Chapter 1
Introduction

1.1 Data Analysis in the Brain Sciences

The brain sciences seek to discover mechanisms by which neural activity is generated, thoughts are created, and behavior is produced. What makes us see, hear, feel, and understand the world around us? How can we learn intricate movements, which require continual corrections for minor variations in path? What is the basis of memory, and how do we allocate attention to particular tasks? Answering such questions is the grand ambition of this broad enterprise and, while the workings of the nervous system are immensely complicated, several lines of now-classical research have made enormous progress: essential features of the nature of the action potential, of synaptic transmission, of sensory processing, of the biochemical basis of memory, and of motor control have been discovered. These advances have formed conceptual underpinnings for modern neuroscience, and have had a substantial impact on clinical practice. The method that produced this knowledge, the scientific method, involves both observation and experiment, but always a careful consideration of the data. Sometimes results from an investigation have been largely qualitative, as in Brenda Milner's documentation of implicit memory retention, together with explicit memory loss, as a result of hippocampal lesioning in patient H.M. In other cases quantitative analysis has been essential, as in Alan Hodgkin and Andrew Huxley's modeling of ion channels to describe the production of action potentials. Today's brain research builds on earlier results using a wide variety of modern techniques, including molecular methods, patch clamp recording, two-photon imaging, single and multiple electrode studies producing spike trains and/or local field potentials (LFPs), optical imaging, electroencephalography (producing EEGs), and functional imaging—positron emission tomography(PET), functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG)—as well as psychophysical and behavioral studies. All of these rely, in varying ways, on vast improvements in data storage, manipulation, and display technologies, as well as corresponding advances in analytical techniques. As a result, data sets from current investigations are often much larger, and more com-

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_1,

plicated, than those of earlier days. For a contemporary student of neuroscience, a working knowledge of basic methods of data analysis is indispensable.

The variety of experimental paradigms across widely ranging investigative levels in the brain sciences may seem intimidating. It would take a multi-volume encyclopedia to document the details of the myriad analytical methods out there. Yet, for all the diversity of measurement and purpose, there are commonalities that make analysis of neural data a single, circumscribed and integrated subject. A relatively small number of principles, together with a handful of ubiquitous techniques—some quite old, some much newer—lay a solid foundation. One of our chief aims in writing this book has been to provide a coherent framework to serve as a starting point in understanding all types of neural data.

In addition to providing a unified treatment of analytical methods that are crucial to progress in the brain sciences, we have a secondary goal. Over many years of collaboration with neuroscientists we have observed in them a desire to learn all that the data have to offer. Data collection is demanding, and time-consuming, so it is natural to want to use the most efficient and effective methods of data analysis. But we have also observed something else. Many neuroscientists take great pleasure in displaying their results not only because of the science involved but also because of the *manner in which* particular data summaries and displays are able to shed light on, and explain, neuroscientific phenomenon; in other words, they have developed a refined appreciation for the data-analytic process itself. The often-ingenious ways investigators present their data have been instructive to us, and have reinforced our own aesthetic sensibilities for this endeavor. There is deep satisfaction in comprehending a method that is at once elegant and powerful, that uses mathematics to describe the world of observation and experimentation, and that tames uncertainty by capturing it and using it to advantage. We hope to pass on to readers some of these feelings about the role of analytical techniques in illuminating and articulating fundamental concepts.

A third goal for this book comes from our exposure to numerous articles that report data analyzed largely by people who lack training in statistics. Many researchers have excellent quantitative skills and intuitions, and in most published work statistical procedures appear to be used correctly. Yet, in examining these papers we have been struck repeatedly by the absence of what we might call statistical thinking, or application of *the statistical paradigm*, and a resulting loss of opportunity to make full and effective use of the data. These cases typically do not involve an incorrect application of a statistical method (though that sometimes does happen). Rather, the lost opportunity is a failure to follow the *general approach* to the analysis of the data, which is what we mean by the label "the statistical paradigm." Our final pedagogical goal, therefore, is to lay out the key features of this paradigm, and to illustrate its application in diverse contexts, so that readers may absorb its main tenets.

To begin, we will review several essential points that will permeate the book. Some of these concern the nature of neural data, others the process of statistical reasoning. As we go over the basic issues, we will introduce some data that will be used repeatedly.

1.1.1 Appropriate analytical strategies depend crucially on the purpose of the study and the way the data are collected.

The answer to the question, "How should I analyze my data?" always depends on what you want to know. Convenient summaries of the data are used to convey apparent tendencies. Particular summaries highlight particular aspects of the data—but they ignore other aspects. At first, the purpose of an investigation may be stated rather vaguely, as in "I would like to know how the responses differ under these two experimental conditions." This by itself, however, is rarely enough to proceed. Usually there are choices to be made, and figuring out what analysis should be performed requires a sharpening of purpose.

Example 1.1 SEF neural activity under two conditions Olson et al. (2000) examined the behavior of neurons in the supplementary eye field (SEF), which is a frontal lobe region anterior to, and projecting to, the eye area in motor cortex. The general issue was whether the SEF merely relays the message to move the eyes, or whether it is involved in some higher-level processing. To distinguish these two possibilities, an experiment was devised in which a monkey moved its eyes in response to either an explicit external cue (the point to which the eyes were to move was illuminated) or an internally-generated translation of a complex cue (a particular pattern at fixation point determined the location to which the monkey was to move its eyes). If the SEF simply transmits the movement message to motor cortex and other downstream areas, one would expect SEF neurons to behave very similarly under the two experimental conditions. On the other hand, distinctions between the neural responses in the two conditions would indicate that the SEF is involved in higher-level cognitive processing. While an individual neuron's activity was recorded from the SEF of an alert macaque monkey, one of the two conditions was chosen at random and applied. This experimental protocol was repeated many times, for each of many neurons. Thus, for each recorded neuron, under each of the two conditions, there were many *trials*, which consist of experimental repetitions designed to be as close to identical as possible.

Results for one neuron are given in Fig. 1.1. The figure displays a pair of raster plots and peri-stimulus time histograms (PSTHs). Each line in each raster plot contains results from a single trial, which consist of a sequence of times at which action potentials or *spikes* occur. The sequence is usually called a *spike train*. Note that for each condition the number and timing of the spikes, displayed on the many lines of each raster plot, vary from trial to trial. The PSTH is formed by creating time bins (here, each bin is 10 ms in length), counting the total number of spikes that occur across all trials within each bin, and then normalizing (by dividing by the number of trials and the length of each bin in seconds) to convert count to firing rate in units of spikes per second. The PSTH is used to display firing-rate trends across time, which are considered to be common to[1] the many separate trials.

[1] One source of variation across trials is that the behavior of the monkey is not identical on every trial. For instance, the eyes may move along slightly different paths and at different rates. Even

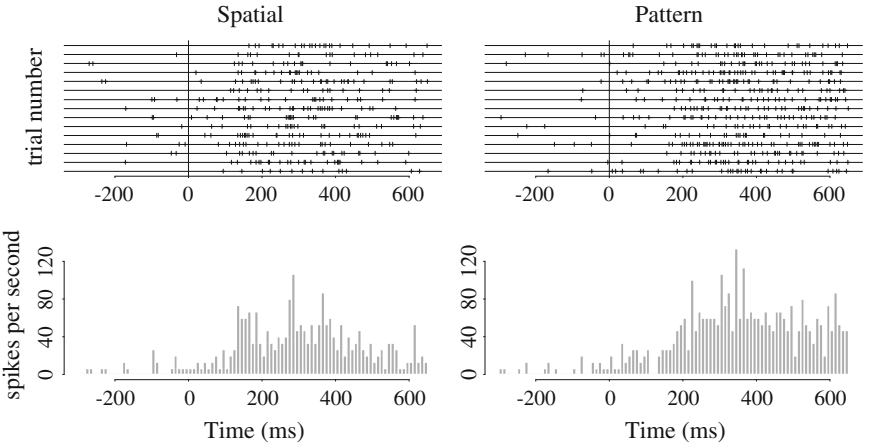

Fig. 1.1 Raster plot (*Top*) and PSTH (*Bottom*) for an SEF neuron under both the external-cue or "spatial" condition (*Left*) and the complex cue or "pattern" condition (*Right*). Each line in each raster plot contains data from a single trial, that is, a single instance in which the condition was applied. (There are 15 trials for each condition.) The tick marks represent spike times, i.e., times at which the neuron fired. The PSTH contains normalized spike counts within 10 ms time bins: for each time bin the number of spikes occurring in that bin is counted across all trials; this count is then divided by the number of trials, and the width of the time bin in seconds, which results in firing rate in units of spikes per second. Time is measured relative to presentation of a visual cue, which is considered time $t = 0$. This neuron is more active under the pattern condition, several hundred milliseconds post cue. The increase in activity may be seen from the raster plots, but is more apparent from comparison of the PSTHs.

Visual comparison of the two raster plots and two PSTHs in Fig. 1.1 indicates that this neuron tends to respond more strongly under the pattern condition than under the spatial condition, at least toward the end of the trial. But such qualitative impressions are often insufficiently convincing even for a single neuron; furthermore, results for many dozens of neurons need to be reported. How should they be summarized? Should the firing rates be averaged over a suitable time interval, and then compared? If so, which interval should be used? Might it be useful to display the firing-rate histograms on top of each other somehow, for better comparison, and might the distinctions between them be quantified and then summarized across all neurons? Might it be useful to compare the peak firing rates for the two neurons, or the time at which the peaks occurred? All of these variations involve different ways to look at the data, and each effectively defines differently the purpose of the study.

The several possible ways of examining firing rate, just mentioned, have in common the aggregation of data across trials. A quite different idea would be to examine the relationship of neural spiking activity and reaction time, on a trial-by-trial

(Footnote 1 continued)
in preparations in vitro, however, identical current inputs to a neuron do not necessarily produce identical spiking outputs. This is due, at least in part, to the stochastic behavior of the movements of ions and molecules that govern the spiking mechanism.

basis, and then to see how that changes across conditions. This intriguing possibility, however, would require a different experiment: in the experiment of Olson et al. the eye movement occurred long after[2] the cue, so there was no observed behavior corresponding to reaction time. This is an extreme case of the way analytical alternatives depend on the purpose of the experiment. □

Example 1.1 illustrates the way a particular purpose shaped the design of the experiment, the way the data were collected, and the possible analytic strategies. In thinking about the way the data are collected, one particular distinction is especially important: that of *steady-state* versus systematically evolving conditions. In many studies, an experimental manipulation leads to a measured response that evolves in a more-or-less predictable way over time. In Example 1.1 the neuronal firing rate, as represented by the PSTH, evolves over time, with the firing rate increasing roughly 200 ms after the cue. This may be contrasted with observation of a phenomenon that has no predictable time trend, experimentally-induced or otherwise. Typically, such situations arise when one is making baseline measurements, in which some indicator of neural activity is observed while the organism or isolated tissue is at rest and receives no experimental stimulus.[3] Sometimes a key piece of laboratory apparatus must be observed in steady state to establish background conditions. Here is an important example.

Example 1.2 MEG background noise Magnetoencephalography (MEG) is an imaging technique used to measure the magnetic fields produced by electrical activity in the brain. MEG recordings are used clinically to localize a brain tumor or to identify the site of an epileptic focus; they are used by neuroscientists to study such things as language production, memory formation, and the neurological basis of diseases such as schizophrenia.

The MEG signals are generated from the net effect of ionic currents flowing in the dendrites of cortical neurons during synaptic transmission. From Maxwell's equations, any electrical current produces a magnetic field oriented orthogonally (perpendicularly) to the current flow, according to the right-hand rule. MEG measures this magnetic field. Magnetic fields are relatively unaffected by the tissue through which the signal passes on the way to a detector, but the signals are very weak. Two things make detection possible. One is that MEG uses highly sensitive detectors called superconducting quantum interference devices (SQUIDs). The second is that currents from many neighboring neuronal dendrites have similar orientations, so that their magnetic fields reinforce each other. The dendrites of pyramidal cells in the cortex are generally perpendicular to its surface and, in many parts of the brain, their generated fields are oriented outward, toward the detectors sitting outside the head.

---

[2] They used a random delay followed by a separate cue to move; this helped ensure that movement and anticipatory effects would not contaminate the processing effects of interest.

[3] Analyses of brain activity when the subject is resting (e.g., during passive eye fixation or with eyes closed) have been reported by many groups. See, for example, Fox et al. (2005), who used fMRI to describe two distinct resting-state networks.

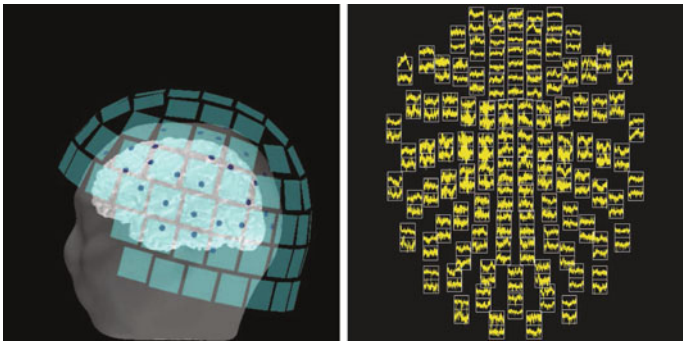

Fig. 1.2 MEG imaging. *Left* drawing of the way the SQUID detectors sit above the head in a MEG machine. *Right* plots of sensor signals laid out in a two-dimensional configuration to correspond, roughly, to their three-dimensional locations as shown in the *left* panel of the figure.

A detectable MEG signal is produced by the net effects of currents from approximately 50,000 active neurons. See Fig. 1.2.

Because the signals are weak, and the detectors extremely sensitive, it is important to assess MEG activity prior to imaging patients. Great pains are taken to remove sources of magnetic fields from the room in which the detector is located. Nonetheless, there remains a background signal that must be identified under steady-state conditions. □

Many analytical methods assume a steady state exists. The mathematical formulation of "steady state," based on *stationarity*, will be discussed in Chapter 18.

1.1.2 Many investigations involve a response to a stimulus or behavior.

In contrast to the steady state conditions in Example 1.2, many experiments involve perturbation or stimulation of a system, producing a temporally evolving response. This does *not* correspond to a steady state. The SEF experiment was a stimulus-response study. Functional imaging also furnishes good examples.

Example 1.3 fMRI in a visuomotor experiment Functional magnetic resonance imaging (fMRI) uses change in magnetic resonance (MR) to infer change in neural activity, within small patches (voxels) of brain tissue. When neurons are active they consume oxygen from the blood, which produces a local increase in blood flow after a delay of several seconds. Oxygen in the blood is bound to hemoglobin, and the magnetic resonance of hemoglobin changes when it is oxygenated. By using an appropriate MR pulse sequence, the change in oxygenation can be detected as the blood-oxygen-level dependent (BOLD) signal, which follows a few seconds after the increase in neural activity. The relationship between neural activity and BOLD is not

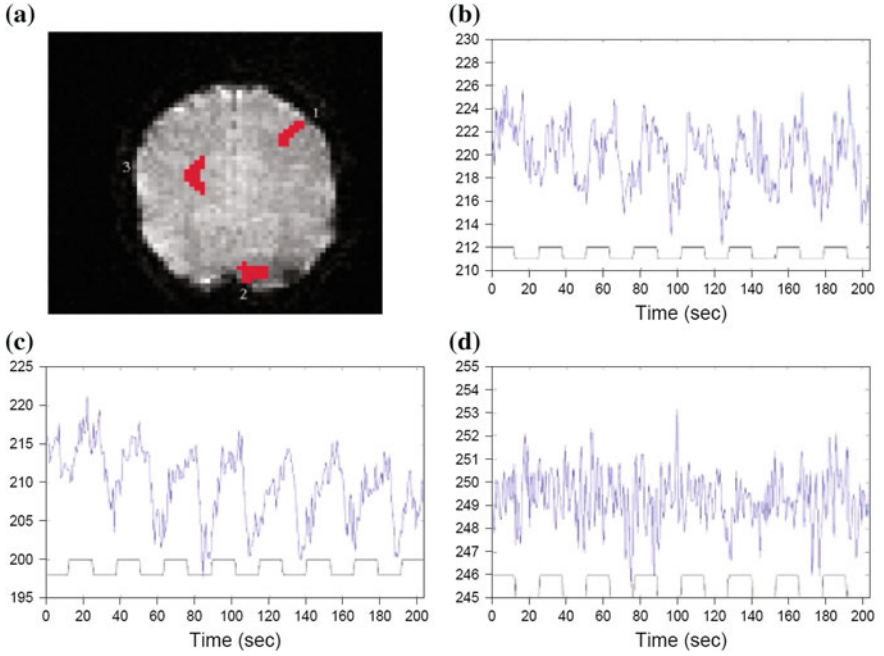

Fig. 1.3 An fMRI image with several traces of the signal across time. Panel A displays an image indicating three locations, shown in *red*, from which voxel signals were examined. Panels B-D display the signals themselves, averaged across the voxels. They correspond, respectively, to motor cortex, primary visual cortex, and white matter.

known in detail, but since the 1990s fMRI has been used to track changes in BOLD in relation to the execution of a task, giving at least a rough guide to the location of sustained functional neural activity.

Figure 1.3 displays images from one subject in a combined visual and motor fMRI experiment. The subject was presented with a full-field flickering checkerboard, in a repeating pattern of 12.8 s (seconds) OFF followed by 12.8 s ON. This was repeated 8 times. Alternating out of phase with the flickering checkerboard pattern the subject also executed a finger tapping task (12.8 s ON followed by 12.8 s OFF). The brain was imaged once every 800 ms for the duration of the experiment. The slice shown was chosen to transect both the visual and motor cortices. Three regions of interest have been selected, corresponding to (1) motor (2) visual cortex, and (3) white matter. Parts B through D of the figure illustrate the raw time series taken from each of these regions, along with timing diagrams of the input stimuli. As expected, the motor region is more active during finger tapping (but the BOLD signal responds several seconds after the tapping activity commences) while the visual region is more active during the flickering visual image (again with several seconds lag). The response within white matter serves as a control. □

The focus of stimulus-response experiments is usually the relationship between stimulus and response. This may suggest strategies for analysis of the data. If we

let X denote the stimulus and Y the response, we might write the relationship as follows:

$$Y \longleftarrow X \tag{1.1}$$

where the arrow indicates that X leads to Y. Chapters 12, 14, and 15 are devoted to regression methods, which are designed for situations in which X might lead to Y.

In Example 1.1, Y could be the average firing rate in a specified window of time, such as 200–600 ms following the cue, and X could represent the experimental condition. In other words, the particular experimental condition leads to a corresponding average firing rate. In Example 1.3, Y could be the value of the BOLD response, and X could represent whether the checkerboard was on or off 5 s prior to the response Y.

The arrow in (1.1) suggests a mechanistic relationship (the stimulus occurred, and that made Y occur), but it is often wise to step back and remain agnostic about a causal connection. A more general notion is that the variables X and Y are *associated*, meaning that they tend to vary together. A wide variety of neuroscientific studies seek to establish associations among variables. Such studies might relate a pair of behavioral measures, for example, or they might involve spike trains from a pair of neurons recorded simultaneously, EEGs from a pair of electrodes on the scalp, or MEG signals from a pair of SQUID detectors. Many statistical tools apply to both causal and non-causal relationships. Measures of association are discussed in Chapters 10 and 12. Chapter 13 also contains a brief discussion of the distinction between association and causation, and some issues to consider when one wishes to infer causation from an observed association.

1.2 The Contribution of Statistics

Many people think of statistics as a collection of particular data-analytic techniques, such as analysis of variance, chi-squared goodness-of-fit, linear regression, etc. And so it is. But the field of statistics, as an academically specialized discipline, strives for something much deeper, namely, the development and characterization of data collection and analysis methods according to well-defined principles, as a means of quantifying knowledge about underlying phenomena and rationalizing the learning and decision-making process. As we said above, one of the main pedagogical goals of this book is to impart to the reader some sense of the way data analytic issues are framed within the discipline of statistics. In trying to achieve this goal, we find it helpful to articulate the nature of the statistical paradigm as concisely as possible. After numerous conversations with colleagues, we have arrived at the conclusion that among many components of the statistical paradigm, summarized below, two are the most fundamental.

Two Fundamental Tenets of the Statistical Paradigm:

1. Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning.
2. Statistical methods may be analyzed to determine how well they are likely to perform.

In the remainder of this section we will elaborate, adding a variety of comments and clarifications.

1.2.1 Statistical models describe regularity and variability of data in terms of probability distributions.

When data are collected, repeatedly, under conditions that are as nearly identical as an investigator can make them, the measured responses nevertheless exhibit variation. The spike trains generated by the SEF neuron in Example 1.1 were collected under experimental conditions that were essentially identical; yet, the spike times, and the number of spikes, varied from trial to trial. The most fundamental principle of the statistical paradigm, its starting point, is that this variation may be described by probability. Chapters 3 and 5 are devoted to spelling out the details, so that it will become clear what we mean when we say that probability describes variation. But the idea is simple enough: probability describes familiar games of chance, such as rolling dice, so when we use probability also to describe variation, we are making an analogy; we do not know all the reasons why one measurement is different than another, so it is *as if* the variation in the data were generated by a gambling device. Let us consider a simple but interesting example.

Example 1.4 Blindsight in patient P.S. Marshall and Halligan (1988) reported an interesting neuropsychological finding from a patient, identified as P.S. This patient was a 49 year-old woman who had suffered damage to her right parietal cortex that reduced her capacity to process visual information coming from the left side of her visual space. For example, she would frequently read words incorrectly by omitting left-most letters ("smile" became "mile") and when asked to copy simple line drawings, she accurately drew the right-hand side of the figures but omitted the left-hand side without any conscious awareness of her error. To show that she could actually see what was on the left but was simply not responding to it—a phenomenon known as *blindsight*—the examiners presented P.S. with a pair of cards showing identical green line drawings of a house, except that on one of the cards bright red flames were depicted on the left side of the house. They presented to P.S. both cards, one above the other (the one placed above being selected at random), and asked her to choose which house she would prefer to live in. She thought this was silly "because they're the same" but when forced to make a response chose the non-burning house on 14 out of 17 trials. This would seem to indicate that she did, in

fact, see the left side of the drawings but was unable to fully process the information. But how convincing is it that she chose the non-burning house on 14 out of 17 trials? Might she have been guessing?

If, instead, P.S. had chosen 17 out of 17 trials there would have been very strong evidence that her processing of the visual information affected her decision-making, while, on the other hand, a choice of 9 out of 17 clearly would have been consistent with guessing. The intermediate outcome 14 out of 17 is of interest as a problem in data analysis and scientific inference precisely because it feels fairly convincing, but leaves us unsure: a thorough, quantitative analysis of the uncertainty would be very helpful.

The standard way to begin is to recognize the variability in the data, namely, that P.S. did not make the same choice on every trial; we then say that the choice made by P.S. on each trial was a random event, that the probability of her choosing the non-burning house on each trial was a value p, and that the responses on the different trials were independent of each other. These three assumptions use probability to describe the variability in the data. Once these three assumptions are made it becomes possible to quantify the uncertainty about p and the extent to which the data are inconsistent with the value $p = .5$, which would correspond to guessing. In other words, it becomes possible to make statistical inferences. □

The key step in Example 1.4 is the introduction of probability to describe variation. Once that first step is taken, the second step of making inferences about the phenomenon becomes possible. Because the inferences are statistical in nature, and they require the introduction of probability, we usually refer to the probability framework—with its accompanying assumptions—as a *statistical model.* Statistical models provide a simple formalism for describing the way the repeatable, regular features of the data are combined with the variable features. In Example 1.4 we may think of p as the propensity for P.S. to choose the non-burning house. According to this statistical model, p is a kind of regularity in the data in the sense that it is unchanging across trials. The variation in the data comes from the probabilistic nature of the choice: what P.S. will choose is somewhat unpredictable, so we attribute a degree of uncertainty to unknown causes and describe it as if predicting her choice were a game of chance. We elaborate on the statistical model, and the inferences drawn from the data of Example 1.4 in Chapters 5 and 7.

Probability is also often introduced to describe small fluctuations around a specified formula or "law." We typically consider such fluctuations "noise," in contrast to the systematic part of the variation in some data, which we call the "signal." For instance, as we explain in Chapter 12, when the underlying, systematic mathematical specification (the signal) has the form

$$y = f(x)$$

we will replace it with a statistical model having the form

$$Y = f(x) + \epsilon \tag{1.2}$$

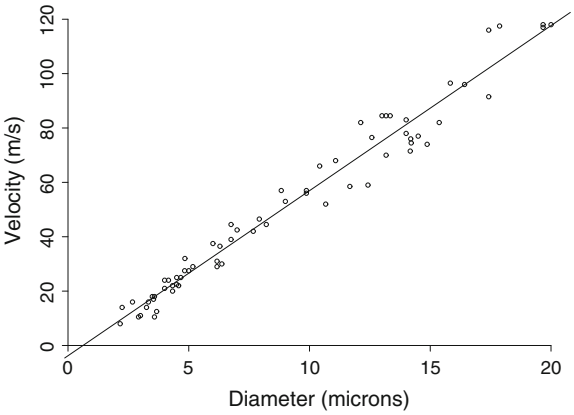

Fig. 1.4 Conduction velocity of action potentials, as a function of diameter. The x-axis is diameter in microns; the y-axis is velocity in meters per second. Based on Hursh (1939, Fig. 2). Also shown is the least-squares regression line.

where ϵ represents noise and the variable Y is capitalized to indicate its now-random nature: it becomes "signal plus noise." The simplest case occurs when $f(x)$ is a line, having the form $f(x) = \beta_0 + \beta_1 x$, where we use coefficients β_0 and β_1 (instead of writing $f(x) = a + bx$) to conform to statistical convention. Here is an example.

Example 1.5 Neural conduction velocity Hursh (1939) presented data on the relationship between a neuron's conduction velocity and its axonal diameter, in adult cats. Hursh measured maximal velocity among fibers in several nerve bundles, and then also measured the diameter of the largest fiber in the bundle. The resulting data, together with a fitted line, are shown in Fig. 1.4. In this case the line $y = \beta_0 + \beta_1 x$ represents the approximate linear relationship between maximal velocity y and diameter x. The data follow the line pretty closely, with the intercept β_0 being nearly equal to zero. This implies, for example, that if one fiber has twice the diameter of another, the first will propagate an action potential roughly twice as fast as the second. □

Before we conclude our introductory remarks about statistical models, by elaborating on (1.2), let us digress for a moment to discuss the method used to fit the line to the data in Fig. 1.4, which is called *least squares regression*, It is one of the core conceptions of statistics, and we discuss it at length in Chapter 12.

Suppose we have a line that is fit by some method, possibly least-squares or possibly another method, and let us write this line as $y = \beta_0^* + \beta_1^* x$. It is customary, in statistics, to use the notations β_0 and β_1 for the intercept and slope. Here we have included the asterisk * in β_0^* and β_1^* because it will simplify some additional notations later on. The important thing is that β_0^* and β_1^* are coefficients that define the line we fit to the data, using whatever method we might choose. Suppose there are n data pairs of the form (x, y) and let us label them with a subscript so that they take the form (x_i, y_i) with $i = 1, 2, \ldots, n$. That is, (x_1, y_1) would be the first data

pair, (x_2, y_2) the second, and so forth. The y-coordinate on the line $y = \beta_0^* + \beta_1^* x$ corresponding to x_i is

$$\hat{y}_i^* = \beta_0^* + \beta_1^* x_i.$$

The number $\hat{y}_i^*$ is called the *fitted value* at x_i and we may think of it as predicting y_i. We then define the ith *residual* as

$$e_i = y_i - \hat{y}_i^*.$$

The value e_i is the error at x_i in fitting, or the error of prediction, i.e., it is the vertical distance between the observation (x_i, y_i) and the line at x_i. We wish to find the line that best predicts the y_i values, which means we want to make the e_i's as small as possible, in aggregate. To do this, we have to minimize some measure of the size of all the e_i's taken together. In choosing such a measure we assume positive and negative values of the residuals are equally important. Two alternative aggregate measures that treat e_i and $-e_i$ equally are the following:

$$\begin{aligned} \text{sum of absolute deviations} &= \sum_{i=1}^{n} |e_i| \\ \text{sum of squares} &= \sum_{i=1}^{n} e_i^2. \end{aligned} \tag{1.3}$$

Data analysts sometimes choose β_0^* and β_1^* to minimize the sum of absolute deviations, but the solution can not be obtained in closed form, and it is harder to analyze mathematically. Instead, the method of least squares works with the sum of squares, where the solution may be found using calculus (see Chapter 12).

The least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values of β_0^* and β_1^* that minimize the sum of squares in (1.3). The least-squares line is then

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Having motivated least-squares with (1.2) let us return to that equation and note that it is not yet a statistical model. If we write

$$Y_i = f(x_i) + \epsilon_i, \tag{1.4}$$

take

$$f(x) = \beta_0 + \beta_1 x$$

and, crucially, let the noise term ϵ_i be a *random variable*, then we obtain a *linear regression model*. Random variables are introduced in Chapter 3. The key point in

the present discussion is that linear regression describes the regularity of the data by a straight line and the variability (the deviations from the line) by a probability distribution (the distribution of the noise random variable ϵ_i).

1.2.2 Statistical models are used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning.

The introduction of a statistical model not only provides guidance in determining fits to data, as in Example 1.5, but also assessments of uncertainty.

Example 1.4 (continued from page 9) Let us return to the question of whether the responses of P.S. were consistent with guessing. In this framework, guessing would correspond to $p = .5$ and the problem then becomes one of assessing what these data tell us about the value of p. As we will see in Chapter 7, standard statistical methods give an approximate 95 % confidence interval for p of (.64, 1.0). This is usually interpreted by saying we are 95 % confident the value of p lies in the interval (.64, 1.0), which is a satisfying result: while this interval contains a range of values, indicating considerable uncertainty, we are nonetheless highly confident that the value of p is inconsistent with guessing. □

The confidence interval we have just reported in Example 1.4 illustrates the expression of "knowledge and uncertainty." It is an example of *inductive reasoning* in the sense that we reason from the data back to the quantity p assumed in the model. Many mathematical arguments begin with a set of assumptions and *prove* some consequence. This is often called *deductive reasoning*. As described in Chapter 7, statistical theory uses deductive reasoning to provide the formalism for confidence intervals. However, when we interpret the result as providing *knowledge* about the unknown quantity p based on experience (the data), the argument is usually called "inductive." Unlike deductive reasoning, inductive reasoning is uncertain. We use probability to calibrate the degree to which a statement is likely to be true. In reporting confidence intervals, the convention is to use a probability of .95, representing a high degree of confidence.

In fact, as a conceptual advance, this expression of knowledge and uncertainty via probability is highly nontrivial: despite quite a bit of earlier mathematical attention to games of chance, it was not until the late 1700s that there emerged any clear notion of inductive, or what we would now call *statistical* reasoning; it was not until the first half of the twentieth century that modern methods began to be developed systematically; and it was only in the second half of the twentieth century that their properties were fully understood. From a contemporary perspective the key point is that the confidence interval is achieved by uniting two distinct uses of probability. The first is descriptive: saying P.S. will choose the non-burning house with probability p is analogous to saying the probability of rolling an even number with an apparently fair six-sided die is $1/2$. The second use of probability is often called "epistemic,"

and involves a statement of knowledge: saying we have 95 % confidence that p is in the interval (.64, 1.0) is analogous to someone saying they are 90 % sure that the capital of Louisiana is Baton Rouge. The fundamental insight, gained gradually over many years, is that the descriptive probability in statistical models may be used to produce epistemic statements for scientific inference. We will emphasize the contrast between the descriptive and epistemic roles of statistical models by saying that models describe the *variation* in data and produce *uncertain* inferences. Technically, there are alternative frameworks for bringing descriptive and epistemic probability together, the two principal ones being *Bayesian* and *frequentist*. We will discuss the distinction in Section 7.3.9, and develop the Bayesian approach to inference at greater length in Chapter 16.

While we wish to stress the importance of statistical models in data analysis, we also want to issue several qualifications and caveats: first, the notion of "model" we intend here is very general, the only restriction being that it must involve a probabilistic description of the data; second, modeling is done in conjunction with summaries and displays that do not introduce probability explicitly; third, it is very important to assess the fit of a model to a given set of data; and, finally, statistical models are mathematical abstractions, imposing structure on the data by introducing explicit assumptions. The next three subsections explain these points further.

1.2.3 Statistical models may be either parametric or nonparametric.

In emphasizing statistical models, our only restriction is that probability must be used to express the way regularity and variability in the data are to be understood. One very important distinction is that of *parametric* versus *nonparametric* models.

The terminology comes from the representation of a probability distribution in terms of an unknown parameter. A *parameter* is a number, or vector of numbers, that is used in the definition of the distribution; the probability distribution is characterized by the parameter in the sense that once the value of the parameter is known, the probability distribution is completely determined. In Example 1.4, p. 9, the parameter is p. In Example 1.5, p. 11, the parameter includes the pair (β_0, β_1), together with a noise variation parameter σ, explained in Chapter 12. In both of these cases the values of the unknown parameters determine the probability distribution of the random variables, as in (1.4). Parametric probability distributions are discussed in Chapter 5.

A related distinction arises in the context of y versus x models of the type considered in Example 1.5. That example involved a linear relationship. As we note in Chapters 14 and 15, the methods used to fit linear models can be generalized for nonlinear relationships. The methods in Chapter 15 are also called nonparametric because the fitted relationship is not required to follow a pre-specified form.

Example 1.6 Excitatory post-synaptic current As part of a study on spike-timing-dependent plasticity (Dr. David Nauen, personal communication), rat hippocampal

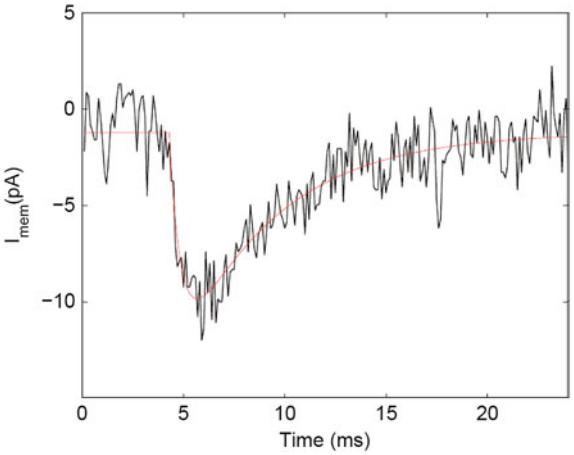

Fig. 1.5 Excitatory post-synaptic current. Current recorded from a rat hippocampal neuron, together with smoothed version (shown as the *thin red line* within the noisy current trace) obtained by fitting a suitable function of time, given in the text. The current values are connected by the *dark line*. When values recorded sequentially in time are plotted it is a common practice to connect them. (Figure courtesy of David Nauen.)

neurons were held in voltage clamp and post-synaptic currents were recorded following an action potential evoked in a presynaptic cell. Figure 1.5 displays a plot of membrane current as a function of time. One measurement of size of the current is found by integrating the current across time (which is implemented by summing the current values and multiplying by the time between observations), giving the total charge transmitted. Other quantities of interest include the onset delay, the rate at which the curve "rises" (here, a negative rise) from onset to peak current, and the rate at which the curve decays from peak current back toward steady state. The current trace is clearly subject to measurement noise, which would contaminate the calculations. A standard way to reduce the noise is to fit the data by a suitable function of time. Such a fit is also shown in the figure. It may be used to produce values for the various constants needed in the analysis. To produce the fit a statistical model of the form (1.4) was used where the function $y = f(x)$, with y being post-synaptic current and x being time, was defined as

$$f(x) = A_1(1-\exp((x-t_0)/\tau_1))\,(A_2\exp((x-t_0)/\tau_2)-(1-A_2)\exp((x-t_0)/\tau_3)).$$

This was based on a suggestion by Nielsen et al. (2004). Least squares was then applied, as defined in Section 1.2.1. The fit is good, though it distorts slightly the current trace in the dip and at the end. The advantage of using this function is that its coefficients may be interpreted and compared across experimental conditions. □

The simple linear fit in Example 1.5, p. 11, is an example of linear regression, discussed in Chapter 12, while the fit based on a combination of exponential functions

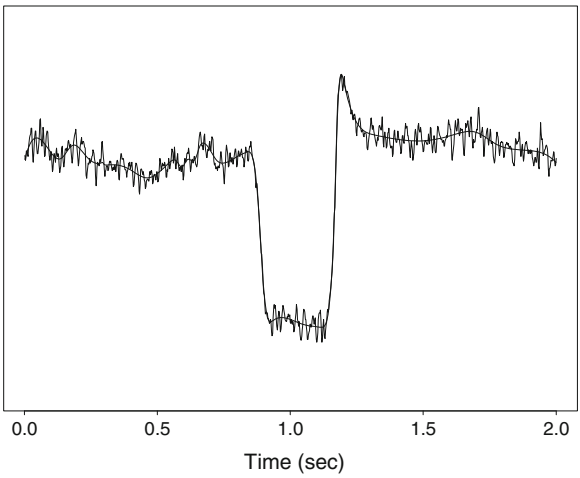

Fig. 1.6 Electrooculogram together with a smoothed, or "filtered" version that removes the noise. The method used for smoothing is an example of nonparametric regression.

in Example 1.6 is an example of *nonlinear regression* discussed in Section 14.2. Both are examples of *parametric regression* because both use specified functions based on formulas that involve a few parameters. In Example 1.5 the parameters were β_0 and β_1 while in Example 1.6 they were A_1, A_2, τ_1, τ_2, τ_3, t_0. *Nonparametric regression* is used when the formula for the function is not needed. Nonparametric regression is a central topic of Chapter 15. Here is an example.

Example 1.7 Electrooculogram smoothing for EEG artifact removal EEG recordings suffer from a variety of artifacts, one of which is their response to eye blinks. A good way to correct for eye-blink artifacts is to record potentials from additional leads in the vicinity of the eyes; such electrooculograms (EOGs) may be used to identify eye blinks, and remove their effects from the EEGs. Wallstrom et al. (2002, 2004) investigated methods for removing ocular artifacts from EEGs using the EOG signals. In Chapter 15 it will become clear how to use a general smoothing method to remove high-frequency noise. This does not require the use of a function having a specified form. Figure 1.6 displays an EOG recording together with a smoothed version of it, obtained using a nonparametric regression method known as BARS (Dimatteo et al. 2001). □

1.2.4 Statistical model building is an iterative process that incorporates assessment of fit and is preceded by exploratory analysis.

Another general point about the statistical paradigm is illustrated in Fig. 1.7. This figure shows where the statistical work fits in. Real investigations are far less sequential than depicted here, but the figure does provide a way of emphasizing two components of the process that go hand-in-hand with statistical modeling: exploratory analysis and assessment of fit. Exploratory analysis involves informal investigation of the data based on numerical or graphical summaries, such as a histogram. Exploratory results, together with judgment based on experience, help guide construction of an initial probability model to represent variability in observed data.

Every such model, and every statistical method, makes some assumptions, leading, as we have already seen, to a reduction of the data in terms of some small number of interpretable quantities. As shown in Fig. 1.7, the data may be used, again, to check the probabilistic assumptions, and to consider ramifications of departures from them. Should serious departures from the assumptions be found, a new model may be formed. Thus, probability modeling and model assessment are iterative, and only when a model is considered adequate are statistical inferences made. This process is embedded into the production of scientific conclusions from experimental results (Box et al. 1978).

1.2.5 All models are wrong, but some are useful.

The simple representation in Fig. 1.7 is incomplete and may be somewhat misleading. Most importantly, while it is true that there are standard procedures for model assessment, some of which we will discuss in Chapter 10, there is no uniformly-applicable rule for what constitutes a good fit. Statistical models, like scientific models, are abstractions and should not be considered perfect representations of the data. As examples of scientific models in neuroscience we might pick, at one extreme, the Hodgkin-Huxley model for action potential generation in the squid giant axon and, at the other extreme, being much more vague, the theory that vision is created via separate ventral and dorsal streams corresponding loosely to "what" and "where." Neither model is perfectly accurate—in fact, every scientific model fails[4] under certain conditions. Models are helpful because they capture important intuitions and can lead to specific predictions and inferences. The same is true of statistical models. On the other hand, statistical models are often driven primarily by raw empiricism—they are produced to fit data and may have little or no other justification or explanatory power. Thus, experienced data analysts carry with them a strong sense of both the

[4] For a discussion of some ways that great equations of physics remain fundamental while only approximating the real world, see Weinberg (2002). An entry into the philosophical literature on statistical inference and modeling is Mayo and Spanos (2010).

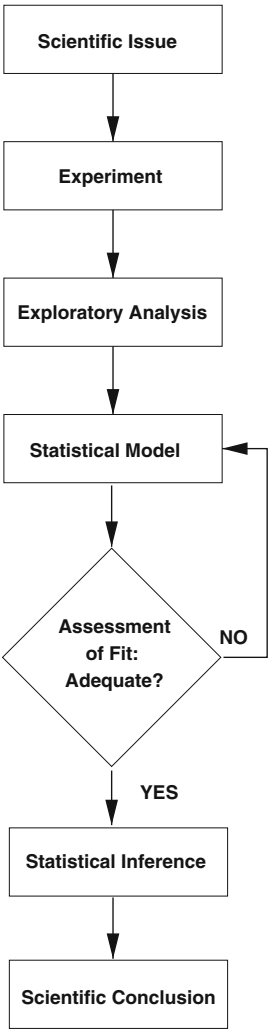

Fig. 1.7 Formal statistical inference within the process of drawing scientific conclusions. Statistical model building is a prerequisite to formal inference procedures. Model building is iterative in the sense that tentative models must be assessed and, if necessary, improved or abandoned. The figure is something of a caricature because the process is not as neat as depicted here. Furthermore, there are typically multiple aspects of the data, which bear on several different issues. A single scientific conclusion may rely on many distinct statistical inferences.

inaccuracies in statistical models and their lingering utility. This sentiment is captured well by the famous quote from George Box, "All models are wrong, but some are useful" (Box 1979).

To emphasize further the status of statistical models we have created Fig. 1.8. Pictured in the left column is the "real world" of data, i.e., the observables, obtained by recording in some form, often by measurement. In the right column is the

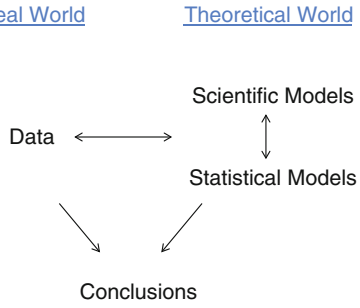

Fig. 1.8 The role of statistical models and methods in scientific inference. Statistical procedures are abstractly defined in terms of mathematics, but are used, in conjunction with scientific models and methods, to explain observable phenomena. Adapted from Kass (2011).

"theoretical world" where both scientific and statistical models live. Scientific models help us organize facts with explanations. They can be high-level or detailed, but they should not, at least in principle, be confused with the observations themselves. The theoretical world seeks to make statements and predictions, often using a precise but abstract mathematical framework, which may be applied to things in the real world that may be observed. In a domain where theory works well, the theoretical world would be judged to be very close to the real world and, therefore, its predictions would be highly trustworthy. Statistical models are used to describe the imperfect predictability of phenomena, the regularity and variability of data, in terms of probability distributions.

A second aspect of the flow diagram in Fig. 1.7 may be misleading. The diagram fails to highlight the way the judgment of adequate fit depends on context. When we say "All models are wrong, but some are useful," part of the point is that a model can be useful *for a specified inferential purpose*. Thus, in judging adequacy of a model, one must ask, "How might the reasonably likely departures from this model affect scientific conclusions?"

We illustrate the way statistical models lead to scientific conclusions in numerous examples throughout this book.

1.2.6 Statistical theory is used to understand the behavior of statistical procedures under various probabilistic assumptions.

The second of the two major components of the statistical paradigm is that methods may be *analyzed* to determine how well they are likely to perform. As we describe briefly in Sections 4.3.4 and 7.3.9, and more fully in Chapters 8 and 11, a series of general principles and criteria are widely used for this purpose. Statistical theory has been able to establish good performance of particular methods under certain probabilistic assumptions. In Chapters 3–6 we provide the necessary background for

the theory we develop. When we wish to add arguments that are not essential to the flow of material we highlight them as *details* and indent them, as follows.

> *Details*: We indent, like this, the paragraphs containing mathematical details we feel may be safely skipped. □

One easy and useful method of checking the effectiveness of a procedure, which is applicable in certain predictive settings, is *cross-validation*. The simplest form of cross-validation involves splitting the data set into two subsets, applying and refining a method using one of the subsets, and then judging its predictive performance (predicting the value of some response) on the second subset. Sometimes the second subset involves entirely new data. For example, in a behavioral study, a new set of subjects may be recruited and examined. Methods that perform well with this kind of cross-validation are often quite compelling. In addition to being intuitive, cross-validation has a theoretical justification discussed briefly in Chapter 12.

1.2.7 Important data analytic ideas are sometimes implemented in many different ways.

The usual starting point in books about data analysis is measures of central tendency, such as mean and median, which we review in Section 2.1.1. There are three reasons for putting a discussion of central tendency at the beginning. First, the use of a single representative value (such as the mean) to summarize a bunch of numbers is ubiquitous. Second, it is an excellent example of the process of data summary; data analysis as a whole may be considered a kind of generalization of this simple method. Third, the mean and median are both single-number summaries but they behave very differently. This last point, that it matters how a general data analytic idea (a single-number summary of central tendency) is defined (mean or median), has become ingrained into teaching about statistical reasoning. The crucial observation[5] is that it can be important to separate the general idea from any specific implementation; as a useful concept, the general idea may transcend any specific definition. For example, in Section 4.3.2 we discuss the deep notion that information represents reduction of uncertainty. As we explain there, the general idea of information could be defined, technically, in terms of a quantity called *mutual information*, but it could also be defined using the squared correlation. Mutual information and squared correlation have very different properties. The definition matters, but with either definition we can think of information as producing a reduction of uncertainty.

1.2.8 Measuring devices often pre-process the data.

Measurements of neural signals are often degraded by noise. A variety of techniques are used to reduce the noise and increase the relative strength of the signal, some

[5] This point was emphasized by Mosteller and Tukey (1977, Section 1F).

of which will be discussed in Chapter 7. In many cases, methods such as these are applied by the measurement software to produce the data the investigator will analyze. For example, fMRI data are acquired in terms of frequency and software is used to reconstruct a signal in time; MEG sensors must be adjusted to ensure detection above background noise; and extracellular electrode signals are thresholded and filtered to isolate action potentials, which then must be "sorted" to identify those from particular neurons. In each of these cases the data that are to be analyzed are not in the rawest form possible. Such pre-processing may be extremely useful, but its effects are not necessarily benign. Inaccurate spike sorting, for example, is a notorious source of problems in some contexts. (See Bar-Gad et al. (2001) and Wood et al. (2004).) The wise analyst will be aware of possible distortions that might arise before the data have been examined.

1.2.9 Data analytic techniques are rarely able to compensate for deficiencies in data collection.

A common misconception is that flaws in experimental design, or data collection, can be fixed by statistical methods after the fact. It is true that an alternative data analytic technique may be able to be help avoid some presumed difficulty an analyst may face in trying to apply a particular method—especially when associated with a particular piece of software. But when a measured variable does not properly capture the phenomenon it is supposed to be measuring, post hoc manipulation will be almost never be able to rectify the situation; in the rare cases that it can, much effort and very strong assumptions will typically be required. For example, we already mentioned that inaccurate spike sorting can create severe problems. When these problems arise, no post-hoc statistical manipulation is likely to fix them.

1.2.10 Simple methods are essential.

Another basic point concerning analytical methods is that simple, easily-understood data summaries, particularly visual summaries such as the PSTH, are essential components of analysis. These fit into the diagram of Fig. 1.7 mainly under the heading of exploratory data analysis, though sometimes inferential analyses from simple models are also used in conjunction with those from much more elaborate models. When a complicated data-analytic procedure is applied, it is important to understand the way results agree, or disagree, with those obtained from simpler methods.

1.2.11 It is convenient to classify data into several broad types.

When spike train data, like those in Example 1.1, are summarized by spike counts occurring in particular time intervals, the values taken by the counts are necessarily

non-negative integers. Because the integers are separated from each other, such data are called *discrete*. On the other hand, many recordings, such as MEG signals, or EEGs, can take on essentially all possible values within some range—subject only to the accuracy of the recording instrument. These data are called *continuous*. This is a very important distinction because specialized analytical methods have been developed to work with each kind of data.

Count data form an important subclass within the general category of discrete data. Within count data, a further special case occurs when the only possible counts are 0 or 1. These are *binary* data. The key characterization is that there are only two possible values; it is a matter of analytical convenience to consider the two values to be 0 or 1. As an example, the response of patient P.S. on each trial was binary. By taking the response "non-burning house" to be 1 and "burning house" to be zero, we are able to add up all the coded values (the 1 and 0 s) to get the total number of times P.S. chose the non-burning house. This summation process is easy to deal with mathematically. A set of binary data would almost always be assumed to consist of 0 and 1 s.

Two other kinds of data arising in neuroscience deserve special mention here. They are called *time series* and *point processes*. Both involve sequential observations made across time. MEG signals, EEGs, and LFPs are good examples of time series: at each of many successive points in time, a measurement is recorded. Spike trains are good examples of point processes: neuronal action potentials are recorded as sequences of event times. In each case, the crucial fact is that an observation at time t_1 is related to an observation made at time t_2 whenever t_1 and t_2 are close to each other. Because of this temporal relationship time series and point process data must be analyzed with specialized methods. Statistical methods for analyzing time series and point processes are discussed in Chapters 18 and 19.