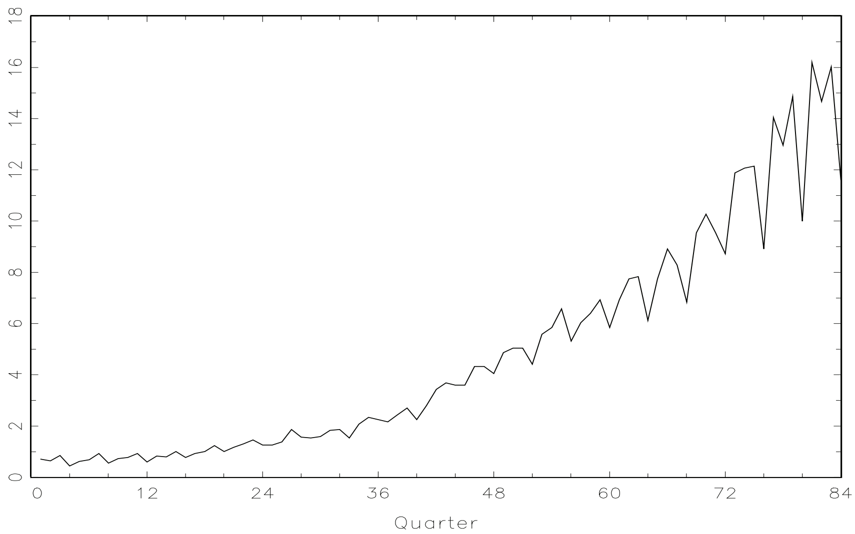

Figure 1.1 Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV.

1.2 The Nature of Time Series Data

Some of the problems and questions of interest to the prospective time series analyst can best be exposed by considering real experimental data taken from different subject areas. The following cases illustrate some of the common kinds of experimental time series data as well as some of the statistical questions that might be asked about such data.

Example 1.1 Johnson & Johnson Quarterly Earnings

Figure 1.1 shows quarterly earnings per share for the U.S. company Johnson & Johnson, furnished by Professor Paul Griffin (personal communication) of the Graduate School of Management, University of California, Davis. There are 84 quarters (21 years) measured from the first quarter of 1960 to the last quarter of 1980. Modeling such series begins by observing the primary patterns in the time history. In this case, note the gradually increasing underlying trend and the rather regular variation superimposed on the trend that seems to repeat over quarters. Methods for analyzing data such as these are explored in Chapter 2 (see Problem 2.1) using regression techniques, and in Chapter 6, §6.5, using structural equation modeling.

To plot the data using the R statistical package, suppose you saved the data as `jj.dat` in the directory `mydata`. Then use the following steps to read in the data and plot the time series (the `>` below are prompts, you

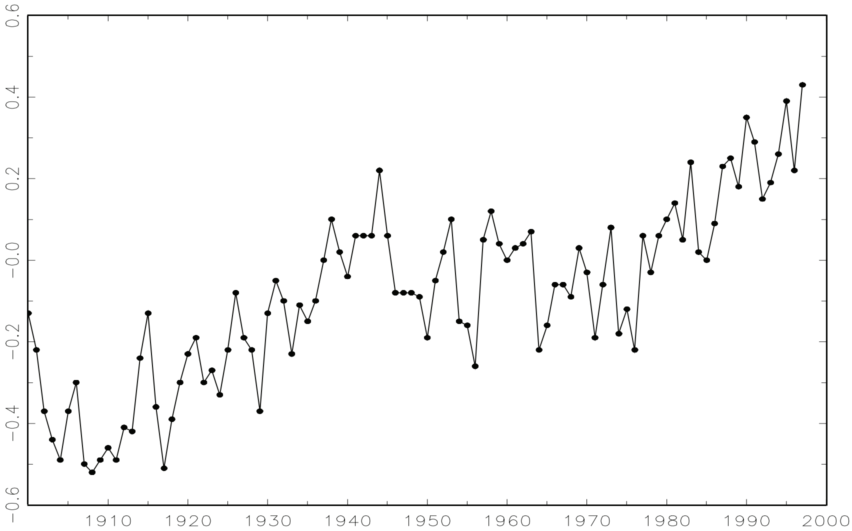

Figure 1.2 Yearly average global temperature deviations (1900–1997) in degrees centigrade.

do not type them):

```
> jj = scan("/mydata/jj.dat")   # yes forward slash
> jj=ts(jj,start=1960, frequency=4)
> plot(jj,  ylab="Quarterly Earnings per Share")
```

You can replace `scan` with `read.table` in this example.

Example 1.2 Global Warming

Consider a global temperature series record, discussed in Jones (1994) and Parker et al. (1994, 1995). The data in Figure 1.2 are a combination of land-air average temperature anomalies (from 1961-1990 average), measured in degrees centigrade, for the years 1900-1997. We note an apparent upward trend in the series that has been used as an argument for the global warming hypothesis. Note also the leveling off at about 1935 and then another rather sharp upward trend at about 1970. The question of interest for global warming proponents and opponents is whether the overall trend is natural or whether it is caused by some human-induced interface. Problem 2.8 examines 634 years of glacial sediment data that might be taken as a long-term temperature proxy. Such percentage changes in temperature do not seem to be unusual over a time period of 100 years. Again, the question of trend is of more interest than particular periodicities.

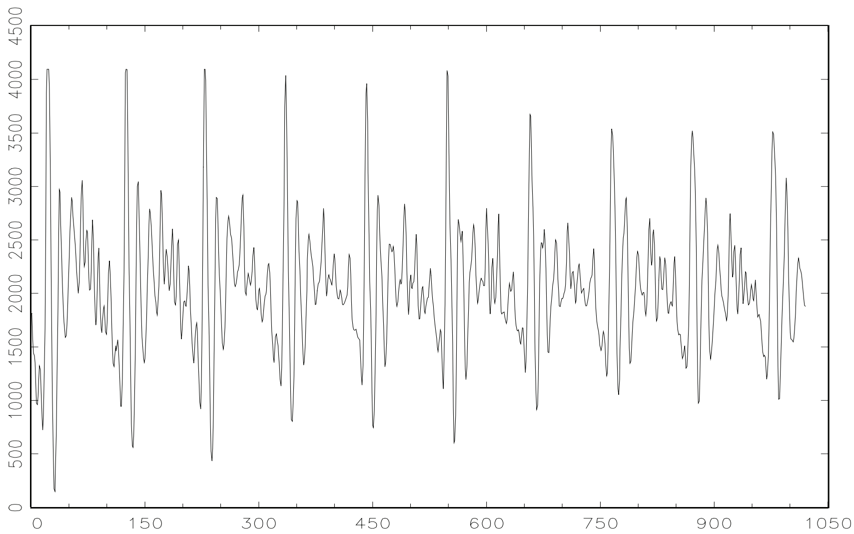

Figure 1.3 Speech recording of the syllable $aaa \cdots hhh$ sampled at 10,000 points per second with $n = 1020$ points.

Example 1.3 Speech Data

More involved questions develop in applications to the physical sciences. Figure 1.3 shows a small .1 second (1000 point) sample of recorded speech for the phrase $aaa \cdots hhh$, and we note the repetitive nature of the signal and the rather regular periodicities. One current problem of great interest is computer recognition of speech, which would require converting this particular signal into the recorded phrase $aaa \cdots hhh$. Spectral analysis can be used in this context to produce a signature of this phrase that can be compared with signatures of various library syllables to look for a match. One can immediately notice the rather regular repetition of small wavelets. The separation between the packets is known as the pitch period and represents the response of the vocal tract filter to a periodic sequence of pulses stimulated by the opening and closing of the glottis.

Example 1.4 New York Stock Exchange

As an example of financial time series data, Figure 1.4 shows the daily returns (or percent change) of the New York Stock Exchange (NYSE) from February 2, 1984 to December 31, 1991. It is easy to spot the crash of October 19, 1987 in the figure. The data shown in Figure 1.4 are typical of return data. The mean of the series appears to be stable

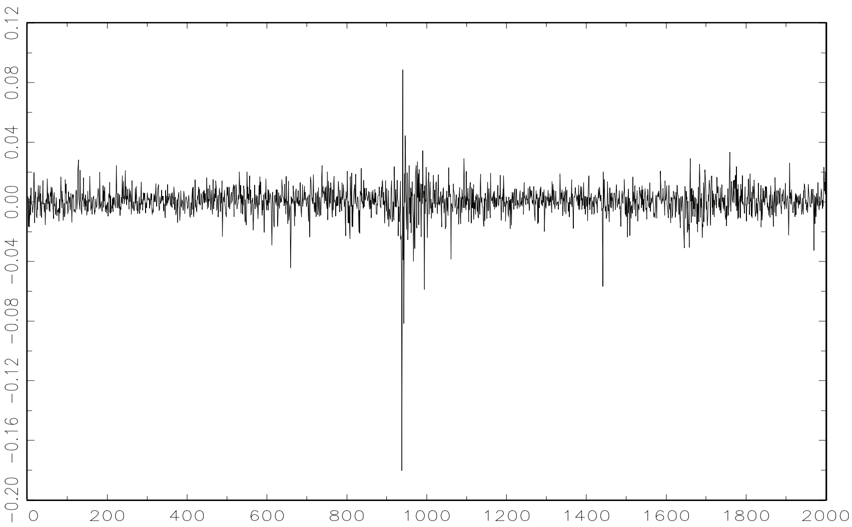

Figure 1.4 Returns of the NYSE. The data are daily value weighted market returns from February 2, 1984 to December 31, 1991 (2000 trading days). The crash of October 19, 1987 occurs at $t = 938$.

with an average return of approximately zero, however, the volatility (or variability) of data changes over time. In fact, the data show volatility clustering; that is, highly volatile periods tend to be clustered together. A problem in the analysis of these type of financial data is to forecast the volatility of future returns. Models such as ARCH and GARCH models (Engle, 1982; Bollerslev, 1986) and stochastic volatility models (Harvey, Ruiz and Shephard, 1994) have been developed to handle these problems. We will discuss these models and the analysis of financial data in Chapters 5 and 6.

Example 1.5 El Niño and Fish Population

We may also be interested in analyzing several time series at once. Figure 1.5 shows monthly values of an environmental series called the Southern Oscillation Index (SOI) and associated Recruitment (number of new fish) furnished by Dr. Roy Mendelssohn of the Pacific Environmental Fisheries Group (personal communication). Both series are for a period of 453 months ranging over the years 1950-1987. The SOI measures changes in air pressure, related to sea surface temperatures in the central Pacific. The central Pacific Ocean warms every three to seven years due to the El Niño effect, which has been blamed, in particular, for the 1997 floods in the midwestern portions of the U.S. Both series in Figure 1.5 tend to exhibit repetitive behavior, with regularly repeating cycles that are easily

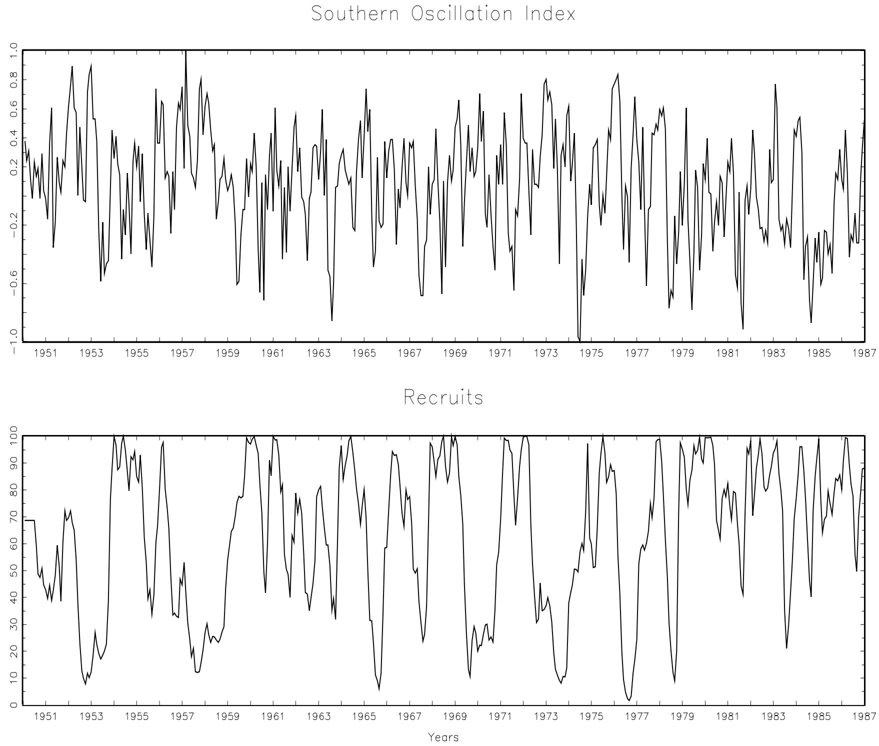

Figure 1.5 Monthly SOI and Recruitment (Estimated new fish), 1950-1987.

visible. This periodic behavior is of interest because underlying processes of interest may be regular and the rate or frequency of oscillation characterizing the behavior of the underlying series would help to identify them. One can also remark that the cycles of the SOI are repeating at a faster rate than those of the Recruitment series. The Recruitment series also shows several kinds of oscillations, a faster frequency that seems to repeat about every 12 months and a slower frequency that seems to repeat about every 50 months. The study of the kinds of cycles and their strengths is the subject of Chapter 4. The two series also tend to be somewhat related; it is easy to imagine that somehow the fish population is dependent on the SOI. Perhaps, even a lagged relation exists, with the SOI signaling changes in the fish population. This possibility suggests trying some version of regression analysis as a procedure for relating the two series. Transfer function modeling, as considered in Chapter 5, can be applied in this case to obtain a model relating Recruitment to its own past and the past values of the SOI Index.

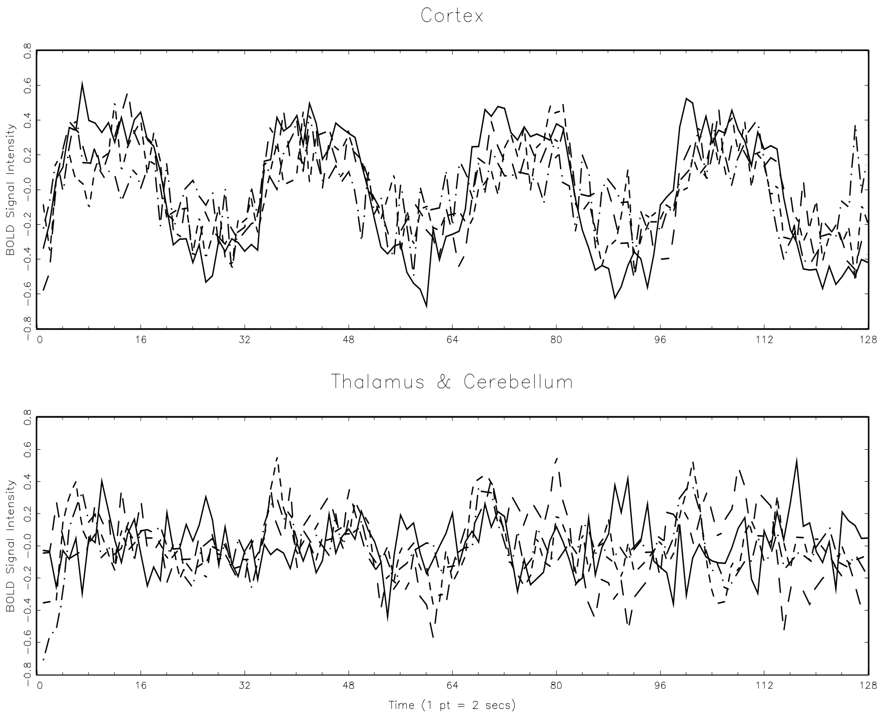

Figure 1.6 fMRI data from various locations in the cortex, thalamus, and cerebellum; $n = 128$ points, one observation taken every 2 seconds.

Example 1.6 fMRI Imaging

A fundamental problem in classical statistics occurs when we are given a collection of independent series or vectors of series, generated under varying experimental conditions or treatment configurations. Such a set of series is shown in Figure 1.6, where we observe data collected from various locations in the brain via functional magnetic resonance imaging (fMRI). In this example, five subjects were given periodic brushing on the hand. The stimulus was applied for 32 seconds and then stopped for 32 seconds; thus, the signal period is 64 seconds. The sampling rate was one observation every 2 seconds for 256 seconds ($n = 128$). For this example, we averaged the results over subjects (these were evoked responses, and all subjects were in phase). The series shown in Figure 1.6 are consecutive measures of blood oxygenation-level dependent (BOLD) signal intensity, which measures areas of activation in the brain. Notice that the periodicities appear strongly in the motor cortex series and less strongly in the thalamus and cerebellum. The fact that one has series from different areas of the brain suggests testing whether the areas are

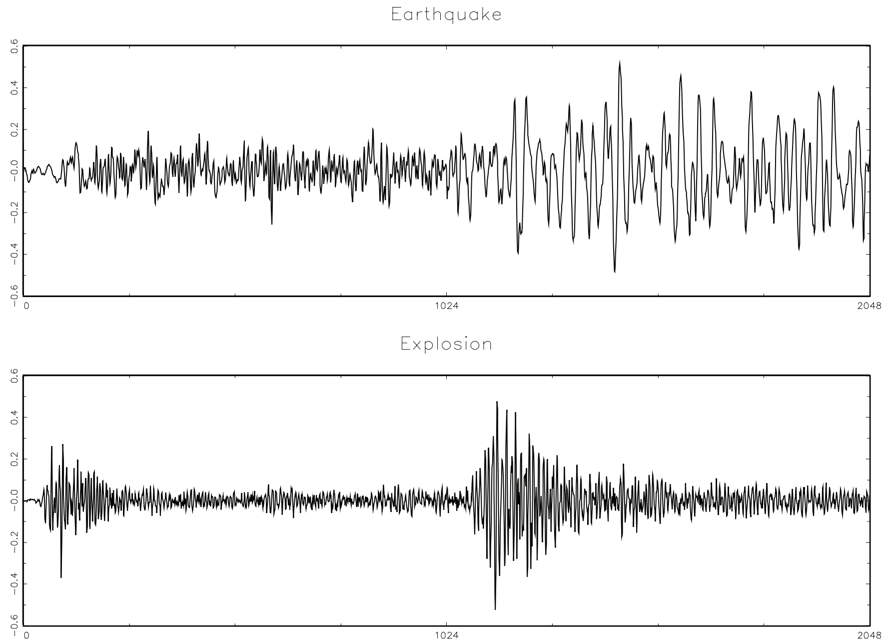

Figure 1.7 Arrival phases from an earthquake (top) and explosion (bottom) at 40 points per second.

responding differently to the brush stimulus. Analysis of variance techniques accomplish this in classical statistics, and we show in Chapter 7 how these classical techniques extend to the time series case, leading to a spectral analysis of variance.

The data are in a file called `fmri.dat`, which consists of nine columns; the first column represents time, whereas the second through ninth columns represent the BOLD signals at eight locations. Assuming the data are located in the directory `mydata`, use the following commands in R to plot the data as in this example.

```
> fmri = read.table("/mydata/fmri.dat")
> par(mfrow=c(2,1))   # sets up the graphics
> ts.plot(fmri[,2:5], lty=c(1,4), ylab="BOLD")
> ts.plot(fmri[,6:9], lty=c(1,4), ylab="BOLD")
```

Example 1.7 Earthquakes and Explosions

As a final example, the series in Figure 1.7 represent two phases or arrivals along the surface, denoted by P ($t = 1, \dots, 1024$) and S ($t =$

$1025, \ldots, 2048$), at a seismic recording station. The recording instruments in Scandinavia are observing earthquakes and mining explosions with one of each shown in Figure 1.7. The general problem of interest is in distinguishing or discriminating between waveforms generated by earthquakes and those generated by explosions. Features that may be important are the rough amplitude ratios of the first phase P to the second phase S, which tend to be smaller for earthquakes than for explosions. In the case of the two events in Figure 1.7, the ratio of maximum amplitudes appears to be somewhat less than .5 for the earthquake and about 1 for the explosion. Otherwise, note a subtle difference exists in the periodic nature of the S phase for the earthquake. We can again think about spectral analysis of variance for testing the equality of the periodic components of earthquakes and explosions. We would also like to be able to classify future P and S components from events of unknown origin, leading to the time series discriminant analysis developed in Chapter 7.

The data are in the file `eq5exp6.dat` as one column with 4096 entries, the first 2048 observations correspond to an earthquake and the next 2048 observations correspond to an explosion. To read and plot the data as in this example, use the following commands in R:

```
> x = matrix(scan("/mydata/eq5exp6.dat"), ncol=2)
> par(mfrow=c(2,1))
> plot.ts(x[,1], main="Earthquake", ylab="EQ5")
> plot.ts(x[,2], main="Explosion", ylab="EXP6")
```

1.3 Time Series Statistical Models

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data, like that encountered in the previous section. In order to provide a statistical setting for describing the character of data that seemingly fluctuate in a random fashion over time, we assume a time series can be defined as a collection of random variables indexed according to the order they are obtained in time. For example, we may consider a time series as a sequence of random variables, $x_1, x_2, x_3, \ldots$, where the random variable x_1 denotes the value taken by the series at the first time point, the variable x_2 denotes the value for the second time period, x_3 denotes the value for the third time period, and so on. In general, a collection of random variables, $\{x_t\}$, indexed by t is referred to as a stochastic process. In this text, t will typically be discrete and vary over the integers $t = 0, \pm1, \pm2, ...$, or some subset of the integers. The observed values of a stochastic process are referred to as a realization of the stochastic process. *Because it will be clear from the context of our discussions, we use the term time series whether we are referring generically to the process or to a particular realization and make no notational distinction between the two concepts.*