Chapter 12
Random processes

This chapter is devoted to the idea of a random process and to models for random processes. The problems of estimating model parameters and of testing the adequacy of the fit of a hypothesized model to a set of data are addressed. As well as revisiting the familiar Bernoulli and Poisson processes, a model known as the Markov chain is introduced.

In this chapter, attention is turned to the study of random processes. This is a wide area of study, with many applications. We shall discuss again the familiar Bernoulli and Poisson processes, considering different situations where they provide good models for variation and other situations where they do not. Only one new model, the Markov chain, is introduced in any detail in this chapter (Sections 12.2 and 12.3). This is a refinement of the Bernoulli process. In Section 12.4 we revisit the Poisson process.

However, in order for you to achieve some appreciation of the wide application of the theory and practice of random processes, Section 12.1 is devoted to examples of different contexts. Here, questions are posed about random situations where some sort of random process model is an essential first step to achieving an answer. You are not expected to do any more in this section than read the examples and try to understand the important features of them. First, let us consider what is meant by the term 'random process'.

In Chapter 2, you rolled a die many times, recording a 1 if you obtained either a 3 or a 6 (a success) and 0 otherwise. You kept a tally of the running total of successes with each roll, and you will have ended up with a table similar to that shown in Table 2.3. Denoting by the random variable \( X_n \) the number of successes achieved in \( n \) rolls, you then calculated the proportion of successes \( X_n/n \) and plotted that proportion against \( n \), finishing with a diagram something like that shown in Figure 2.2. For ease, that diagram is repeated here in Figure 12.1.

The reason for beginning this chapter with a reminder of past activities is that it acts as a first illustration of a random process. The sequence of random variables

\[
\{X_n; n = 1, 2, 3, \ldots\},
\]

or the sequence plotted in Figure 12.1,

\[
\{X_n/n; n = 1, 2, 3, \ldots\},
\]

are both examples of random processes. You can think of a random process evolving in time: repeated observations are taken on some associated random
variable, and this list of measurements constitutes what is known as a realization (that is, a particular case) of the random process. Notice that the order of observations is preserved and is usually important. In this example, the realization shows early fluctuations that are wide; but after a while, the process settles down (cf. Figure 2.3 which showed the first 500 observations on the random variable $X_n/n$). It is often the case that the long-term behaviour of a random process $\{X_n\}$ is of particular concern.

Here is another example.

**Example 12.1 More traffic data**

The data in Table 12.1 were collected by Professor Toby Lewis as an illustration of a random process which might possibly be modelled as a Poisson process. They show the 40 time intervals (in seconds) between 41 vehicles travelling northwards along the M1 motorway in England past a fixed point near Junction 13 in Bedfordshire on Saturday 23 March 1985. Observation commenced shortly after 10.30 pm.

One way of viewing the data is shown in Figure 12.2. This shows the passing vehicles as 41 blobs on a time axis, assuming the first vehicle passed the observer at time 0. You have seen this type of diagram already.

A different representation is to denote by the random variable $X(t)$ the number of vehicles (after the one that initiated the sequence) to have passed the observer by time $t$. Thus, for instance,

$$X(10) = 0, \quad X(20) = 3, \quad X(30) = 4, \quad \ldots$$

A graph of $X(t)$ against $t$ is shown in Figure 12.3. It shows a realization of the random process

$$\{X(t); t \geq 0\}$$
and could be used in a test of the hypothesis that events (vehicles passing a fixed point) occur as a Poisson process in time.

Figure 12.3 A realization of the random process \( \{X(t); t \geq 0\} \)

Again, the notion of ordering is important. Here, the random process is an integer-valued continuous-time process. Notice that there is no suggestion in this case that the random variable \( X(t) \) settles down to some constant value.

In this chapter we shall consider three different models for random processes. You are already familiar with two of them: the Bernoulli and the Poisson processes. The third model is a refinement of the Bernoulli process, which is useful in some contexts when the Bernoulli process proves to be an inadequate model. We shall consider the problem of parameter estimation given data arising from realizations of random processes, and we shall see how the quality of the fit of a hypothesized model to data is tested.

These models are discussed in Sections 12.2 to 12.4, and variations of the models are briefly examined in Section 12.5. There are no exercises in Sections 12.1 or 12.5.

12.1 Examples of random processes

So far in the course we have seen examples of many kinds of data set. The first is the random sample. Examples of this include leaf lengths (Table 2.1), heights (Table 2.15), skull measurements (Table 3.3), waiting times (Tables 4.7 and 4.10), leech counts (Table 6.2) and wages (Table 6.9). In Chapter 10 we looked at regression data such as height against age (Figure 10.1), boiling point of water against atmospheric pressure (Tables 10.2 and 10.4) and paper strength against hardwood content (Table 10.8). In Chapter 11 we explored bivariate data such as systolic and diastolic blood pressure measurements (Table 11.1), snoring frequency and heart disease (Table 11.2), and height and weight (Table 11.4).
Elements of Statistics

In many cases we have done more than simply scan the data: we have hypothesized models for the data and tested hypotheses about model parameters. For instance, we know how to test that a regression slope is zero; or that there is zero correlation between two responses measured on the same subject.

We have also looked at data in the context of the Poisson process. For the earthquake data (Table 4.7) we decided (Figure S9.6) that the waiting times between successive serious earthquakes world-wide could reasonably be modelled by an exponential distribution. In this case, therefore, the Poisson process should provide an adequate model for the sequence of earthquakes, at least over the years that the data collection extended. For the coal mine disaster data in Table 7.4 there was a distinct curve to the associated exponential probability plot (Figure 9.6). Thus it would not be reasonable to model the sequence of disasters as a Poisson process. (Nevertheless, the probability plot suggests some sort of systematic variation in the waiting times between disasters, and therefore the random process could surely be satisfactorily modelled without too much difficulty.) For the data on eruptions of the Old Faithful geyser (Table 3.18) we saw in Figure 3.18 that the inter-event times were not exponential: there were two pronounced modes. So the sequence of eruptions could not sensibly be modelled as a Poisson process.

Here are more examples.

**Example 12.2 Times of system failures**

The data in Table 12.2 give the cumulative times of failure, measured in CPU seconds and starting at time 0, of a command and control software system.

<table>
<thead>
<tr>
<th>Table 12.2</th>
<th>Cumulative times of failure (CPU seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>33 146 227 342 351 353 444 556</td>
</tr>
<tr>
<td>571</td>
<td>709 759 836 860 968 1056 1726 1846</td>
</tr>
<tr>
<td>1872</td>
<td>1986 2311 2366 2608 2676 3098 3278 3288</td>
</tr>
<tr>
<td>4434</td>
<td>5034 5049 5085 5089 5089 5097 5324 5389</td>
</tr>
<tr>
<td>5565</td>
<td>5623 6080 6380 6477 6740 7192 7447 7644</td>
</tr>
<tr>
<td>7837</td>
<td>7843 7922 8738 10089 10237 10258 10491 10625</td>
</tr>
<tr>
<td>10982</td>
<td>11175 11411 11442 11811 12559 12559 12791 13121</td>
</tr>
<tr>
<td>13486</td>
<td>14708 15251 15261 15277 15806 16185 16229 16358</td>
</tr>
<tr>
<td>17168</td>
<td>17458 17758 18287 18568 18728 19556 20567 21012</td>
</tr>
<tr>
<td>21308</td>
<td>23063 24127 25910 26770 27753 28460 28493 29361</td>
</tr>
<tr>
<td>30085</td>
<td>32408 35338 36799 37642 37654 37915 39715 40580</td>
</tr>
<tr>
<td>42015</td>
<td>42045 42188 42296 42296 45406 46653 47596 48296</td>
</tr>
<tr>
<td>49171</td>
<td>49146 50145 52042 52489 52875 53321 53443 54433</td>
</tr>
<tr>
<td>55381</td>
<td>56463 56485 56560 57042 62551 62681 62661 63732</td>
</tr>
<tr>
<td>64103</td>
<td>64893 71043 74364 75409 76057 81542 82702 84566</td>
</tr>
<tr>
<td>88862</td>
<td></td>
</tr>
</tbody>
</table>

Here, the random process \( \{T_n; n = 1, 2, 3, \ldots\} \) gives the time of the \( n \)th failure. If the system failures occur at random, then the sequence of failures will occur as a Poisson process in time. In this case, the times between failures will constitute independent observations from an exponential distribution. An exponential probability plot of the time differences \((3, 30, 113, 81, \ldots, 4116)\) is shown in Figure 12.4.

You can see from the figure that an exponential model does not provide a good fit to the variation in the times between failures. In fact, a more
extended analysis of the data reveals that the failures are becoming less frequent with passing time. The system evidently becomes less prone to failure as it ages.

In Example 12.2 the system was monitored continuously so that the time of occurrence of each failure was recorded accurately. Sometimes it is not practicable to arrange continuous observation, and system monitoring is intermittent. An example of this is given in Example 12.3.

**Example 12.3 Counts of system failures**

The data in Table 12.3 relate to a DEC-20 computer which was in use at the Open University during the 1980s. They give the number of times that the computer broke down in each of 128 consecutive weeks of operation, starting in late 1983. For these purposes, a breakdown was defined to have occurred if the computer stopped running, or if it was turned off to deal with a fault. Routine maintenance was not counted as a breakdown. So Table 12.3 shows a realization of the random process \( \{X_n; n = 1, 2, \ldots, 128\} \), an integer-valued random process developing in discrete time where \( X_n \) denotes the number of failures during the \( n \)th week.

| \( X \) | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 2 | 3 | 4 | 5 | 6 | 10 | 11 | 13 | 14 | 15 | 16 | 22 | 25 | 31 |
| \( n \) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |

Again, if breakdowns occur at random in time, then these counts may be regarded as independent observations from a Poisson distribution with constant mean. In fact, this is not borne out by the data (a chi-squared test of goodness-of-fit against the Poisson model with estimated mean 4.02 gives a SP which is essentially zero). The question therefore is: how might the incidence of break-
downs be modelled instead? (You can see occasional peaks of activity—weeks where there were many failures—occurring within the data.)

In the next example, this idea is explored further. The data are reported in a paper about a statistical technique called **bump-hunting**, the identification of peaks of activity.

**Example 12.4 Scattering activity**

A total of 25,752 scattering events called scintillations was counted from a particle-scattering experiment over 172 bins each of width 10 MeV. There are peaks of activity called *bumps* and it is of interest to identify where these bumps occur.

<table>
<thead>
<tr>
<th>Bin number, n</th>
<th>Scattering reactions X_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>11 17 21 15 17 23 25 30 22 36 29 33</td>
</tr>
<tr>
<td>43</td>
<td>54 55 59 44 58 66 59 55 67 75 82 98</td>
</tr>
<tr>
<td>94</td>
<td>85 92 102 113 122 153 155 193 197 207 258 305</td>
</tr>
<tr>
<td>332</td>
<td>318 378 457 540 592 646 773 787 783 695 774 759</td>
</tr>
<tr>
<td>692</td>
<td>559 499 431 421 353 315 343 306 262 265 254</td>
</tr>
<tr>
<td>225</td>
<td>246 225 196 150 118 114 99 121 106 112 122 120</td>
</tr>
<tr>
<td>126</td>
<td>126 141 122 122 115 119 166 135 154 120 162 156</td>
</tr>
<tr>
<td>175</td>
<td>193 162 178 201 214 230 216 229 214 197 170 181</td>
</tr>
<tr>
<td>183</td>
<td>144 114 120 132 109 108 97 102 89 71 92 58</td>
</tr>
<tr>
<td>65</td>
<td>55 53 40 42 46 47 37 49 38 29 34 42</td>
</tr>
<tr>
<td>45</td>
<td>42 40 59 42 35 41 35 48 41 47 49 37</td>
</tr>
<tr>
<td>40</td>
<td>33 33 37 29 26 38 22 27 27 13 18 25</td>
</tr>
<tr>
<td>24</td>
<td>21 16 24 14 23 21 17 17 21 10 14 18</td>
</tr>
</tbody>
</table>

Just glancing at the numbers, the peaks of activity do not seem consistent with independent observations from a Poisson distribution. The data in Table 12.4 become more informative if portrayed graphically, and this is done in Figure 12.5. The number of scattering reactions $X_n$ is plotted against bin number $n$.

**Figure 12.5** Scattering reactions in consecutive bins of constant width
The figure suggests that there are two main peaks of activity. (There are more than 30 smaller local peaks which do not appear to be significant, and are merely evidence of random variation in the data.)

Another example where mean incidence is not constant with passing time is provided by the road casualty data in Example 12.5. It is common to model this sort of unforeseen accident as a Poisson process in time, but, as the data show, the rate is in fact highly time-dependent.

**Example 12.5 Counts of road casualties**

Table 12.5 gives the number of casualties in Great Britain due to road accidents on Fridays in 1986, for each hour of the day from midnight to midnight.

A plot of these data (frequency of casualties against the midpoint of the associated time interval) is given in Figure 12.6.

<table>
<thead>
<tr>
<th>Time of day</th>
<th>Casualties</th>
</tr>
</thead>
<tbody>
<tr>
<td>00–01</td>
<td>938</td>
</tr>
<tr>
<td>01–02</td>
<td>621</td>
</tr>
<tr>
<td>02–03</td>
<td>455</td>
</tr>
<tr>
<td>03–04</td>
<td>207</td>
</tr>
<tr>
<td>04–05</td>
<td>138</td>
</tr>
<tr>
<td>05–06</td>
<td>215</td>
</tr>
<tr>
<td>06–07</td>
<td>526</td>
</tr>
<tr>
<td>07–08</td>
<td>1933</td>
</tr>
<tr>
<td>08–09</td>
<td>3377</td>
</tr>
<tr>
<td>09–10</td>
<td>2045</td>
</tr>
<tr>
<td>10–11</td>
<td>2078</td>
</tr>
<tr>
<td>11–12</td>
<td>2351</td>
</tr>
<tr>
<td>12–13</td>
<td>3015</td>
</tr>
<tr>
<td>13–14</td>
<td>2966</td>
</tr>
<tr>
<td>14–15</td>
<td>2912</td>
</tr>
<tr>
<td>15–16</td>
<td>4305</td>
</tr>
<tr>
<td>16–17</td>
<td>4923</td>
</tr>
<tr>
<td>17–18</td>
<td>4427</td>
</tr>
<tr>
<td>18–19</td>
<td>3164</td>
</tr>
<tr>
<td>19–20</td>
<td>2950</td>
</tr>
<tr>
<td>20–21</td>
<td>2601</td>
</tr>
<tr>
<td>21–22</td>
<td>2420</td>
</tr>
<tr>
<td>22–23</td>
<td>2557</td>
</tr>
<tr>
<td>23–00</td>
<td>4319</td>
</tr>
</tbody>
</table>

**Figure 12.6** Hourly casualties on Fridays

There are two clear peaks of activity (one in the afternoon, particularly between 4 p.m. and 5 p.m., and another just before midnight). Possibly there is a peak in the early morning between 8 a.m. and 9 a.m. A formal analysis of these data would require some sort of probability model (not a Poisson process, which would provide a very bad fit) for the incidence of casualties over time.

There is a random process called the **simple birth process** used to model the size of growing populations. One of the assumptions of the model is that the birth rate remains constant with passing time; one of its consequences is that (apart from random variation) the size of the population increases exponentially, without limit. For most populations, therefore, this would be a useful model only in the earliest stages of population development: usually the birth rate would tail off as resources become more scarce and the population attains some kind of steady state.

One context where the simple birth process might provide a useful model is the duckweed data in Chapter 10, Table 10.7 and Figure 10.11. The question is: is the birth rate constant or are there signs of it tailing off with passing...
time? In the case of these data, we saw in Figure 10.22 that a logarithmic plot of population size against time was very suggestive of a straight line. At least over the first eight weeks of the development of the duckweed colony, there appears to be no reduction in the birth rate.

Similarly, some populations only degrade, and there are many variations on the random process known under the general term of death process.

**Example 12.6 Green sunfish**

A study was undertaken to investigate the resistance of green sunfish *Lepomis cyanellus* to various levels of thermal pollution. Warming of water occurs, for instance, at the seaside outflows of power stations: it is very important to discover the consequences of this for the local environment. Twenty fish were introduced into water heated at 39.7°C and the numbers of survivors were recorded at pre-specified regular intervals. This followed a five-day acclimatization at 35°C. The data are given in Table 12.6. In fact, the experiment was repeated a number of times with water at different temperatures. The sort of question that might be asked is: how does mean survival time depend on water temperature? In order to answer this question, some sort of probability model for the declining population needs to be developed. Here, the random process \{X_n; n = 0, 5, 10, \ldots\} gives the numbers of survivors counted at five-second intervals.

Example 12.7 gives details of some very old data indeed! They come from Daniel Defoe's *Journal of the Plague Year* published in 1722, describing the outbreak of the bubonic plague in London in 1665.

**Example 12.7 The Great Plague**

Defoe used published bills of mortality as indicators of the progress of the epidemic. Up to the outbreak of the disease, the usual number of burials in London each week varied from about 240 to 300; at the end of 1664 and into 1665 it was noticed that the numbers were climbing into the 300s and during the week 17–24 January there were 474 burials. Defoe wrote: 'This last bill was really frightful, being a higher number than had been known to have been buried in one week since the preceding visitation [of the plague] of 1656.' The data in Table 12.7 give the weekly deaths from all diseases, as well as from the plague, during nine weeks from late summer into autumn 1665.

<table>
<thead>
<tr>
<th>Table 12.6</th>
<th>Survival times of green sunfish (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (seconds)</td>
<td>Survivors</td>
</tr>
<tr>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>45</td>
<td>20</td>
</tr>
<tr>
<td>50</td>
<td>19</td>
</tr>
<tr>
<td>55</td>
<td>19</td>
</tr>
<tr>
<td>60</td>
<td>18</td>
</tr>
<tr>
<td>65</td>
<td>17</td>
</tr>
<tr>
<td>70</td>
<td>14</td>
</tr>
<tr>
<td>75</td>
<td>13</td>
</tr>
<tr>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>85</td>
<td>11</td>
</tr>
<tr>
<td>90</td>
<td>8</td>
</tr>
<tr>
<td>95</td>
<td>7</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
</tr>
<tr>
<td>105</td>
<td>6</td>
</tr>
<tr>
<td>110</td>
<td>4</td>
</tr>
<tr>
<td>115</td>
<td>3</td>
</tr>
<tr>
<td>120</td>
<td>3</td>
</tr>
<tr>
<td>125</td>
<td>1</td>
</tr>
<tr>
<td>130</td>
<td>1</td>
</tr>
<tr>
<td>135</td>
<td>0</td>
</tr>
</tbody>
</table>

The question arises: is there a peak of activity before which the epidemic may still be said to be taking off, and after which it may be said to be fading out?
In order for the timing of such a peak to be identified, a model for deaths due to the disease would need to be formulated. The diagram in Figure 12.7 informally identifies the peak for deaths due to the plague. However, there is a small diminution at Week 5, making conclusions not quite obvious. (For instance, it is possible that had the data been recorded midweek-to-midweek rather than weekend-to-weekend, then the numbers might tell a somewhat different story.)

The example is interesting because, at the time of writing, much current effort is devoted to Aids research, including the construction of useful models for the spread of the associated virus within and between communities. Different diseases require fundamentally different models to represent the different ways in which they may be spread.

For most populations, allowance has to be made in the model to reflect fluctuations in the population size due to births and deaths. Such models are called birth-and-death processes. Such a population is described in Example 12.8.

Example 12.8 Whooping cranes

The whooping crane \((Grus americana)\) is a rare migratory bird which breeds in the north of Canada and winters in Texas. The data in Table 12.8 give annual counts of the number of whooping cranes arriving in Texas each autumn between 1938 and 1972.

A plot of the fluctuating crane population \(\{X(t); t = 1938, 1939, \ldots, 1972\}\) against time is given in Figure 12.8.

For most populations, allowance has to be made in the model to reflect fluctuations in the population size due to births and deaths. Such models are called birth-and-death processes. Such a population is described in Example 12.8.

You can see from the plot that there is strong evidence of an increasing trend in the crane population. The data may be used to estimate the birth and death rates for a birth-and-death model, which may then be tested against the data. (Alternatively, perhaps, some sort of regression model might be attempted.)
There is one rather surprising application of the birth-and-death model: that is, to queueing. This is a phenomenon that has achieved an enormous amount of attention in the statistical literature. The approach is to think of arrivals as 'births' and departures as 'deaths' in an evolving population (which may consist of customers in a bank or at a shop, for instance).

**Example 12.9 A queue**

One evening in 1994, vehicles were observed arriving at and leaving a petrol station in north-west Milton Keynes, England. The station could accommodate a maximum of twelve cars, and at each service point a variety of fuels was available (leaded and unleaded petrol, and diesel fuel). When observation began there were three cars at the station. In Table 12.9 the symbol A denotes an arrival (a vehicle crossed the station entrance) and D denotes a departure (a vehicle crossed the exit line). Observation ceased after 30 minutes.

The sequence of arrivals and departures is illustrated in Figure 12.9, where the random variable $X(t)$ (the number of vehicles in the station at time $t$) is plotted against passing time. The random process $\{X(t); t \geq 0\}$ is an integer-valued continuous-time random process.

Data provided by F. Daly, The Open University.

### Table 12.9 The number of cars at a petrol station

<table>
<thead>
<tr>
<th>Time</th>
<th>Type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0m 00s</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2m 20s</td>
<td>A</td>
<td>4</td>
</tr>
<tr>
<td>4m 02s</td>
<td>D</td>
<td>3</td>
</tr>
<tr>
<td>4m 44s</td>
<td>D</td>
<td>2</td>
</tr>
<tr>
<td>6m 42s</td>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>7m 42s</td>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>8m 15s</td>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>10m 55s</td>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>11m 11s</td>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>14m 40s</td>
<td>D</td>
<td>2</td>
</tr>
<tr>
<td>15m 18s</td>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>16m 22s</td>
<td>A</td>
<td>4</td>
</tr>
<tr>
<td>16m 51s</td>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>19m 18s</td>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>20m 36s</td>
<td>D</td>
<td>3</td>
</tr>
<tr>
<td>20m 40s</td>
<td>D</td>
<td>2</td>
</tr>
<tr>
<td>21m 20s</td>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>22m 23s</td>
<td>D</td>
<td>2</td>
</tr>
<tr>
<td>22m 44s</td>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>24m 32s</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>28m 17s</td>
<td>A</td>
<td>1</td>
</tr>
</tbody>
</table>

The study of queues is still an important research area. Even simple models representing only the most obvious of queue dynamics can become mathematically very awkward; and some situations require complicated models. Elements of a queueing model include the probability distribution of the times between successive arrivals at the service facility, and the probability distribution of the time spent by customers at the facility. More sophisticated models include, for instance, the possibility of 'baulking': potential customers are discouraged at the sight of a long queue and do not join it.

In Example 12.7 the data consisted of deaths due to the plague. It might have been interesting but not practicable to know the numbers of individuals afflicted with the disease at any one time. In modern times, this is very important—for instance, one needs to be able to forecast health demands and
resources. The study of epidemic models is a very important part of the theory of random processes. The next example deals with a much smaller population.

**Example 12.10 A smallpox epidemic**

The data in Table 12.10 are from a smallpox outbreak in the community of Abakaliki in south-eastern Nigeria. There were 30 cases in a population of 120 individuals at risk. The data give the 29 inter-infection times (in days). A zero corresponds to cases appearing on the same day. Notice the long pause between the first and second reported cases.

A plot showing the progress of the disease is given in Figure 12.10.

**Figure 12.10** New cases of smallpox

Here, the numbers are very much smaller than they were for the plague data in Example 12.7. One feature that is common to many sorts of epidemic is that they start off slowly (there are many susceptible individuals, but rather few carriers of the disease) and end slowly (when there are many infected persons, but few survivors are available to catch the disease); the epidemic activity peaks when there are relatively large numbers of both infectives and susceptibles. If this is so, then the model must reflect this feature. Of course, different models will be required for different diseases, because different diseases often have different transmission mechanisms.

Regular observations are often taken on some measure (such as unemployment, interest rates, prices) in order to identify trends or cycles. The techniques of time series analysis are frequently applied to such data.

**Example 12.11 Sales of jeans**

Data were collected on monthly sales of jeans as part of an exercise in forecasting future sales. From the data, is there evidence of trend (a steady increase or decrease in demand) or of seasonal variation (sales varying with

**Table 12.10** Waiting times between 30 cases of smallpox (days)

<table>
<thead>
<tr>
<th>Days</th>
<th>13</th>
<th>7</th>
<th>2</th>
<th>3</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here 'infectives' are those who have and who are capable of spreading the disease, and 'susceptibles' are those who have not yet contracted the disease.
the time of year)? If trend or seasonal variation is identified and removed, how might the remaining variation be modelled?

<table>
<thead>
<tr>
<th>Table 12.11</th>
<th>Monthly sales of jeans in the UK, 1980–1985 (thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>Jan</td>
</tr>
<tr>
<td>1982</td>
<td>1969</td>
</tr>
<tr>
<td>1983</td>
<td>2149</td>
</tr>
<tr>
<td>1984</td>
<td>2319</td>
</tr>
<tr>
<td>1985</td>
<td>2137</td>
</tr>
</tbody>
</table>

These data may be summarized in a time series diagram, as shown in Figure 12.11.

![Time series diagram](image)

**Figure 12.11** Monthly sales of jeans in the UK, 1980–1985 (thousands)

Here, a satisfactory model might be achieved simply by superimposing random normal variation $N(0, \sigma^2)$ on trend and seasonal components (if any). Then the trend component, the seasonal component and the value of the parameter $\sigma$ would need to be estimated from the data.

Meteorology is another common context for time series data.

**Example 12.12 Monthly temperatures**

Average air temperatures were recorded every month for twenty years at Nottingham Castle, England. This example is similar to Example 12.11; but the time series diagram in Figure 12.12 exhibits only seasonal variation and no trend. (One might be worried if in fact there was any trend evident!)
Table 12.12  Average monthly temperatures at Nottingham Castle, England, 1920–1939 (°F)

<table>
<thead>
<tr>
<th>Year</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1920</td>
<td>40.6</td>
<td>40.8</td>
<td>44.4</td>
<td>46.7</td>
<td>54.1</td>
<td>58.5</td>
<td>57.7</td>
<td>56.4</td>
<td>54.3</td>
<td>50.5</td>
<td>42.9</td>
<td>39.8</td>
</tr>
<tr>
<td>1921</td>
<td>44.2</td>
<td>39.8</td>
<td>45.1</td>
<td>47.0</td>
<td>54.1</td>
<td>58.7</td>
<td>66.3</td>
<td>59.9</td>
<td>57.0</td>
<td>54.2</td>
<td>39.7</td>
<td>42.8</td>
</tr>
<tr>
<td>1922</td>
<td>37.5</td>
<td>38.7</td>
<td>39.5</td>
<td>42.1</td>
<td>55.7</td>
<td>57.8</td>
<td>56.8</td>
<td>54.3</td>
<td>54.3</td>
<td>47.1</td>
<td>41.8</td>
<td>41.7</td>
</tr>
<tr>
<td>1923</td>
<td>41.8</td>
<td>40.1</td>
<td>42.9</td>
<td>45.8</td>
<td>49.2</td>
<td>52.7</td>
<td>64.2</td>
<td>59.6</td>
<td>54.4</td>
<td>49.2</td>
<td>36.3</td>
<td>37.6</td>
</tr>
<tr>
<td>1924</td>
<td>39.3</td>
<td>37.5</td>
<td>38.3</td>
<td>45.5</td>
<td>53.2</td>
<td>57.7</td>
<td>60.8</td>
<td>58.2</td>
<td>56.4</td>
<td>49.8</td>
<td>44.4</td>
<td>43.6</td>
</tr>
<tr>
<td>1925</td>
<td>40.0</td>
<td>40.5</td>
<td>40.8</td>
<td>45.1</td>
<td>53.8</td>
<td>59.4</td>
<td>63.5</td>
<td>61.0</td>
<td>53.0</td>
<td>50.0</td>
<td>38.1</td>
<td>36.3</td>
</tr>
<tr>
<td>1926</td>
<td>39.2</td>
<td>43.4</td>
<td>43.4</td>
<td>48.9</td>
<td>50.6</td>
<td>56.8</td>
<td>62.5</td>
<td>62.0</td>
<td>57.5</td>
<td>46.7</td>
<td>41.6</td>
<td>39.8</td>
</tr>
<tr>
<td>1927</td>
<td>39.4</td>
<td>38.5</td>
<td>45.3</td>
<td>47.1</td>
<td>51.7</td>
<td>55.0</td>
<td>60.4</td>
<td>60.5</td>
<td>54.7</td>
<td>50.3</td>
<td>42.3</td>
<td>35.2</td>
</tr>
<tr>
<td>1928</td>
<td>40.8</td>
<td>41.1</td>
<td>42.8</td>
<td>47.3</td>
<td>50.9</td>
<td>56.4</td>
<td>62.2</td>
<td>60.5</td>
<td>55.4</td>
<td>50.2</td>
<td>43.0</td>
<td>37.3</td>
</tr>
<tr>
<td>1929</td>
<td>34.8</td>
<td>31.3</td>
<td>41.0</td>
<td>43.9</td>
<td>53.1</td>
<td>56.9</td>
<td>62.5</td>
<td>60.3</td>
<td>59.8</td>
<td>49.2</td>
<td>42.9</td>
<td>41.9</td>
</tr>
<tr>
<td>1930</td>
<td>41.6</td>
<td>37.1</td>
<td>41.2</td>
<td>46.9</td>
<td>51.2</td>
<td>60.4</td>
<td>60.1</td>
<td>61.6</td>
<td>57.0</td>
<td>50.9</td>
<td>43.0</td>
<td>38.8</td>
</tr>
<tr>
<td>1931</td>
<td>37.1</td>
<td>38.4</td>
<td>38.4</td>
<td>46.5</td>
<td>53.5</td>
<td>58.4</td>
<td>60.6</td>
<td>58.2</td>
<td>53.8</td>
<td>46.6</td>
<td>45.5</td>
<td>40.6</td>
</tr>
<tr>
<td>1932</td>
<td>42.4</td>
<td>38.4</td>
<td>40.3</td>
<td>44.6</td>
<td>50.9</td>
<td>57.0</td>
<td>62.1</td>
<td>63.5</td>
<td>56.3</td>
<td>47.3</td>
<td>43.6</td>
<td>41.8</td>
</tr>
<tr>
<td>1933</td>
<td>36.2</td>
<td>39.3</td>
<td>44.5</td>
<td>48.7</td>
<td>54.2</td>
<td>60.8</td>
<td>65.5</td>
<td>64.9</td>
<td>60.1</td>
<td>50.2</td>
<td>42.1</td>
<td>35.8</td>
</tr>
<tr>
<td>1934</td>
<td>39.4</td>
<td>38.2</td>
<td>40.4</td>
<td>46.9</td>
<td>53.4</td>
<td>59.6</td>
<td>66.5</td>
<td>60.4</td>
<td>59.2</td>
<td>51.2</td>
<td>42.8</td>
<td>45.8</td>
</tr>
<tr>
<td>1935</td>
<td>40.0</td>
<td>42.6</td>
<td>43.5</td>
<td>47.1</td>
<td>50.0</td>
<td>60.5</td>
<td>64.6</td>
<td>64.0</td>
<td>56.8</td>
<td>48.6</td>
<td>44.2</td>
<td>36.4</td>
</tr>
<tr>
<td>1936</td>
<td>37.3</td>
<td>35.0</td>
<td>44.0</td>
<td>43.9</td>
<td>52.7</td>
<td>58.6</td>
<td>60.0</td>
<td>61.1</td>
<td>58.1</td>
<td>49.6</td>
<td>41.6</td>
<td>41.3</td>
</tr>
<tr>
<td>1937</td>
<td>40.8</td>
<td>41.0</td>
<td>38.4</td>
<td>47.4</td>
<td>54.1</td>
<td>58.6</td>
<td>61.4</td>
<td>61.8</td>
<td>56.3</td>
<td>50.9</td>
<td>41.4</td>
<td>37.1</td>
</tr>
<tr>
<td>1938</td>
<td>42.1</td>
<td>41.2</td>
<td>47.3</td>
<td>46.6</td>
<td>52.4</td>
<td>59.0</td>
<td>59.6</td>
<td>60.4</td>
<td>57.0</td>
<td>50.7</td>
<td>47.8</td>
<td>39.2</td>
</tr>
<tr>
<td>1939</td>
<td>39.4</td>
<td>40.9</td>
<td>42.4</td>
<td>47.8</td>
<td>52.4</td>
<td>58.0</td>
<td>60.7</td>
<td>61.8</td>
<td>58.2</td>
<td>46.7</td>
<td>46.6</td>
<td>37.8</td>
</tr>
</tbody>
</table>

The time series diagram in Figure 12.12 shows the seasonal variation quite clearly.

Average monthly temperatures (°F)

![Graph of average monthly temperatures at Nottingham Castle, 1920–1939]

Figure 12.12  Average monthly temperatures at Nottingham Castle, England, 1920–1939 (°F)

Here, it is possible that superimposing random normal variation on seasonal components would not be enough. There is likely to be substantial correlation between measurements taken in successive months, and a good model ought to reflect this. ■
Example 12.13 is an example where one might look for trend as well as seasonal features in the data.

**Example 12.13 Death statistics, lung diseases**

The data in Table 12.13 list monthly deaths, for both sexes, from bronchitis, emphysema and asthma in the United Kingdom between 1974 and 1979. The sorts of research questions that might be posed are: what can be said about the way the numbers of deaths depend upon the time of year (seasonal variation)? What changes are evident, if any, between 1974 and 1979 (trend)?

<table>
<thead>
<tr>
<th>Year</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1974</td>
<td>3035</td>
<td>2552</td>
<td>2704</td>
<td>2554</td>
<td>2014</td>
<td>1655</td>
<td>1721</td>
<td>1524</td>
<td>1596</td>
<td>2074</td>
<td>2199</td>
<td>2512</td>
</tr>
<tr>
<td>1975</td>
<td>2933</td>
<td>2889</td>
<td>2938</td>
<td>2497</td>
<td>1870</td>
<td>1726</td>
<td>1607</td>
<td>1545</td>
<td>1396</td>
<td>1787</td>
<td>2076</td>
<td>2837</td>
</tr>
<tr>
<td>1976</td>
<td>2787</td>
<td>3891</td>
<td>3179</td>
<td>2011</td>
<td>1636</td>
<td>1580</td>
<td>1489</td>
<td>1300</td>
<td>1356</td>
<td>1653</td>
<td>2013</td>
<td>2823</td>
</tr>
<tr>
<td>1977</td>
<td>2996</td>
<td>2523</td>
<td>2540</td>
<td>2520</td>
<td>1994</td>
<td>1641</td>
<td>1691</td>
<td>1479</td>
<td>1596</td>
<td>1877</td>
<td>2032</td>
<td>2484</td>
</tr>
<tr>
<td>1978</td>
<td>2899</td>
<td>2990</td>
<td>2890</td>
<td>2379</td>
<td>1933</td>
<td>1734</td>
<td>1617</td>
<td>1495</td>
<td>1440</td>
<td>1777</td>
<td>1970</td>
<td>2745</td>
</tr>
<tr>
<td>1979</td>
<td>2841</td>
<td>3535</td>
<td>3010</td>
<td>2091</td>
<td>1667</td>
<td>1589</td>
<td>1518</td>
<td>1349</td>
<td>1392</td>
<td>1619</td>
<td>1954</td>
<td>2633</td>
</tr>
</tbody>
</table>

The diagram in Figure 12.13 exhibits some seasonal variation but there is little evidence of substantial differences between the start and end of the series. In order to formally test hypotheses of no change, it would be necessary to formulate a mathematical model for the way in which the data are generated.

![Deaths from lung diseases](image)

**Figure 12.13** Monthly deaths from lung diseases in the UK, 1974–1979

Interestingly, most months show a decrease in the counts from 1974 to 1979. However, February shows a considerable rise. Without a model, it is not easy to determine whether or not these differences are significant.

Often time series exhibit cycles that are longer than a year, and for which there is not always a clear reason. A famous set of data is the so-called ‘lynx data’, much explored by statisticians. The data are listed in Table 12.14.

---

Example 12.14 The lynx data

The data in Table 12.14 were taken from the records of the Hudson's Bay Company. They give the annual number of Canadian lynx trapped in the Mackenzie River district of north-west Canada between 1821 and 1934.

Table 12.14 Lynx trappings, 1821–1934

<table>
<thead>
<tr>
<th>Year</th>
<th>1821</th>
<th>1841</th>
<th>1861</th>
<th>1881</th>
<th>1901</th>
<th>1921</th>
<th>1941</th>
</tr>
</thead>
<tbody>
<tr>
<td>1820</td>
<td>269</td>
<td>321</td>
<td>585</td>
<td>871</td>
<td>1475</td>
<td>2821</td>
<td>3928</td>
</tr>
<tr>
<td>1830</td>
<td>523</td>
<td>98</td>
<td>184</td>
<td>279</td>
<td>409</td>
<td>2285</td>
<td>2685</td>
</tr>
<tr>
<td>1840</td>
<td>151</td>
<td>45</td>
<td>68</td>
<td>213</td>
<td>546</td>
<td>1033</td>
<td>2129</td>
</tr>
<tr>
<td>1850</td>
<td>377</td>
<td>225</td>
<td>360</td>
<td>731</td>
<td>1638</td>
<td>2725</td>
<td>2871</td>
</tr>
<tr>
<td>1860</td>
<td>236</td>
<td>245</td>
<td>552</td>
<td>1623</td>
<td>3311</td>
<td>6721</td>
<td>4254</td>
</tr>
<tr>
<td>1870</td>
<td>358</td>
<td>784</td>
<td>1594</td>
<td>1676</td>
<td>2251</td>
<td>1426</td>
<td>756</td>
</tr>
<tr>
<td>1880</td>
<td>469</td>
<td>736</td>
<td>2042</td>
<td>2811</td>
<td>4431</td>
<td>2511</td>
<td>389</td>
</tr>
<tr>
<td>1890</td>
<td>59</td>
<td>188</td>
<td>377</td>
<td>1292</td>
<td>4031</td>
<td>3495</td>
<td>587</td>
</tr>
<tr>
<td>1900</td>
<td>758</td>
<td>1307</td>
<td>3465</td>
<td>6991</td>
<td>6313</td>
<td>3794</td>
<td>1836</td>
</tr>
<tr>
<td>1910</td>
<td>1388</td>
<td>2713</td>
<td>3800</td>
<td>3091</td>
<td>2985</td>
<td>3790</td>
<td>374</td>
</tr>
<tr>
<td>1920</td>
<td>229</td>
<td>399</td>
<td>1132</td>
<td>2432</td>
<td>3574</td>
<td>2935</td>
<td>1537</td>
</tr>
<tr>
<td>1930</td>
<td>469</td>
<td>736</td>
<td>2042</td>
<td>2811</td>
<td>4431</td>
<td>2511</td>
<td>389</td>
</tr>
<tr>
<td>1940</td>
<td>59</td>
<td>188</td>
<td>377</td>
<td>1292</td>
<td>4031</td>
<td>3495</td>
<td>587</td>
</tr>
</tbody>
</table>

The counts are shown in Figure 12.14.

![Lynx trappings, 1821–1934](image)

There is evidence of a ten-year cycle in the population counts. Can you see any indications of a 40-year cycle? If the data set were more extended, this could be explored.

Recently researchers have started exploring meteorological data for evidence of long-term climatic changes. The data in Table 1.12 on annual snowfall in Buffalo, NY (USA) are an example. The 30-year period of the data in Example 12.15 is relatively brief for such purposes.
**Example 12.15 March rainfall**

These are data on 30 successive values of March precipitation (in inches) for Minneapolis St Paul. Again, to determine whether there are significant changes with time, one approach would be to formulate a model consistent with there being no fundamental change, and test its fit to the data.

Table 12.15  March precipitation (inches)

<table>
<thead>
<tr>
<th>Year</th>
<th>0.77</th>
<th>1.74</th>
<th>0.81</th>
<th>1.20</th>
<th>1.95</th>
<th>1.20</th>
<th>0.47</th>
<th>1.43</th>
<th>3.37</th>
<th>2.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.00</td>
<td>3.09</td>
<td>1.51</td>
<td>2.10</td>
<td>0.52</td>
<td>1.62</td>
<td>1.31</td>
<td>0.32</td>
<td>0.59</td>
<td>0.81</td>
</tr>
<tr>
<td>1</td>
<td>2.81</td>
<td>1.87</td>
<td>1.18</td>
<td>1.35</td>
<td>4.75</td>
<td>2.48</td>
<td>0.96</td>
<td>1.89</td>
<td>0.90</td>
<td>2.05</td>
</tr>
</tbody>
</table>

The data are shown in Figure 12.15.

![March precipitation, Minneapolis St Paul (inches)](image)

Over this 30-year period there is little evidence of change in rainfall patterns. ■

Finally, collections of two-dimensional data are often of interest.

**Example 12.16 Presence-absence data**

The data illustrated in Figure 12.16 show the presence or absence of the plant *Carex arenaria* over a spatial region divided into a 24 x 24 square lattice. The data may be examined for evidence of clustering or competition between the plants. The plants appear to be less densely populated at the lower left of the region than elsewhere; however, for a formal test of the nature of plant coverage, a model would have to be constructed. ■

In this section a large number of different data sets have been introduced and some modelling approaches have been indicated: unfortunately, it would take another book as long as this one to describe the models and their uses in any detail! For the rest of this chapter, we shall look again at the Bernoulli and Poisson processes, and a new model, the *Markov chain*, will be introduced. The Markov chain is useful under some circumstances for modelling sequences of successes and failures, where the Bernoulli process provides an inadequate model.

![Figure 12.16](image)
12.2 The Bernoulli process as an inadequate model

The Bernoulli process was defined in Chapter 4. For ease, the definition is repeated here.

The Bernoulli process

A Bernoulli process is the name given to a sequence of Bernoulli trials in which

(a) trials are independent;
(b) the probability of success is the same from trial to trial.

Notice the notion of ordering here: it is crucial to the development of the ideas to follow that trials take place sequentially one after the other. Here, in Example 12.17, are two typical realizations of a Bernoulli process.

Example 12.17 Two realizations of a Bernoulli process

Each realization consists of 40 trials.

1111010110110000011010001111001100110
00100011011101100110111110111101010

Each realization is the result of a computer simulation. In both cases the indexing parameter \( p \) for the Bernoulli process (the probability of obtaining a 1) was set equal to 0.7.

The Bernoulli process is an integer-valued random process evolving in discrete time. It provides a useful model for many practical contexts. However, there are several ways in which a random process, consisting of a sequence of Bernoulli trials (that is, a sequence of trials in each of which the result might be success or failure), does not constitute a Bernoulli process. For instance, the probability of success might alter from trial to trial. A situation where this is the case is described in Example 12.18.

Example 12.18 The collector's problem

Packets of breakfast cereals sometimes contain free toys, as part of an advertising promotion designed to appeal primarily to children. Often the toys form part of a family, or collection. The desire on the part of the children to complete the collection encourages purchases of the cereal.

Every time a new packet of cereal is purchased, the toy that it contains will be either new to the consumer's collection, or a duplicate. If each purchase of a new box of cereal is considered as the next in a sequence of trials, then the acquisition of a new toy may be counted as a success, while the acquisition of a duplicate is a failure. Thus each successive purchase may be regarded as a Bernoulli trial.
However, at the beginning of the promotion, the first purchase is bound to produce a new toy, for it will be the first of the collection: so, at trial number 1, the Bernoulli parameter \( p \) is equal to 1. If the number of toys in a completed collection is (say) 8, then at this stage the collector still lacks 7 toys. Assuming that no toy is any more likely to occur than any other in any given packet of cereal, then the probability of success at trial number 2 is \( p = \frac{7}{8} \) (and the probability of failure is \( q = \frac{1}{8} \), corresponding to the case where the second cereal packet contains the same toy as the first). The success probability \( p = \frac{7}{8} \) will remain constant at all subsequent trials until a new toy is added to the collection; then there will be 6 toys missing, and the probability of success at subsequent trials changes again, from \( \frac{7}{8} \) to \( p = \frac{6}{8} = \frac{3}{4} \).

The process continues (assuming the promotion continues long enough) until the one remaining toy of the 8 is acquired (with probability \( p = \frac{3}{4} \) at each trial). Here are two realizations of the collection process. Each trial represents a purchase; a 1 indicates the acquisition of a new toy to the collection, a 0 indicates a duplicate. Necessarily, the process begins with a 1.

```
1 1 1 0 0 1 0 1 1 0 0 0 0 0 1 1
1 1 1 1 0 1 1 1 0 0 0 0 1
```

Actually, each of these realizations is somewhat atypical. In the first case the final toy in the collection was obtained immediately after the seventh, though in general it is likely that several duplicates would be obtained at this point. In the second case the collection was completed remarkably quickly, with seven toys of the collection acquired within the first eight purchases. In the first case it was necessary to make sixteen purchases to complete the collection; in the second case, twelve purchases were necessary. It can be shown that the expected number of purchases is 21.7. In the context of cereal purchase, this is a very large number! In practice, if the collection is sufficiently desirable, children usually increase the rate of new acquisitions by swapping their duplicates with those of others, and add to their collection in that way.

Another example where trials remain independent but where the probability of success alters is where the value of \( p \) depends on the number of the trial.

**Example 12.19 Rainy days**

In Chapter 4, Example 4.1, where the incidence of rainy days was discussed, the remark was made that a Bernoulli process was possibly not a very good model for rain day to day, and that ‘...a better model would probably include some sort of mechanism that allows the parameter \( p \) to vary with the time of the year’. (That is, that the value of \( p \), the probability of rain on any given day in the year, should vary as \( n \), the number of the day, varies from 1 to 365.) It would not be entirely straightforward to express the dependence of \( p \) on \( n \) in mathematical terms, but the principle is clear.

In fact, rainfall and its incidence day to day has been studied probabilistically in some detail, as we shall see in Example 12.20.

There is one very common model for dependence from trial to trial, and that is where the probability of success at any trial depends on what occurred at the trial immediately preceding it. Here are two examples where this sort of model has proved useful in practice.
Example 12.20 The Tel Aviv rainfall study

An alternative rainfall model to that described in Example 12.19 is one where the probability of rain on any given day depends on what occurred the day before. An extensive study was carried out on the weather in Tel Aviv and the following conclusions were reached.

Successive days were classified as wet (outcome 1, a success) or dry (outcome 0, a failure), according to some definition. After analysing weather patterns over many days it was concluded that the probability of a day being wet, given that the previous day was wet, is 0.662; in contrast to this, the probability of a day being wet, given that the preceding day was dry, is only 0.250.

(It follows that the probability that a day is dry, given that the preceding day was dry, is $1 - 0.250 = 0.750$; and the probability that a day is dry, given the preceding day was wet, is $1 - 0.662 = 0.338$.)

We shall need a small amount of mathematical notation at this point. Denoting by the random variable $X_j$ the weather outcome on day $j$, the findings of the study may be written formally as

$$P(X_{j+1} = 1 | X_j = 1) = 0.662; \quad P(X_{j+1} = 1 | X_j = 0) = 0.250. \quad (12.1)$$

These probabilities are called transition probabilities. Notice that they are quite different: the weather of the day before has a substantial effect on the chances of rain on the following day. There is, or appears to be, a very strong association between the weather on consecutive days. For a Bernoulli process, these probabilities would be the same, equal to the parameter $p$—for in a Bernoulli process, success at any trial is independent of the history of the process.

The random process \{X_j, j = 1, 2, 3, \ldots\} gives the sequence of wet and dry days in Tel Aviv. Any actual day-to-day data records are realizations of that random process.

Starting with a dry day on Day 1 (so $X_1 = 0$) here is a sample realization of 40 days' weather in Tel Aviv, using the probabilities given in Example 12.20. The realization was generated using a computer.

$$000011001101011000000000111110011100000 \quad (12.2)$$

You can probably appreciate that, if you did not know, it is not entirely easy to judge whether or not this sequence of trial results was generated as a Bernoulli process. In Section 12.3 we shall see some sample realizations which clearly break the assumptions of the Bernoulli process, and learn how to test the adequacy of the Bernoulli model for a given realization.

Example 12.21 provides a second context where this kind of dependence proves a more useful model than a simple Bernoulli process.

Example 12.21 The sex of children in a family

In Chapter 3, Example 3.13 a Bernoulli process was suggested as a useful model for the sequence of boys and girls in a family, and the value $p = 0.514$ was put forward for the probability of a boy. However, it was almost immediately remarked that ‘... nearly all the very considerable mass of data collected on the sex of children in families suggests that the Bernoulli model would be a very bad model to adopt ... some statistical analyses seem to show that the Gabriel, K.R. and Neumann, J. (1962) A Markov chain model for daily rainfall occurrence at Tel Aviv. Quarterly Journal of the Royal Meteorological Society, 88, 90–95.

Remember the vertical slash notation ‘/’ from Chapter 11: it is read ‘given that’. 


independence assumption of the Bernoulli model breaks down: that is, that Nature has a kind of memory, and the sex of a previous child affects to some degree the probability distribution for the sex of a subsequent child.'

One model has been developed in which transition probabilities are given for boy-boy and girl-boy sequences. The estimated probabilities (writing 1 for a boy and 0 for a girl) are (for one large data set)

\[ P(X_{j+1} = 1|X_j = 1) = 0.4993; \quad P(X_{j+1} = 1|X_j = 0) = 0.5433. \]  \hspace{1cm} (12.3)

These probabilities are quite close; their relative values are very interesting. A preceding girl makes a boy more likely, and vice versa. If Nature has a memory, then it can be said to operate to this effect: in the interests of achieving a balance, Nature tries to avoid long runs of either boys or girls.

In that respect, Example 12.21 is different from Example 12.20: there, wet weather leads to more wet weather, and dry weather leads to more dry weather.

This section ends with two brief exercises.

**Exercise 12.1**

For each of the following contexts, say whether you think a Bernoulli process might provide an adequate model:

(a) the order of males and females in a queue;
(b) the sequence of warm and cool days at one location;
(c) a gambler's luck (whether he wins or loses) at successive card games.

You have already seen at (12.2) a typical realization of Tel Aviv weather day to day. Exercise 12.2 involves computer simulation. You will need to understand how your computer can be made to mimic the outcome of a sequence of Bernoulli trials with altering probabilities.

**Exercise 12.2**

Assuming in both cases that the corresponding model is an adequate representation for the underlying probability mechanisms, use the transition probabilities given in Example 12.20 and Example 12.21 to simulate

(a) a week's weather in Tel Aviv, assuming the first day was wet;
(b) the sequence of boys and girls in a family of four children, assuming the oldest child was a girl.

Development of this model is continued in Section 12.3, where methods are explored for parameter estimation, and for testing the adequacy (or otherwise) of the fit of a Bernoulli model to data.
12.3 Markov chains

12.3.1 Definitions

The probability model for a sequence of Bernoulli trials in which the success probability at any one trial depends on the outcome of the immediately preceding trial is called a Markov chain.

Markov chains
A Markov chain is a random process
\[ \{X_n; n = 1, 2, 3, \ldots \} \]
where
(a) the random variables \( X_n \) are integer-valued (0 or 1);
(b) the transition probabilities defining the process are given by
\[
\begin{align*}
P(X_{j+1} = 0|X_j = 0) &= 1 - \alpha, \\
P(X_{j+1} = 1|X_j = 0) &= \alpha, \\
P(X_{j+1} = 0|X_j = 1) &= \beta, \\
P(X_{j+1} = 1|X_j = 1) &= 1 - \beta;
\end{align*}
\]
(c) \( P(X_1 = 1) = p_1. \)

The transition probabilities may be conveniently written in the form of a square array as
\[
M = \begin{bmatrix}
0 & 1 - \alpha \\
1 - \beta & \alpha \\
\beta & 1 - \beta
\end{bmatrix}
\]
where \( M \) is called the transition matrix.

Notice the labelling of the rows of \( M \); it is helpful to be reminded of which probabilities correspond to which of the two outcomes, 0 and 1. A more extended notation is sometimes used: \( M \) can be written
\[
M = \begin{bmatrix}
0 & 1 \\
1 - \alpha & \alpha \\
1 & \beta \\
\beta & 1 - \beta
\end{bmatrix}
\]
with the row labels indicating the 'current' state \( X_j \) and the column labels the 'next' state \( X_{j+1} \), respectively. In this course we shall use the less elaborate representation, that is,
\[
M = \begin{bmatrix}
0 & 1 - \alpha \\
1 - \beta & \alpha \\
\beta & 1 - \beta
\end{bmatrix}
\]

Over a long period of observation, any realization of a Markov chain will exhibit a number of 0s and 1s. It is possible to show that in the long term, and with the transition probabilities given, the average proportions of 0s and 1s in a long realization will be
\[
q = \frac{\beta}{\alpha + \beta}, \quad p = \frac{\alpha}{\alpha + \beta}
\]
respectively. This result is not complicated to show, but it is notationally awkward, and you need not worry about the details. A 'typical' realization

Markov chains are named after the prominent Russian mathematician A.A. Markov (1856–1922). The term Markov chain actually covers rather a wide class of models, but the terminology will suffice for the process described here. In general, more than two states (0 and 1) are possible; and what has been called a Markov chain here is more precisely known as a 'two-state Markov chain'.

We need to know the value of \( X_1 \) before we can use the transition probabilities to 'drive' the rest of the process.

This particular identification of the transition probabilities may appear to be a little eccentric to you, but it turns out to be convenient.
of a Markov chain may therefore be achieved by taking the starting success probability for \( X_1, p_1 \), as

\[
p_1 = \frac{\alpha}{\alpha + \beta}.
\]

It is useful to give a name to such chains: they are called stationary.

A Markov chain \( \{X_n, n = 1, 2, 3, \ldots\} \) with transition probability matrix

\[
M = \begin{bmatrix}
1 - \alpha & \alpha \\
\beta & 1 - \beta
\end{bmatrix}
\]

and where the probability distribution of \( X_1 \) is given by

\[
P(X_1 = 0) = \frac{\beta}{\alpha + \beta}, \quad P(X_1 = 1) = \frac{\alpha}{\alpha + \beta}
\]

is said to be stationary.

**Example 12.20 continued**

For the Tel Aviv rainfall study, the transition matrix \( M \) is written

\[
M = \begin{bmatrix}
0.750 & 0.250 \\
0.338 & 0.662
\end{bmatrix}
\]

(where 0 indicates a dry day, 1 a wet day). In this sort of situation you might find it clearer to write

\[
M = \begin{bmatrix}
\text{Dry} & 0.750 & 0.250 \\
\text{Wet} & 0.338 & 0.662
\end{bmatrix}.
\]

Based on these transition probabilities, the underlying proportion of wet days in Tel Aviv is given (using (12.4)) by

\[
p = \frac{\alpha}{\alpha + \beta} = \frac{0.250}{0.250 + 0.338} = 0.425.
\]

You could simulate a typical week’s weather in Tel Aviv by taking your first day’s weather from the Bernoulli distribution \( X_1 \sim \text{Bernoulli}(0.425) \) and simulating the next six days’ weather using \( M \).

**Exercise 12.3**

(a) Write down the transition matrix \( M \) for sequences of girls and boys in a family using the figures given at (12.3) in Example 12.21, and estimate the underlying proportion of boys.

(b) Simulate a typical family of five children.

Finally, you should note that a Bernoulli process is a special case of a stationary Markov chain, with transition matrix

\[
M = \begin{bmatrix}
q & p \\
1 & q
\end{bmatrix}.
\]

So the two rows of transition probabilities are the same.
12.3.2 Estimating transition probabilities

A question that now arises is: how do we estimate the transition matrix $M$ from a given realization (assuming a Markov chain provides an adequate model)?

Assume that the given realization includes the outcomes of a total of $n + 1$ trials; this is for ease because then there are $n$ transitions. We need to count the frequency of each of the four possible types of transition: two types denote no change, 00 and 11, and two types denote a switch, 01 and 10.

Let us denote by $n_{00}$ the number of 00 transitions in the sequence, by $n_{01}$ the number of 01 transitions, by $n_{10}$ the number of 10 transitions and by $n_{11}$ the number of 11 transitions. (Then you will find that $n_{00} + n_{01} + n_{10} + n_{11} = n$)

—a useful check on your tally.) If you write these totals in matrix form as

$$N = \begin{bmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{bmatrix}$$

and then write in the row totals

$$N = \begin{bmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{bmatrix} \begin{bmatrix} n_{00} + n_{01} \\ n_{10} + n_{11} \end{bmatrix}$$

(simply to aid your arithmetic) then the estimate $\hat{M}$ of the transition matrix $M$ is given by

$$\hat{M} = \begin{bmatrix} n_{00}/(n_{00} + n_{01}) & n_{01}/(n_{00} + n_{01}) \\ n_{10}/(n_{10} + n_{11}) & n_{11}/(n_{10} + n_{11}) \end{bmatrix}.$$  \hspace{1cm} (12.5)

It turns out that the elements of $\hat{M}$ are the maximum likelihood estimates of the transition probabilities.

Example 12.20 continued

The original data source shows that the Tel Aviv data included altogether 2438 days (so there were 2437 transitions), with the matrix of transition frequencies given by

$$N = \begin{bmatrix} 1049 & 350 \\ 351 & 687 \end{bmatrix} \begin{bmatrix} 1399 \\ 1038 \end{bmatrix}$$

and so

$$\hat{M} = \begin{bmatrix} 1049/1399 & 350/1399 \\ 351/1038 & 687/1038 \end{bmatrix} \begin{bmatrix} 0.750 & 0.250 \\ 0.338 & 0.662 \end{bmatrix};$$

these estimates are the transition probabilities that have been used in all our work. ■

Example 12.22 Counting the transitions

For the sequence of simulated Tel Aviv weather described at (12.2) and repeated here,

$$00001110011001011000000001111110011100000$$

the matrix of transition frequencies is given by

$$N = \begin{bmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{bmatrix} \begin{bmatrix} n_{00} + n_{01} \\ n_{10} + n_{11} \end{bmatrix} = \begin{bmatrix} 0 & 6 \\ 1 & 11 \end{bmatrix} \begin{bmatrix} 16 \\ 6 \end{bmatrix} = 22.$$
Elements of Statistics

(Notice that \( n_{00} + n_{01} + n_{10} + n_{11} = 16 + 6 + 6 + 11 = 39 \), one less than the number of trials.) The corresponding estimated transition matrix \( \hat{M} \) is given by

\[
\hat{M} = \begin{bmatrix}
0 & \frac{16}{22} & \frac{6}{22} \\
1 & \frac{6}{17} & \frac{11}{17}
\end{bmatrix} = \begin{bmatrix}
0.727 & 0.273 \\
0.353 & 0.647
\end{bmatrix}.
\]

You can compare this estimate with the matrix of transition probabilities used to drive the simulation, given by the probabilities at (12.1):

\[
M = \begin{bmatrix}
0.750 & 0.250 \\
0.338 & 0.662
\end{bmatrix}.
\]

The estimate is quite good. 

The following exercise is about estimating transition probabilities.

**Exercise 12.4**

For the following realizations of Markov chains, find the matrix of transition frequencies \( N \), and hence calculate an estimate \( \hat{M} \) of the transition matrix \( M \).

(a) 11110110011001000001111100101100101101
(b) 00010011010110010001001101000000100
(c) 1110000000000000000001111111111111111111101111000
(d) 00001000000000000111110000000100000011
(e) 101001010101010101010101010101010011000000100
(f) 100010101010010001000100100010100110101

Give your estimated probabilities to three decimal places in each case.

In fact, realizations (a) and (b) in Exercise 12.4 are those of a Bernoulli process with parameter \( p = 0.45 \)—equivalently, of a Markov chain with transition matrix

\[
M = \begin{bmatrix}
0.55 & 0.45 \\
0.55 & 0.45
\end{bmatrix}.
\]

Realizations (c) and (d) were each driven by the transition matrix

\[
M = \begin{bmatrix}
0.8 & 0.2 \\
0.3 & 0.7
\end{bmatrix}.
\]

Here, a 0 is very likely to be followed by another 0, and a 1 is likely to be followed by another 1. As you can see, the sample realizations exhibit long runs of 0s and 1s.

Finally, realizations (e) and (f) were generated by the transition matrix

\[
M = \begin{bmatrix}
0.4 & 0.6 \\
0.9 & 0.1
\end{bmatrix}.
\]
Here, a 0 is just more likely than not to be followed by a 1, and a 1 is very likely to be followed by a 0. Both sample realizations exhibit very short runs of 0s and 1s: there is a lot of switching.

The estimates found in Exercise 12.4 show some variation, as might be expected, but in general they are reasonably close.

### 12.3.3 The runs test

We have identified two major types of alternative to the Bernoulli process. One corresponds to realizations exhibiting a relatively small number of long runs of 1s and 0s; and the other to realizations exhibiting a large number of short runs. A run is an unbroken sequence of identical outcomes. The length of a run can be just 1.

It seems intuitively reasonable that a test for the quality of the fit of a Bernoulli model to a set of trials data could therefore be developed using the observed number of runs in the data as a test statistic. An exact test based on the observed number of runs is known as the runs test.

#### Example 12.23 Counting runs

The number of runs in the realization given in Exercise 12.4(d), for instance, may be found as follows. Simply draw lines under consecutive sequences of 0s and 1s and count the total number of lines.

$\begin{align*}
000001000000000011111100000010000011
\end{align*}$

*Figure 12.17* Counting runs

Here, there are 8 runs. Notice that in this case the matrix of transition counts is given by

$$
N = \begin{bmatrix}
0 & 25 & 4 \\
1 & 3 & 7
\end{bmatrix} ;
$$

an alternative way to tally the number of runs is to observe that it is equal to the number of 01 switches plus the number of 10 switches plus 1: that is (denoting the number of runs by $r$),

$$
r = n_{01} + n_{10} + 1 . \quad (12.6)
$$

In this case,

$$
r = 4 + 3 + 1 = 8 .
$$

Next, we shall need to determine whether the observed number of runs is significantly greater or less than the number that might have been expected under a Bernoulli model.

In Exercise 12.5 you should check that your results confirm the formula at (12.6).


**Exercise 12.5**

Count the number of runs in each of the realizations in Exercise 12.4(c) and (e).

The observed number of runs $R$ in a sequence of trials may be used as the test statistic for the hypothesis that the sequence arises as a Bernoulli process. If the observed number of runs is very small or very large, then the hypothesis will be rejected. Like many of the exact tests to which you have been exposed in this course, the runs test is based on a complicated enumeration of possibilities, and only a computer renders the test feasible. You should explore the facilities available on your computer in attempting Exercise 12.6.

**Exercise 12.6**

For each of the six realizations given in Exercise 12.4, test whether they might reasonably be regarded as arising as a Bernoulli process.

For very long sequences of trials, the distribution of $R$ is approximately normally distributed, with mean and variance under the null hypothesis given by

\[
E(R) = \frac{2n_0 n_1}{n_0 + n_1} + 1, \quad V(R) = \frac{2n_0 n_1 (2n_0 n_1 - n_0 - n_1)}{(n_0 + n_1)^2 (n_0 + n_1 - 1)} \quad (12.7)
\]

where $n_0$ is the number of 0s in the sample realization and $n_1$ is the number of 1s.

The Tel Aviv data provided a very long realization indeed, and this approximate test should yield useful results.

**Example 12.20 continued**

For the Tel Aviv data, we only have the matrix of transition counts as follows.

\[
N = \begin{bmatrix}
0 & 1049 & 350 \\
1 & 351 & 687 \\
\end{bmatrix} \begin{bmatrix}
1399 \\
1038 \\
\end{bmatrix}
\]

The number of runs is therefore

\[
r = n_{01} + n_{10} + 1 = 350 + 351 + 1 = 702
\]

(using (12.6)). We also need the overall number of 0s and 1s in the sequence of 2438 trials. In fact, it is in general not possible to deduce the values of $n_0$ and $n_1$ from the matrix $N$ of transition frequencies, without more information. Reference to the original data gives

\[
n_0 = 1400, \quad n_1 = 1038.
\]

Thus the observed value $r = 702$ of $R$ needs to be tested against the normal distribution with mean

\[
E(R) = \frac{2n_0 n_1}{n_0 + n_1} + 1 = \frac{2 \times 1400 \times 1038}{1400 + 1038} + 1 = 1193.12,
\]
and variance

\[ V(R) = \frac{2n_0n_1(2n_0n_1 - n_0 - n_1)}{(n_0 + n_1)^2(n_0 + n_1 - 1)} \]

\[ = \frac{2 \times 1400 \times 1038 \times (2 \times 1400 \times 1038 - 1400 - 1038)}{2438^2 \times 2437} \]

\[ = 582.67, \]

using (12.7). The corresponding \( z \)-score is

\[ z = \frac{r - E(R)}{SD(R)} = \frac{702 - 1193.12}{582.67} = -20.34, \]

to be compared against the standard normal distribution \( N(0,1) \). You can see that the \( SP \) is quite negligible, and there is a strong suggestion (\( r \) much smaller than expected) that the realization exhibits long runs of dry days, and long runs of wet days, which is inconsistent with a Bernoulli model.

The normal approximation to the null distribution of \( R \) is useful as long as \( n_0 \) and \( n_1 \) are each about 20 or more.

**Exercise 12.7**

In a small study of the dynamics of aggression in young children, the following binary sequence of length 24 was observed.

\[ 101111011100001001110000 \]

The binary scores were derived from other aggression scores by coding 1 for high aggression, and 0 otherwise.

(a) Obtain the matrix \( N \) of transition frequencies for these data, and hence obtain an estimate \( \hat{M} \) for the transition matrix \( M \).

(b) Find the number of runs in the data.

(c) Perform an exact test of the hypothesis that the scores could reasonably be regarded as resulting from a Bernoulli process, and state your conclusions.

(d) Notwithstanding the dearth of data, test the same hypothesis using an asymptotic test based on the normal distribution, and state your conclusions.

In Section 12.4 we revisit the Poisson process as a model for random events occurring in continuous time.

**12.4 The Poisson process**

Earlier in the course, we met several examples of realizations of random processes that may be reasonably supposed to be well modelled by a Poisson process. These include the earthquake data of Table 4.7 (inter-event times), the nerve impulse data of Table 4.10 (inter-event times) and the coal mine disaster data of Table 7.4 (inter-event times—in fact, these do not follow an
Elements of Statistics

The data of Table 4.4 constitute radiation counts over $\frac{7}{2}$-second intervals. The Poisson distribution provides a good fit. (See Chapter 9, Example 9.4.)

In Tables 12.2 and 12.3 data are given on computer breakdowns, again exhibiting incidentally the two forms in which such data may arise: as inter-event waiting times, or as counts.

Here is another example of Poisson counts.

Example 12.24 The Prussian horse-kick data

This data set is one of the most well-known data collections. The data are known as Bortkewitsch’s horse-kick data, after their collector. The data in Table 12.16 summarize the numbers of Prussian Militärpersonen killed by the kicks of a horse for each of 14 corps in each of 20 successive years 1875–1894. You can read down the columns to find the annual deaths in each of the 14 corps; sum across the rows for the annual death toll overall.

<table>
<thead>
<tr>
<th>Year</th>
<th>G</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1875</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1876</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1877</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1878</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>1879</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1880</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1881</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>1882</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1883</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>1884</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>1885</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1886</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1887</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1888</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1889</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1890</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1891</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1892</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>1893</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1894</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Total 16 16 12 12 8 11 17 12 7 13 15 25 24 8 196

Most often these appear as summary data. Table 12.17 shows the 196 deaths to have occurred during 280 corps-years. Here, the agreement with the Poisson distribution is moderately good.

<table>
<thead>
<tr>
<th>Deaths</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>&gt;4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>144</td>
<td>91</td>
<td>32</td>
<td>11</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Bortkewitsch had noted that four of the corps (G, 1, 6 and 11) were atypical. If these columns are deleted, the remaining 122 deaths over 200 corps-years are summarized as shown in Table 12.18. Here, the Poisson agreement is very good indeed.

Ladislaus von Bortkewitsch (1868–1931) was born in St Petersburg. After studying in Russia he travelled to Germany, where later he received a Ph. D. from the University of Göttingen. The horse-kick data appear in a monograph of 1898 entitled Das Gesetz der kleinen Zahlen—that is, The Law of Small Numbers. The data are discussed in detail in Preece, D.A., Ross, G.J.S. and Kirby, S.P.J. (1988) Bortkewitsch’s horse-kicks and the generalised linear model. The Statistician, 37, 313–318.

The labelling of the corps is not sequential.
It is common to model the occurrence of such random haphazard accidents as deaths due to horse-kicks, as a Poisson process in time. Apart from the four corps which (for whatever reason) are atypical, this model seems to be a very useful one here.

### 12.4.1 Properties of the Poisson process

We first met the Poisson process in Chapter 4, Section 4.4. Events occur at random in continuous time, as portrayed in Figure 12.18.

![Figure 12.18](image-url)

**Figure 12.18** Events occurring as a Poisson process in time

This is a continuous-time process rather than a discrete-time process. Denoting by $X(t)$ the total number of events to have occurred in any time interval of duration $t$, then the random variable $X(t)$ has a Poisson distribution with parameter $\lambda t$, where $\lambda > 0$ is the average rate of events.

The time $T$ between events has an exponential distribution with mean $\mu = 1/\lambda$: this is written $T \sim M(\lambda)$. These characteristics may be summarized as follows:

The Poisson process

For recurrent events occurring as a Poisson process in time at average rate $\lambda$, the number of events $X(t)$ occurring in time intervals of duration $t$ follows a Poisson distribution with mean $\lambda t$: that is, $X(t) \sim \text{Poisson}(\lambda t)$. The waiting times between consecutive events are independent observations on an exponential random variable $T \sim M(\lambda)$.

Several examples of events that might reasonably be modelled as a Poisson process are listed in Chapter 4, Example 4.17.

### 12.4.2 Testing the adequacy of the fit of a Poisson process

Estimation procedures depend on the form in which the data arise. They may arise as a consequence of continuous observation, in which case they will consist of inter-event waiting times. Commonly, observation will be regular but intermittent: then the data will consist of counts of events over equal time intervals.
You are already very familiar with the estimation of parameters for exponential and Poisson models.

Example 12.25 Estimating the rate of earthquakes

For the earthquake data in Table 4.7 each time interval $t_i, i = 1, 2, \ldots, 62$ may be regarded under the Poisson process model as a data value independently chosen from the exponential distribution with parameter $\lambda$. The mean of such an exponential distribution is $1/\lambda$ (see (4.25)). So, it seems quite reasonable that a point estimate of $\lambda$ should be $1/\bar{t}$, where $\bar{t} = 437$ days is the sample mean waiting time between serious earthquakes. The maximum likelihood estimate of $\lambda$ (the earthquake rate) is $1/437 = 0.0023$ major earthquakes per day.

A 95% confidence interval for the mean $1/\lambda$ (again, assuming an exponential model) is given by $(346, 570)$ days. The corresponding confidence interval for the earthquake rate is $(0.0018, 0.0029)$ earthquakes per day. These confidence intervals are exact. Using the methods in Chapter 7, you could also obtain approximate confidence intervals based on the normal distribution.

A Poisson process also forms a good probability model for certain types of traffic process. Suppose an observer is situated at a particular point at the side of a road and records the time intervals between vehicles passing that spot. A Poisson process will be a reasonable model for such data if the traffic is free-flowing. That is, each vehicle which travels along the road does so at essentially the speed its driver wishes to maintain, with no hindrance from other traffic. In saying this, many traffic situations are ruled out. For example, any road in which traffic is packed densely enough so that traffic queues form is not allowed: an adequate model for such a ‘heavy-traffic’ process would have to make provision for the occurrence of clusters of vehicles close together. Likewise, a main street with many intersections controlled by traffic lights also makes for traffic clumping. On the other hand, a convoy of, say, army trucks travelling long-distance may move in regular spacing, each keeping more or less the same distance from each other. Situations where free-flowing traffic can be assumed therefore reduce either to rather a quiet road where only occasional vehicles pass (singly) along, or, perhaps more usefully, to multi-lane roads such as motorways, which, when not overfull, allow vehicles to progress largely unimpeded by others.

Example 12.26 Yet more traffic data

The data in Table 12.19 are thought to correspond to free-flowing traffic. They consist of 50 vehicle inter-arrival times for vehicles passing an observation point on Burgess Road, Southampton, England, on 24 June 1981. Data provided by Dr J.D. Griffiths of the University of Wales Institute of Science and Technology.

<table>
<thead>
<tr>
<th>Table 12.19 Inter-arrival times (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.2</td>
</tr>
<tr>
<td>3.0</td>
</tr>
<tr>
<td>7.0</td>
</tr>
<tr>
<td>7.5</td>
</tr>
<tr>
<td>12.9</td>
</tr>
</tbody>
</table>

The passage times are illustrated in Figure 12.19.
As we shall see, the data here can be modelled usefully by an exponential distribution.

Sometimes data come in the form of counts: for example, Rutherford and Geiger’s radiation data in Table 4.4. If the counts arise as events in a Poisson process, these data will follow a Poisson distribution. We have seen that the Poisson model provides a good fit. On the other hand, it was remarked that for Table 12.3 in Example 12.3, the Poisson distribution did not provide a good fit to the system breakdown data summarized there.

You could confirm this lack of fit using a chi-squared test, which is an asymptotic test. In this section an exact test is described for testing the quality of the fit of a Poisson process model. The test assumes that the times of occurrence of the events are known—in other words, the test assumes continuous observation.

Now, if events occur according to a Poisson process, then over the whole interval of observation, there should be no identifiable peak of activity where there is an unusually high incidence of events. (Of course, there will be random variation exhibited in the sample realization.) Similarly, there should be no identifiable intervals over which the incidence of events is unusually sparse.

Denoting by \((0, \tau)\) the period of observation, and by

\[ 0 < t_1 < t_2 < \cdots < t_n < \tau \]

the times of occurrence of the \(n\) events observed, then it turns out that the fractions

\[ w(1) = \frac{t_1}{\tau}, w(2) = \frac{t_2}{\tau}, \ldots, w(n) = \frac{t_n}{\tau} \]

may be regarded as a random sample from the standard uniform distribution \(U(0, 1)\), if the underlying process generating the events may be modelled as a Poisson process. This is the only continuous distribution consistent with the idea of 'no preferred time of occurrence'.

A test for whether a list of times of occurrence may be regarded as arising from a Poisson process therefore reduces to testing whether a list of \(n\) numbers all between 0 and 1 may be regarded as a random sample from the standard uniform distribution \(U(0, 1)\). One of the advantages of this test is that no problems of parameter estimation are raised. The test is called the Kolmogorov test. In general, it may be used to test data against any hypothesized continuous distribution. (For example, there is a version of the Kolmogorov test appropriate for testing normality.) We shall only use it as a test of uniformity.

The test consists of comparing the uniform cumulative distribution function

\[ F(w) = w, \quad 0 \leq w \leq 1, \]

with the sample distribution function, or empirical distribution function obtained from the data, defined as follows.
The empirical distribution function

For any random sample \(w_1, w_2, \ldots, w_n\) from a continuous distribution with unknown c.d.f. \(F(w)\), an estimate \(\hat{F}(w)\) of \(F(w)\) called the empirical distribution function may be obtained as follows.

Arrange the sample in ascending order
\[ w_{(1)} \leq w_{(2)} \leq \cdots \leq w_{(n)}. \]

Then the graph of \(\hat{F}(w)\) is constructed as follows:
- for all \(0 < w < w_{(1)}\), \(\hat{F}(w) = 0\);
- for all \(w_{(1)} < w < w_{(2)}\), \(\hat{F}(w) = 1/n\);
- for all \(w_{(2)} < w < w_{(3)}\), \(\hat{F}(w) = 2/n\);
- and so on; finally
- for all \(w_{(n-1)} < w < w_{(n)}\), \(\hat{F}(w) = (n-1)/n\);
- for all \(w_{(n)} < w < 1\), \(\hat{F}(w) = 1\).

At the 'jump points' \(w_{(1)}, w_{(2)}, \ldots, w_{(n)}\), the flat components of the graph of \(\hat{F}(w)\) are joined by vertical components to form a 'staircase'.

The Kolmogorov test consists of superimposing on the graph of the empirical distribution function the cumulative distribution function of the hypothesized model \(F(w)\) (which, in this case, is \(F(w) = w\)), and comparing the two.

Example 12.27 illustrates the technique. The small data set used is artificial, in the interests of simplicity.

**Example 12.27 The empirical distribution function**

Observation is maintained on an evolving system between times 0 and \(\tau = 8\). Over that period of time, five events are observed, occurring at times 0.56, 2.40, 4.08, 4.32, and 7.60. Figure 12.20 shows the occurrence of events over time.

![Figure 12.20](image_url) Five points, possibly occurring as a Poisson process in time

In the nature of things, the times of occurrence are already ordered. If the events occur according to a Poisson process, then the ordered fractions
\[
\begin{align*}
 w_{(1)} &= \frac{t_1}{\tau} = \frac{0.56}{8} = 0.07, \\
 w_{(2)} &= \frac{t_2}{\tau} = \frac{2.40}{8} = 0.30, \\
 w_{(3)} &= \frac{t_3}{\tau} = \frac{4.08}{8} = 0.51, \\
 w_{(4)} &= \frac{t_4}{\tau} = \frac{4.32}{8} = 0.54, \\
 w_{(5)} &= \frac{t_5}{\tau} = \frac{7.60}{8} = 0.95,
\end{align*}
\]
may be regarded as a random sample from the standard uniform distribution $U(0, 1)$. The empirical distribution function is drawn from 0 to 1 with jumps of size $1/n = 0.2$ occurring at the points 0.07, 0.30, 0.51, 0.54, 0.95. This is shown in Figure 12.21.

![Figure 12.21 The empirical distribution function](image)

Now the graph of the c.d.f. $F(w) = w$ of the standard uniform distribution $U(0, 1)$ is superimposed on the figure. This is shown in Figure 12.22.

![Figure 12.22 The empirical distribution function and the hypothesized distribution function](image)

You can see in this case that there are some substantial discrepancies between the empirical distribution function $\hat{F}(w)$ obtained for our data set and the theoretical model $F(w) = w$. Are these differences statistically significant?

If the hypothesis of uniformity holds, the two distribution functions will not be too different. They are equal at the point $(0, 0)$ and at the point $(1, 1)$. Our measure of difference between them, the Kolmogorov test statistic, is defined to be the maximum vertical distance between the two,

$$D = \max_{0 < w < 1} |\hat{F}(w) - F(w)| = \max_{0 < w < 1} |\hat{F}(w) - w|.$$  \hspace{1cm} (12.8)

The distance $D$ is sometimes known as the Kolmogorov distance.

This will always occur at one of the jump points. In Figure 12.22 the maximum difference between the two distribution functions occurs at the point $w = 0.54$, and the size of the difference is $0.8 - 0.54 = 0.26$. This is shown in Figure 12.23.

![Figure 12.23 The Kolmogorov distance $D$: here, $d = 0.26$](image)
So, for Example 12.27, the value of the Kolmogorov test statistic is $d = 0.26$, based on a sample of size $n = 5$. This seems quite large, but to explore whether it is significant (that is, whether the hypothesis of a Poisson process should be rejected) the null distribution of the test statistic $D$ is required. This is extremely complicated, and it is usual at this point to employ statistical software. In fact, the associated $SP$ here is 0.812, which is very high, reflecting the small size of the data set. There is no reason here to reject the hypothesis that the data arise from a Poisson process.

Like the $\chi^2$ test statistic in a test of goodness-of-fit, large values of the Kolmogorov test statistic $D$ lead to rejection of the hypothesized model; small values of $D$ suggest that the fit is good, and that the model is adequate.

**Example 12.1 continued**

In this example, no value is given for $\tau$. Observation ceased exactly with the passage of the last of the 40 cars in Table 12.1: this vehicle passed the observer after $\tau = 12 + 2 + 6 + \cdots + 16 + 2 = 312$ seconds. The first 39 passage times yield observations

$$w(1) = \frac{t_1}{\tau} = \frac{12}{312} = 0.038;$$

$$w(2) = \frac{t_2}{\tau} = \frac{12 + 2}{312} = 0.045;$$

$$w(3) = \frac{t_3}{\tau} = \frac{12 + 2 + 6}{312} = 0.064;$$

and so on. The Kolmogorov test statistic takes the value $d = 0.138$ with $n = 39$. The corresponding $SP$ is 0.412. The exponential fit seems adequate.

**Exercise 12.8**

(a) Use the Kolmogorov test against the null hypothesis of a Poisson process for the traffic data of Example 12.26 (Table 12.19). Interpret your conclusion. (Note here, again, that observation ceased only with the passage of the 50th vehicle.)

(b) Table 12.20 gives the dates (in the form of months after the start of 1851) of the major explosive volcanic eruptions in the northern hemisphere after 1851. The last recorded eruption (corresponding to $t_{56} = 1591$ in the table) occurred in July 1983. In fact, observation was continued until the start of 1985 (so $\tau = 1608$).

<table>
<thead>
<tr>
<th>Table 12.20</th>
<th>Times of major volcanic eruptions (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>38</td>
</tr>
<tr>
<td>392</td>
<td>394</td>
</tr>
<tr>
<td>699</td>
<td>721</td>
</tr>
<tr>
<td>942</td>
<td>1037</td>
</tr>
<tr>
<td>1377</td>
<td>1384</td>
</tr>
<tr>
<td>1553</td>
<td>1564</td>
</tr>
</tbody>
</table>

Use these data to explore whether volcanic eruptions may be assumed to occur as a Poisson process (at least over the period of time that the phenomenon was researched).
12.5 Variations

This chapter concludes with some examples of chance situations where models other than those described so far would be most appropriate. The literature on the subject of model construction and testing for random processes is vast, and all we can do in an introductory text such as this is hope to obtain some notion of the wide area of application.

**Example 12.28 Poisson processes with varying rate**

A time-dependent Poisson process is a Poisson process in which the average rate of occurrence of events alters with passing time. This might be appropriate, for example, in a learning situation where the events recorded are errors—these will become less frequent as time passes. One example where such a model might be appropriate is where the times of system failures were recorded (see Table 12.2). There the remark was made that "... an exponential model does not provide a good fit to the variation in the times between failures ... the failures are becoming less frequent with passing time."

A sketch of a failure rate \( \lambda(t) \) with approximately the right properties is shown in Figure 12.24. In this case the failure rate is always decreasing, but never attains zero. In fact, the rate may be represented by the function

\[
\lambda(t) = \alpha e^{-\beta t}, \quad t \geq 0,
\]

where \( \alpha > 0 \) and \( \beta > 0 \). Once a functional form for \( \lambda(t) \) has been decided (linear, quadratic, trigonometric or, as in this case, exponential decay) then the parameters of the model may be estimated using maximum likelihood. If the model provides a good fit, then it can be used to forecast future failure events. For instance, the number of failures to occur during some future interval \( t_1 < t < t_2 \) follows a Poisson distribution whose mean is given by the integral

\[
\int_{t_1}^{t_2} \lambda(t) \, dt,
\]

shown in Figure 12.25. The selection of an appropriate functional form for \( \lambda(t) \) is not always straightforward; but the ideas of parameter estimation and model testing are not different from those with which you are already familiar.

Example 12.29 is based on the Bernoulli process.

**Example 12.29 Identifying a change in the indexing parameter**

Researchers involved in a grassland experiment were interested in the presence or absence of plants in a protected habitat, year after year. For any given species, they recorded 1 if its presence was observed, 0 otherwise. This was repeated each year and so the data took the form of a sequence of 0s and 1s, possibly consistent with an underlying Bernoulli process, or a Markov chain (or neither of these). The researchers were particularly interested in one species. It certainly appeared that the plant was an intermittent visitor.

A complication was that the plant was easy to mistake: it could be missed (score 0) when it was present; and it could be 'seen' (score 1) when in fact it
was not present—other plants looked very similar to it. So the data were not entirely reliable.

It became known that in reality this particular plant was present in the habitat every year until, one year, it disappeared completely. The problem was to use the sequence of recorded 0s and 1s to estimate when this had happened.

Again, this is a problem involving quite a subtle estimation procedure (not just of the change point, but also of the probabilities of error—recording a 0 when the plant was present and 1 when it was not). It is important to have in mind a useful underlying model, and to be aware of the year-to-year dependencies that need to be built in to it.

A similar but unrelated problem faced researchers studying different plants. The annual visits made by these plants to the habitat were known to be intermittent, but it was also known that after a while there was a possibility of extinction. The problem was to determine from a sequence of recent 0s what probability to attach to the possibility that extinction had taken place. Here, there were no problems of identification. Possibly the year-by-year presence or absence could be reasonably modelled as a Markov chain, for at least as long as the plant survived.

**Example 12.30 Competing species**

Birth-and-death models were briefly explored in Section 12.1. Biologists and ecologists often have to consider populations of two or more species living together and interacting. Sometimes these interactions are co-operative, and sometimes they are antagonistic (though not necessarily detrimental to the future existence of a species). Sometimes such communities are stable, with all species coexisting in a moderately happy association, and sometimes they are not—one species or another eventually disappears because it is not possible indefinitely to sustain life in the presence of another. The literature on the topic is vast, with much work for instance on predator-prey models and host-parasite models.

One important application is to human conflict, and there are some fascinating models to explore.

**Example 12.31 Epidemic models**

There are many variations on models for the spread of disease within communities. It is not just a matter of different parameters being required to reflect differences in virulence, the duration of symptoms, the length of the infectious period, and so on: different diseases are spread in entirely different ways, and the models are fundamentally different.

Populations consist of three types of individual: susceptibles (those who have not yet contracted the disease), infectives (those who have, and who are capable of spreading it) and the rest, who have recovered from the disease. Perhaps this third group are immune, or temporarily immune, to further attacks; or, as with some diseases, there is no difference between somebody who has recovered from the disease and somebody who is susceptible to it. This is, or appears to be, the case with the common cold.

Often a model developed in one context can be used in another, quite different application. Some epidemic models are appropriate to studies of the
diffusion of information, where ‘susceptibles’ are those ignorant of certain information, ‘infectives’ are those who have it, and are passing it on to others, while others have known it but are no longer interested in passing it on. One interesting recent variation includes a model on drug addiction, incorporating susceptibles, addicts and dealers.  

Summary

1. A sequence of Bernoulli trials in which trials are independent and the probability of success is the same from trial to trial is known as a Bernoulli process.

2. A Markov chain is a random process \( \{X_n; \ n = 1, 2, \ldots\} \) taking values 0 or 1, in which \( X_1 \sim \text{Bernoulli}(p_1) \) and thereafter transition probabilities are given by the transition matrix

\[
M = \begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix}.
\]

3. In the long term, the proportions of 0s and 1s in a realization of a Markov chain are given by

\[
q = \frac{\beta}{\alpha + \beta}, \quad p = \frac{\alpha}{\alpha + \beta}
\]

respectively.

A Markov chain in which \( p_1 = \alpha / (\alpha + \beta) \) is called stationary. A Bernoulli process is a special case of a stationary Markov chain.

4. In any realization of a Markov chain the transition frequencies may be summarized through the matrix \( N \) where

\[
N = \begin{bmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{bmatrix} = \begin{bmatrix} n_{00} + n_{01} \\ n_{10} + n_{11} \end{bmatrix};
\]

then the estimate \( \widehat{M} \) of the transition matrix \( M \) is given by

\[
\widehat{M} = \begin{bmatrix} n_{00} / (n_{00} + n_{01}) & n_{01} / (n_{00} + n_{01}) \\ n_{10} / (n_{10} + n_{11}) & n_{11} / (n_{10} + n_{11}) \end{bmatrix}.
\]

5. The number of runs \( r \) in a realization of a Markov chain is given by

\[
r = n_{01} + n_{10} + 1.
\]

The observed value \( r \) of \( R \) may be used to test the hypothesis that a given realization is generated by a Bernoulli process. The test is exact; an approximate test is based on the asymptotic normality of \( R \) where

\[
E(R) = \frac{2n_0n_1}{n_0 + n_1} + 1, \quad V(R) = \frac{2n_0n_1(2n_0n_1 - n_0 - n_1)}{(n_0 + n_1)^2(n_0 + n_1 - 1)};
\]

here, \( n_0 \) is the number of 0s in the sample realization and \( n_1 \) is the number of 1s.

6. In a Poisson process the distribution of $X(t)$, the number of events in time intervals of duration $t$, is $\text{Poisson}(\lambda t)$, where $\lambda$ is the average rate at which events occur. The waiting time between consecutive events follows an exponential distribution $M(\lambda)$.

7. In order to test whether events occurring at times
$$0 < t_1 < t_2 < \cdots < t_n < \tau$$
may be assumed to have been generated according to a Poisson process, the Kolmogorov test may be applied. This tests whether the observations
$$w(j) = \frac{t_j}{\tau}, \quad j = 1, 2, \ldots, n$$
may be assumed to have arisen from a standard uniform distribution.
The Kolmogorov test statistic is the 'Kolmogorov distance'
$$D = \max_{0 < w < 1} |\hat{F}(w) - w|.$$ 
The Kolmogorov test is an exact test.