts have to wait for the transfer batch to form. To calculate the average wait-to-batch time after machine 1, note that the output rate of Machine 1 is r_a, so the mean inter-departure time of parts from Machine 1 is $\frac{1}{r_a}$). Hence,

$$\varphi(\text{w2b}) = \frac{k-1}{2r_a}.$$

Then, when the batch arrives at machine 2, it has to wait in the buffer for processing. The mean waiting time of a batch in the buffer of machine 2 is

$$\begin{aligned} \varphi_B(2) &= \frac{1}{2}\left(\frac{c_d^2(1)}{k} + \frac{c_e^2(2)}{k}\right)\frac{u(2)}{1 - u(2)}kt_e(2) \\ &= \frac{1}{2}\left(c_d^2(1) + c_e^2(2)\right)\frac{u(2)}{1 - u(2)}t_e(2) \end{aligned}$$

where $u(2) = \frac{r_a}{k}kt_e(2) = r_a t_e(2)$ and

$$c_d^2(1) = u^2(1)c_e^2(1) + (1 - u^2(1))c_a^2(1).$$

After waiting in the buffer, the parts in the batch is processed one by one by Machine 2. The average wait-to-process time of a part at machine 2 is equal to

$$\begin{aligned} \varphi(\text{w2p}) &= \frac{1}{k}(t_e(2) + 2t_e(2) + \cdots + kt_e(2)) \\ &= \frac{(k+1)t_e(2)}{2}. \end{aligned}$$

Summarizing, the mean flow time of a part is

$$\begin{aligned}\varphi &= \varphi(1) + \varphi(\text{w2b}) + \varphi_B(2) + \varphi(\text{w2p}) \\ &= \frac{1}{2}(c_a^2(1) + c_e^2(1))\frac{u(1)}{1-u(1)}\, t_e(1) + t_e(1) + \frac{k-1}{2r_a} + \frac{1}{2}\left(c_d^2(1) + c_e^2(2)\right)\frac{u(2)}{1-u(2)}t_e(2) + \frac{(k+1)t_e(2)}{2}.\end{aligned}$$

From this expression we observe that the mean total flow time increases proportionally with the transfer batch size k, which has nothing to do with process or arrival variability. In other words, this is bad control! Also note that the impact of transfer batching on the mean flow time is largest when the arrival rate r_a is low.

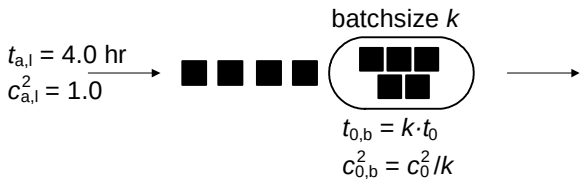


Figure 4.43: Batch machine processing batches of fixed size k

Exercise 54. (*Exercise 3.7.3 [1]*) We have a batch machine that processes batches of fixed size k, see Figure 4.43. Lots arrive with mean inter-arrival time $t_{a,l} = 4.0$ hours and squared coefficient of variation $c^2_{a,l} = 1.0$. The process time for a batch is proportional with the batch size: $t_{0,b} = k \cdot t_0$, the squared coefficient of variation is inversely proportional with the batch size: $c^2_{0,b} = c^2_0/k$. For this machine we have $t_0 = 3.0$ hours and $c^2_0 = 0.50$.

1. Express the mean total flow time for this batch machine in batch size k.
2. For what value of k is the flow time minimal? (Do not forget to assure a stable system.)

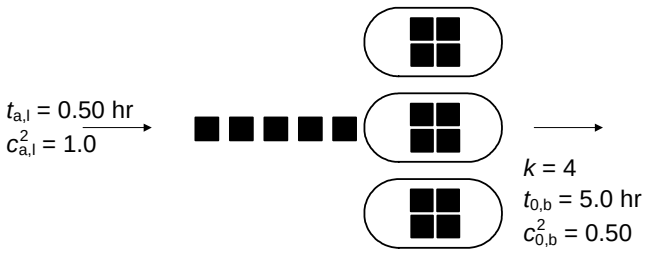


Figure 4.44: Three batch machines in parallel

Exercise 55. (*Exercise 3.7.4 [1]*) We have three identical batch machines in parallel, see Figure 4.44. The batch machines have a single buffer in which batches are formed. The machines process batches of a fixed size $k = 4$ lots. Lots arrive at the buffer with $t_{a,l} = 0.50$ hours and $c^2_{a,l} = 1.0$. The batch process time is characterized by $t_{0,b} = 5.0$ hours and $c^2_{0,b} = 0.5$.

1. Calculate the utilisation of the batch machines.
2. Calculate the wait-to-batch-time in the buffer.
3. Calculate the queueing time for a batch in the buffer.
4. Calculate the mean total flow time.

Bibliography

[1] I.J.B.F. ADAN, A.T. HOFKAMP, J.E. ROODA AND J. VERVOORT, *Analysis of Manufacturing Systems*, 2012.

[2] W.J. HOPP, M.L. SPEARMAN, 2012. *Factory Physics*, 3rd ed., McGraw-Hill, 2008.

[3] D. MORIN, *Probability - For the Enthusiastic Beginner*, 2016.

[4] R. SURI, *Quick Response Manufacturing*, Taylor&Francis Inc, 1998.

[5] H.C. TIJMS, *Understanding Probability*, 3rd ed., Cambridge, 2012.

A

KIVA model

```
type pod = int;

proc G(chan! pod a; int N):
    pod x;

    for i in range(N):
        a!x;
    end
end

proc S(chan? pod a; chan! pod b; real la):
    pod x;

    while true:
        a?x;
        delay sample exponential(1.0/la);
        b!x;
    end
end

proc B(chan? pod a; chan! pod b):
    list pod xs;
    pod x;

    while true:
        select
            a?x:
                xs = xs + [x]
        alt
            size(xs) > 0, b!xs[0]:
                xs = xs[1:]
        end
    end
end

proc P(chan? pod a; chan! pod b; real mu; int n):
    pod x;

    for i in range(n):
        a?x;
        delay sample exponential(1.0/mu);
        b!x;
    end
    writeln("TH = %g", n / time);
end

model KIVA():
    int N = 1;
    real la = 4.0, mu = 20.0;
    chan pod a, b, c;

    run G(a, N),
        unwind j in range(N):
            S(a, b, la)
        end,
```

```
56          B(b, c), P(c, a, mu, 10000)
57  end
```

B

Zero-buffer model

```
 1  type job = int;
 2
 3  proc G(chan! job a; dist real u):
 4      job x;
 5
 6      while true:
 7          a!x;
 8          delay sample u;
 9          x = x + 1;
10      end
11  end
12
13  proc M(chan? job a; chan! job b; dist real u):
14      job x;
15
16      while true:
17          a?x;
18          delay sample u;
19          b!x;
20      end
21  end
22
23  proc real E(chan? job a; int n):
24      job x;
25
26      while x < n:
27          a?x;
28      end;
29      exit x / time
30  end
31
32  model real GME():
33      real ta = 1.0, te = 1.0;
34      int n = 1000;
35      chan job a, b;
36
37      run G(a, constant(ta)), M(a, b, constant(te)), E(b, n)
38  end
```

C

Finite-buffer model

```
type job = int;

proc G(chan! job a; dist real u):
    job x;

    while true:
        a!x;
        delay sample u;
        x = x + 1;
    end
end

proc B(chan? job a; chan! job b; int N):
    list job xs;
    job x;

    while true:
        select
            size(xs) < N, a?x:
                xs = xs + [x]
        alt
            size(xs) > 0, b!xs[0]:
                xs = xs[1:]
        end
    end
end

proc M(chan? job a; chan! job b; dist real u):
    job x;

    while true:
        a?x;
        delay sample u;
        b!x;
    end
end

proc real E(chan? job a; int n):
    job x;

    while x < n:
        a?x;
    end;
    exit x / time
end

model real GBME():
    real ta = 1.0, te = 1.0;
    int n = 1000, N = 10;
    chan job a, b, c;

    run G(a, exponential(ta)), B(a, b, N), M(b, c, exponential(te)), E(c, n)
end
```

D

Single machine model

```
type job = real;

proc G(chan! job a; dist real u):

    while true:
        a!time;
        delay sample u;
    end
end

proc B(chan? job a; chan! job b):
    list job xs;
    job x;

    while true:
        select
            a?x:
                xs = xs + [x]
        alt
            size(xs) > 0, b!xs[0]:
                xs = xs[1:]
        end
    end
end

proc M(chan? job a; chan! job b; dist real u):
    job x;

    while true:
        a?x;
        b!x;
        delay sample u;
    end
end

proc E(chan? job a; int n):
    real sumw;
    job x;

    for i in range(n):
        a?x;
        sumw = sumw + (time - x);
    end;
    writeln("Mean waiting time spent in buffer = %g", sumw / n);
end

model GBME():
    chan job a, b, c;

    run G(a, uniform(0.0, 2.0)),
        B(a, b), M(b, c, uniform(0.0, 1.0)),
        E(c, 100000)
end
```

E

Multi machine model

```
type job = real;

proc G(chan! job a; dist real u):

    while true:
        a!time;
        delay sample u;
    end
end

proc B(chan? job a; chan! job b):
    list job xs;
    job x;

    while true:
        select
            a?x:
                xs = xs + [x]
        alt
            size(xs) > 0, b!xs[0]:
                xs = xs[1:]
        end
    end
end

proc M(chan? job a; chan! job b; dist real u):
    job x;

    while true:
        a?x;
        b!x;
        delay sample u;
    end
end

proc E(chan? job a; int n):
    real sumw;
    job x;

    for i in range(n):
        a?x;
        sumw = sumw + (time - x);
    end;
    writeln("Mean waiting time spent in buffer = %g", sumw / n);
end

model GBMmE():
    int m = 2;
    chan job a, b, c;

    run G(a, exponential(0.111)),
        B(a, b),
        unwind j in range(m):
        M(b, c, uniform(0.0, 0.4))
        end
```

```
56          E(c, 1000000)
57  end
```

F

Serial production line

```
type job = real;

xper X():
    int m = 10;
    real phi;
    real sum1, sum2, smean, svar;

    for i in range(m):
        phi = Mline();
        sum1 = sum1 + phi;
        sum2 = sum2 + phi * phi;
        writeln("Mean flow time in run %d is %g", i+1, phi)
    end;

    smean = sum1 / m;
    svar = sum2 / m - smean * smean;
    writeln("Mean flow time estimate is %g +- %g", smean, 1.96 * sqrt(svar / m));
end;

proc G(chan! job a; dist real u):

    while true:
        a!time;
        delay sample u;
    end
end

proc B(chan? job a; chan! job b):
    list job xs;
    job x;

    while true:
        select
            a?x:
                xs = xs + [x]
        alt
            size(xs) > 0, b!xs[0]:
                xs = xs[1:]
        end
    end
end

proc M(chan? job a; chan! job b; dist real u):
    job x;

    while true:
        a?x;
        delay sample u;
        b!x;
    end
end

proc real E(chan? job a; int n):
    int i;
    real sum;
```

```
56      job x;
57
58      while i < n:
59          a?x;
60          sum = sum + (time - x);
61          i = i + 1;
62      end;
63      exit sum / n
64  end
65
66  model real Mline():
67      list(3) chan job a, b;
68
69      run G(a[0], exponential(0.5)),
70          B(a[0], b[0]), M(b[0], a[1], constant(0.33)),
71          B(a[1], b[1]), M(b[1], a[2], uniform(0.0,0.8)),
72          E(a[2], 100000)
73  end
```