See attached correction for page 61

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, electroenc~~h~~ephalography (producing EEGs), and functional imaging—positron emission tomography(PET), functional magnetic 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

electroenceephalography

remove h

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

indispensable

a

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

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 reptitions designed to be as close to identical as possible.

repetitions
e

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

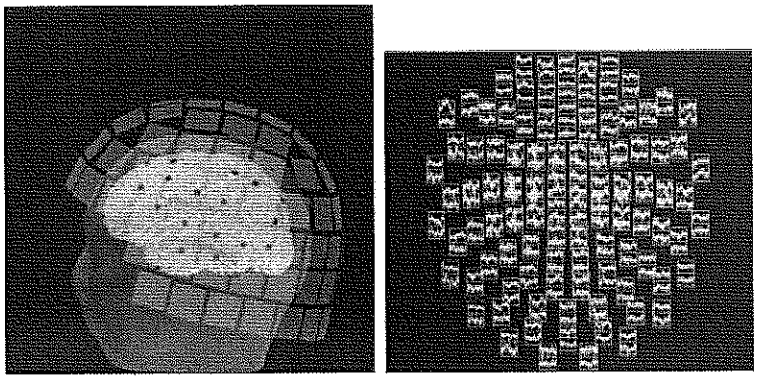


Fig. 1.2 MEG imaging. *Left* drawing of the way the SQUID detectors sit above the head in a MEG machine. *Right* plots of sensor signals laid out in a two-dimensional configuration to correspond, roughly, to their three-dimensional locations as shown in the *left* panel of the figure.

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.*

system

s

In contrast to the steady state conditions in Example 1.2, many experiments involve perturbation or stimulation of a sy^tem, 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 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 i-th *residual* as

ith

italic i

remove hyphen

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

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:

$$\text{sum of absolute deviations} = \sum_{i=1}^{n} |e_i|$$

$$\text{sum of squares} = \sum_{i=1}^{n} e_i^2. \tag{1.3}$$

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

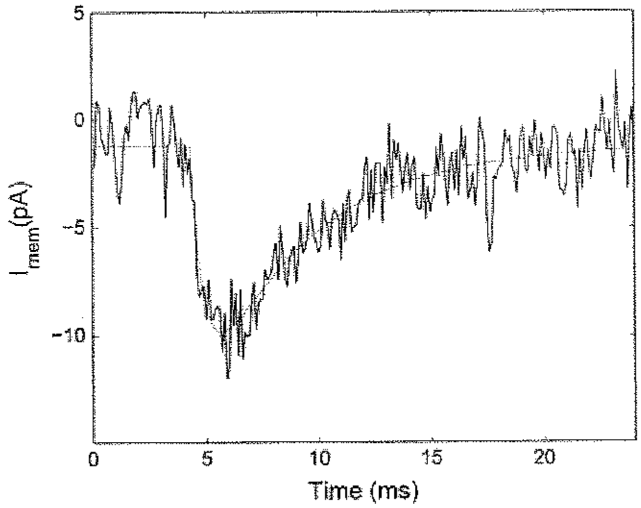
Having motivated least-squares with (1.2) let us return to that equation and note that it is not yet a statistical model. If we write

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

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



hippocampal
?

Fig. 1.5 Excitatory post-synaptic current. Current recorded from a rat hippocamal neuron, together with smoothed version (shown as the *thin 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 signal) at each frequency, for each time bin, indicated in the figure by three different colors representing low, medium, and high power. In Fig. 2.2 the most easily visible oscillations are the alpha rhythm (roughly 8–13 Hz) in the second half of the EEG trace in the awake phase (when the eyes are closed) and the delta rhythm (below 4 Hz) during the surgical phase. Precise scientific statements often require statistical inferences (indications of uncertainy or tests of hypotheses), but spectrograms are very useful in displaying time-frequency information even without formal inferential assessments. □

uncertainty
t

2.2 Data Transformations

2.2.1 Positive values are often transformed by logarithms.

Measurement scales arise from convenience, and need not be considered in any way absolute or immutable; changing the scale often produces a more elegant description. A canonical example involves the acidity of a dilute aqueous solution, which is determined by the concentration of hydrogen ions. The larger the concentration $[H^+]$ of hydrogen ions, the more acidity. Rather than using $[H^+]$ to measure acidity, we use its logarithm, which is known as pH. Specifically, $pH = -\log_{10}([H^+])$, so that an increase in $[H^+]$ corresponds to a decrease in pH. Because the defining property of the logarithm is

$$\log ab = \log a + \log b, \tag{2.1}$$

log transformations are used when multiplicative effects seem more natural than additive. In the case of pH, a solution having a hydrogen ion concentration of 10^{-5} mol l^{-1} is 1 unit greater pH (less acidic) than a solution having a concentration of 10^{-4} mol l^{-1}. Similarly, a solution having a hydrogen ion concentration of 10^{-9} mol l^{-1} is 1 unit greater pH than a solution having a concentration of 10^{-8} mol l^{-1}. In both cases, a 1 unit increase in pH corresponds to a 10-fold decrease in hydrogen ion concentration, regardless of the concentration we started with. In chemical calculations, the log concentration scale is simpler to work with than the concentration scale.

Many other familiar scales are logarithmic. One example is the use of decibels to measure the strength of an auditory signal. Not only are log scales familiar and intuitive but, in addition, some batches of data look more nearly like observations from a normal distribution following a log transformation. In particular, it frequently happens that a batch of data look highly skewed in a given measurement scale, but are much closer to being symmetric in the log scale.

Example 2.1 (continued from p. 24) Figure 2.3 displays the saccadic reaction time data in both the original scale and the log transformed scale. To transform the data to the log scale we have replaced x = reaction time by $\log_{10}(x)$ for each of the 119

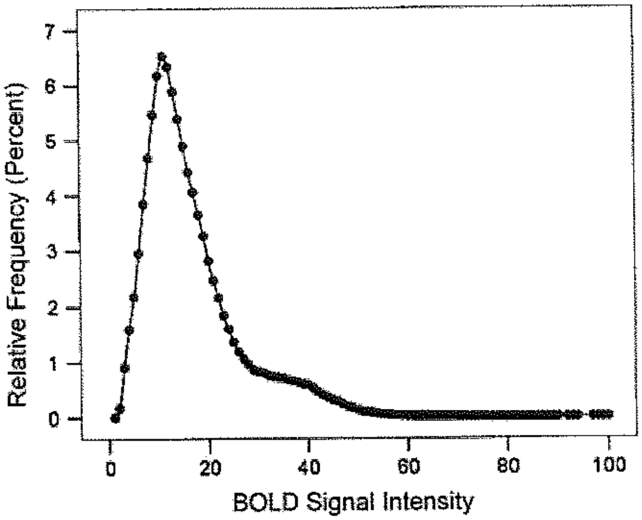


Fig. 2.4 High-field of BOLD signal intensities. The frequencies are plotted as *dots*, rather than bin heights. The distribution across voxels is skewed toward high values. Reprinted with permission from Lewis et al. (2005).

symmetrical. Presumably, this has to do with effects of the Central Limit Theorem. We will discuss this great theorem in Chapter 6. For now let us be content to state it this way: if we add up many small, independent effects their sum will be approximately normally distributed. The empirical observation of approximate normality may then be interpreted as follows: *if we choose the right scale*, the data values may be considered sums of many small, independent effects.

We can understand this a little more deeply by returning to the logarithmic relationship in Eq. (2.1), and considering the role it may play when many small effects are combined to produce variability. The cases where the log transformation is valuable are those where it is natural to think in terms of proportionality. So suppose the reason that two measurements are different is that many small *proportional* effects, of somewhat different sizes in the two measurements, have been combined. For example, the length of a dendritic spine may depend on contributions to the cell membrane and its contents by vast numbers of lipid and protein molecules. If we break the growth process into many thousands of pieces, each might be considered a small effect, so that the net result is a composition of many, many small effects. When we see that one spine is longer than another, we might imagine that the many small effects in the longer spine tended to be *proportionally* larger than those in the shorter spine. Now consider two such small growth effects x_1 and x_2, occurring, respectively, in the shorter and longer dendrites. If we think of the variation as proportional, we may relate the values x_1 and x_2 by writing $x_2 = x_1(1 + \epsilon)$, where ϵ is a small number representing the proportional change (e.g., $\epsilon = .05$, or 5 %) in going from x_1 to x_2. From Eq. (2.1) together with a little calculus, for small ϵ we have $\log(1 + \epsilon) \approx \epsilon$ (see Section ~~19.4~~ A of the Appendix). We then have A.4

Section A.4

If we let $\cup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$ then Axiom 3 may be written instead in the form

3. If $A_1, A_2, \ldots,$ are mutually exclusive events, then $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.

A technical point is that in advanced texts, Axiom 3 would instead involve infinitely many events, and an infinite sum:
3′. If $A_1, A_2, \ldots,$ are infinitely many mutually exclusive events, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Regardless of whether one worries about the possibility of infinitely many events, it is easy to deduce from the axioms the elementary properties we need.

Theorem: Three Properties of Probability For any events A and B we have

(i) $P(A^c) = 1 - P(A)$, where A^c is the complement of A.
(ii) If A and B are mutually exclusive, $P(A \cap B) = 0$.
(iii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: To prove (i) we simply note that $\Omega = A \cup A^c$. From axiom (2) we then have $P(A \cup A^c) = 1$ and because A and A^c are mutually exclusive axiom (3) gives $P(A) + P(A^c) = 1$, which is the same as (i). It is similarly easy to prove (ii) and (iii). □

To illustrate, suppose we pick at random a playing card from a standard 52-card deck. We may compute the probability of drawing a spade or a face card, meaning either a spade that is not a face card, or a face card that is not a spade, or a face card that is also a spade. We take A to be the event that we draw a spade and B to be the event that we draw a face card. Then, because there are 3 face cards that are spades we have $P(A \cap B) = \frac{3}{52}$, and, appying the last formula above, we get $P(A \cup B) = \frac{1}{4} + \frac{3}{13} - \frac{3}{52} = \frac{11}{26}$. This matches a simple enumeration argument: there are 13 spades and 9 non-spade face cards, for a total of 22 cards that are either a spade or a face card, i.e., $P(A \cup B) = \frac{22}{52} = \frac{11}{26}$. The main virtue of such formulas is that they also apply to contexts where probabilities are determined without reference to a decomposition into equally-likely sub-components.

applying
l

Example 3.1 (continued from p. 38) From 1,200 replications of the 100 ms stimulus Kelly calculated the probability that the first neuron would fire at least once was $P(A) = .13$ and the probability that the second neuron would fire at least once was $P(B) = .22$, while the probability that both would fire at least once was $P(A \cap B) = .063$. Applying the formula for the union (property (iii) above), the probability that at least one neuron will fire is $P(A \cup B) = .13 + .22 - .063 = .287$. □

$$P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A).$$

probability
ℓ

Multiplication of probabilities should be very familiar. If a coin has probability .5 of coming up heads when flipped, then we usually say the probabiilty of getting two heads is .25 = .5 × .5, because we usually assume that the two flips are independent.

Example 3.1 (continued from p. 39) For the probabilities $P(A)$, $P(B)$ given on p. 39 we have $P(A)P(B) = .029$ while the probability of the intersection was reported to be $P(A \cap B) = .063$. The latter is more than double the product $P(A)P(B)$. We conclude that the two neurons are not independent. Their tendency to fire much more often together than they would if they were independent could be due to their being connected, to their having similar response properties, or to their both being driven by network fluctuations (see also Kelly et al. (2010)). □

The definition of independence extends immediately to more than two events: if $A_1, A_2, \ldots, A_n$ are independent then

$$P(\cap_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

where $\cap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n$.

Independence is extremely useful. Without it, dependencies represented by conditional probabilities can become very complicated. Independence simplifies calculations and is often assumed in statistical models and methods. On the other hand, as illustrated in Example 3.1, above, if the assumption of independence is wrong, the calculations can be way off: in Example 3.1 the probability $P(A \cap B)$ predicted by independence would be too small by a factor of more than 2. In many situations independence is the most consequential statistical assumption, and therefore must be considered carefully.

3.1.4 Bayes' theorem for events gives the conditional probability P(A|B) in terms of the conditional probability P(B|A).

Bayes' theorem is a very simple identity, which we derive easily below. Yet, it has profound consequences. We can state its purpose formally, without regard to its applications: Bayes' theorem allows us to compute $P(A|B)$ from the reverse conditional probability $P(B|A)$, if we also know $P(A)$. As we will see below, and in Chapter 16, there are more complicated versions of the theorem, and it is especially those that produce the wide range of applications. But the power of the result becomes apparent immediately when we take B to be some data and A to be a scientific hypothesis. In this case, we can use the probability $P(\text{data}|\text{hypothesis})$ from the statistical model to obtain the scientific inference $P(\text{hypothesis}|\text{data})$. In the words used in Chapter 1, p. 14, Bayes' theorem provides a vehicle for obtaining epistemic probabilities from descriptive probabilities (see Section 16.1.1). The inverting of conditional probability statements, together with the recognition that a different notion of probability was involved, led to the name "inverse probablity" during the early 1800s. This has been replaced by the name "Bayes" in the theorem, and the adjective "Bayesian" to describe many of its applications.[4] To derive the theorem we need a preliminary result which is also important.

probability

i

Theorem: Law of Total Probability For events A and B we have

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

Proof: We begin by decomposing B into two pieces: $B = (B \cap A) \cup (B \cap A^c)$. Because A and A^c are disjoint, $(B \cap A)$ and $(B \cap A^c)$ are disjoint. We then have $P(B) = P(B \cap A) + P(B \cap A^c)$. Applying the multiplication rule to $P(B \cap A)$ and $P(B \cap A^c)$ gives the result. □

Bayes' Theorem in the Simplest Case If $P(B) > 0$ then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}. \tag{3.1}$$

Proof: We begin with the definition of conditional probability and then use the multiplication rule in the numerator and the law of total probability in the denominator:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}. \end{aligned} \quad \square$$

The "simplest case" modifier here refers to the statement of the theorem in which the law of total probability is applied to the denomoninator $P(B)$ by decomposing B by intersection with only two events, A and A^c. We discuss other versions of the theorem below.

One interesting class of problems where this simple case is useful is in the interpretation of clinical diagnostic screening tests. These tests are used to indicate that a patient may have a particular disease A, based on a test outcome B, but they are not definitive. The probability $P(B|A)$ that a patient having the disease tests positively is known as the *sensitivity* of the test, the probability $P(B^c|A^c)$ that a patient who does *not* have the disease tests negatively is known as the *specificity* of the test, and the probability $P(A)$ that a patient drawn randomly from the population has the disease

[4] For historical comments see Stigler (1986) and Fienberg (2006).

is known as the *prevalence* of the disease. Good diagnostic screening tests have sensitivity and specificity close to 1 but, as we will describe, Bayes' Theorem serves as a quantitative reminder that when a disease is rare, screening tests are preliminary, and other information will be needed to provide a diagnosis. Specifically, if we let $PPV = P(A|B)$, which stands for *positive predictive value*, we get

$$PPV = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})} \tag{3.2}$$

and, when the prevalence is small, the value of PPV will also typically be small—sometimes surprisingly small.

A famous example involves screening for prostate cancer based on the radioimmunoassay prostatic acid phosphatase (PSA). Even though the test is reasonably accurate, the disease remains sufficiently rare among young men that a random male who tests as positive will still have a low probability of actually having prostate cancer. An application of Bayes' Theorem (with A being the event that a randomly chosen man will have the disease and B the event that he tests positive) to data from Watson and Tang (1980), places the probability of disease given a positive test at about 1/125. The intuition comes from recognizing that, among men under age 65 in the United States, the disease has a prevalence of about 1/1,500. Suppose we were to examine 1,500 men, 1 of whom actually had the disease. If the screening test were 90 % accurate, a 10 % false positive rate would mean that about 150 men would test positively. In other words, about 1/150 of the positively tested men would actually have the disease. Bayes' Theorem refines this very crude calculation. Here is an example drawn from neurology.

Example 3.2 Diagnostic test for vascular dementia Vascular dementia (VD) is the second leading cause of dementia. It is important that it be distinguished from Alzheimer's disease because the prognosis and treatments are different. In order to study the effectiveness of clinical tests for vascular dementia, Gold et al. (1997) examined 113 brains of dementia patients post mortem. One of the clinical tests these authors considered was proposed by the National Institute of Neurological Disorders and Stroke (NINDS, an institute of NIH). Gold et al. found that the proportion of patients with VD who were correctly identified by the NINDS test, its sensitivity, was .58, while the proportion of patients who did not have VD who were correctly so identified by the NINDS test, its specificity, was .80. Using these results, let us consider an elderly patient who is identified as having VD by the NINDS test, and compute the probability that this person will actually have the disease. Let A be the event that the person has the disease and B the event that the NINDS test is positive. We want $P(A|B)$, and we are given $P(B|A) = .58$ and $P(B^c|A^c) = .8$. To apply Bayes' Theorem we need the disease prevalence $P(A)$. Let us take this probability to be $P(A) = .03$ (which seems a reasonable value based on Hébert and Brayne (1995)). We then also have $P(A^c) = .97$ and, in addition, $P(B|A^c) = 1 - P(B^c|A^c) = .2$. Plugging these numbers into the formula gives us

postmortem

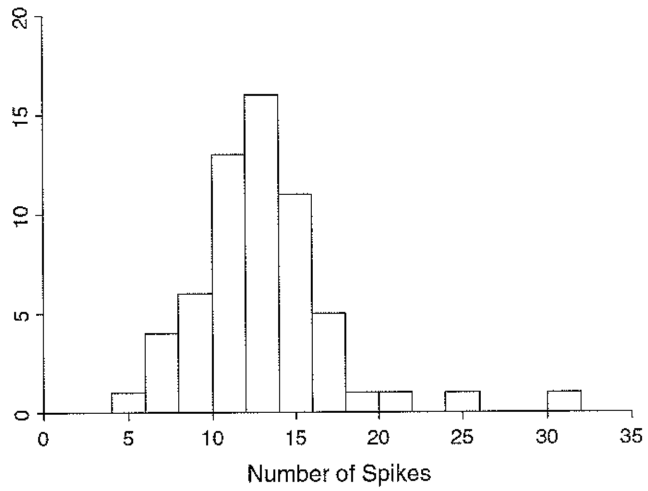


Fig. 3.2 Histogram of spike counts from a motor cortical neuron. The histogram displays 60 spike counts from a particular neuron recorded in primary motor cortex across 60 repetitions of the practiced condition.

together with Bayes' Theorem, we may determine from the spike count B the probability that the saccade will be in each of the four directions. In Bayesian decoding, the signals from many neurons are combined, and the direction A_k having the largest probability $P(A_k|B)$ is considered the "predicted" direction. In unpublished work, our colleague Dr. Valérie Ventura found that, from 55 neurons, Bayesian decoding was able to predict the correct direction more than 95 % of the time. □

3.2 Random Variables

So far we have discussed the basic rules of probability, which apply to sets representing uncertain events. A far more encompassing framework is obtained when we consider quantities measured from those events. For example, the number of times a neuron fires during a particular task may be observed, yielding a spike count. When the behavior is repeated across many trials, the spike counts will vary.

Example 3.4 Spike counts from a motor cortical neuron Matsuzaka et al. (2007) studied cortical correlates of practicing a movement repeatedly by comparing the firing of neurons in primary motor cortex during two sequential button-pressing tasks: one in which the sequence was highly practiced, and the other in which the sequence was determined at random. Figure 3.2 displays spike counts from a single neuron across 60 repititions of the practiced condition. The histogram displays substantial variation among the counts. □

repetitions
e

To describe variation among quantitative measurements, such as that seen in Fig. 3.2, we need to introduce mathematical objects called *random variables*, which assign to each outcome (e.g., neuronal spiking behavior on a particular trial) a number

$$P(X=2)=p^2,$$
$$P(X=1)=p(1-p)+(1-p)p=2p(1-p)$$
$$P(X=0)=(1-p)^2.$$

In this situation X is a random variable and it has a *binomial distribution.* More generally, given a sample space Ω, a *random variable* is a mapping that assigns to every element of Ω a real number. That is, if $\omega \in \Omega$ (see p. 38) then $X(\omega)=x$ is the value of the random variable X when ω occurs. In the context above, $\Omega = \{AA, AA^c, A^cA, A^cA^c\}$ and $X(AA)=2, X(AA^c)=1, X(A^cA)=1, X(A^cA^c)=0$.

In Chapter 1 we discussed the distinction between continuous and discrete data. We may similarly distinguish continuous and discrete random variables: a random variable is continuous if it can take on all values in some interval (A, B), where it is possible that either $A=-\infty$ or $B=\infty$ or both. The mathematical distinctions between discrete and continuous distributions are that (i) discrete distributions assign probabilities to specific values (such as non-negative integers) that can be separated from each other, but continuous distributions assign probabilities to intervals of non-separable numbers (such as numbers in the interval $(0,1)$) and (ii) wherever summation signs appear for discrete distributions, integrals replace them for continuous distributions.

3.2.2 *Distributions of random variables are defined using cumulative distribution functions and probability density functions, from which theoretical means and variances may be computed.*

There are several definitions we need, which will apply to other probability distributions besides the binomial. In the case of two trials from patient P. S., discussed on p. 47, the probabilities $P(X=0)$, $P(X=1)$, and $P(X=2)$ form the *probability mass function.* For convenience, as indicated in Section 3.2.3, we generally instead call the probability mass function a *probability density function (pdf).* We would typically write $P(X=x)$, with x taking the values $0, 1, 2$, and we also use the notation $f(x)=P(X=x)$. The function $F(x)=P(X \leq x)$ is called the *cumulative distribution function (cdf).* Thus, in the case of two trials from patient P.S. we have $F(0)=P(X=0)$, $F(1)=P(X \leq 1)=P(X=0)+P(X=1)$, and $F(2)=P(X \leq 2)=P(X=0)+P(X=1)+P(X=2)$. From the pdf we can obtain the cdf, and vice-versa. When we speak loosely of the "probability distribution of X," or the "distribution of X," we will be referring generically to the range of probabilites attached to X, which could be specified by either the pdf or the cdf.

probabilities

i

that all n values of x are equally likely) we get back[7] $\mu_X = \bar{x}$. Because data are often called *samples*, the data-based mean and standard deviation are often called the *sample mean* and the *sample standard deviation* to differentiate them from μ_X and σ_X, which are often called the *population mean and standard deviation.* This terminolgy distinguishes samples from "populations," rather than distributions, with the word "sample" connoting a batch of observations randomly selected from some large population. Sometimes there is a measurement process that corresponds to such random selection. However, as we have already mentioned, probability is much more general than the population/sample terminology might lead one to expect; specifically, we do not need to have a well-defined population from which we are randomly sampling in order to speak of a probability distribution. So, at least in principle, we might rather avoid calling μ_X a population mean. On the other hand, the "sample" terminology is useful for emphasizing that we are dealing with the observations, as opposed to the theoretical distribution, and it is deeply imbedded in statistical jargon. Similarly, the "population" identifier is frequently used rather than "theoretical." The crucial point is that one must be careful to distinguish between a theoretical distribution and the actual distribution of some sample of data. Many analyses assume that data follow some particular theoretical distribution, and in doing so *hope* that the match between theory and reality is pretty good. We will look at ways of assessing this match in Section 3.3.1.

terminology

o

The following properties are often useful.

Theorem For a discrete random variable X with mean μ_X and standard deviation σ_X we have

$$E(a \cdot X + b) = a \cdot \mu_X + b \tag{3.4}$$

$$\sigma^2_{aX+b} = a^2 \cdot \sigma^2_X \tag{3.5}$$

$$\sigma_{aX+b} = |a| \cdot \sigma_X. \tag{3.6}$$

Proof: We have

$$\begin{aligned} E(aX + b) &= \sum_x (ax + b)f(x) \\ &= a(\sum_x xf(x)) + b\sum_x f(x) \\ &= aE(X) + b \end{aligned}$$

which is the same as (3.4). The derivation of (3.5) is similar, and taking square-roots gives (3.6). □

[7] We also get $\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2}$ which, when we replace μ_X with $\bar{X}$, is not quite the same thing as the *sample standard deviation*; the latter requires a change from n to $n - 1$ as the divisor for certain theoretical reasons, including that the sample variance then becomes an *unbiased* estimator of σ^2_X. See p. 183.

3.2.3 *Continuous random variables are similar to discrete random variables.*

Suppose X is a continuous random variable on an interval (A, B), with $A = -\infty$ and $B = \infty$ both being possible. The *probability density function* (pdf) of X will be written as $f(x)$ where now

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

and, because (from Axiom 2 on p. 38) the total probability is 1, we have

$$\int_A^B f(x)dx = 1.$$

Note that in this continuous case there is no distinction between $P(a \leq X)$ and $P(a < X)$ (we have $P(X = a) = 0$). We may think of $f(x)$ as the probability per unit of x; $f(x)dx$ is the probability that X will lie in an infinitesimal interval about x, that is, $f(x)dx = P(x \leq X \leq x + dx)$. In some contexts there are various random variables being considered and we write the pdf of X as $f_X(x)$.

A technical point is that when either $A > -\infty$ or $B < \infty$ or both, by convention, the pdf $f(x)$ is extended to $(-\infty, \infty)$ by setting $f(x) = 0$ outside (A, B). When we say that X is a continuous random variable on an interval (A, B) we will mean that $f(x) > 0$ on (A, B) and, if either A or B is a number, $f(x) = 0$ outside of (A, B). We next give several examples of continuous distributions.

Illustration: Uniform distribution Perhaps the simplest example is the *uniform distribution*. For instance, if the time of day at which births occurred followed a uniform distribution, then the probability of a birth in any given 30 min period would be the same as that for any other 30 min period throughout the day. In this case the pdf $f(x)$ would be constant over the interval from 0 to 24 h. Because it must integrate to 1, we must have $f(x) = 1/24$ and the probability of a birth in any given 30 min interval starting at a hours is $\int_a^{a+.5} f(x)dx = 1/48$. When a random variable X has a uniform distribution on a finite interval (A, B) we write this as $X \sim U(A, B)$ and the pdf is $f(x) = \frac{1}{B-A}$. □

In this illustration above we have introduced a convention that is ubiquitous, both in this book and throughout statistics: the squiggle "$\sim$" means "is distributed as."

Figure 3.3 displays pdfs for four common distributions. For the two in the top panels, exponential and gamma distributions, X may take on all positive values, i.e., values in $(0, \infty)$. The lower left panel shows a beta distribution, which is confined to the interval $(0, 1)$. A normal distribution, which ranges over the whole real line, is shown in the bottom right panel. We discuss the exponential and normal distributions briefly below and return to them, and to the beta and gamma dis^ributions in Chapter 5.

distributions

t

analogous
o

is the *variance* of X. Note that in each of these formulas we have simply replaced sums by integrals in the analgous definitions for discrete random variables. Note, too, that *pdf* and *cdf* values for certain continuous distributions may be computed with statistical software.[8] We again have

$$\mu_{a \cdot X+b} = a \cdot \mu_X + b \tag{3.8}$$

$$\sigma_{a \cdot X+b} = |a| \cdot \sigma_X. \tag{3.9}$$

These formulas are just as easy to prove as (3.4) and (3.6). Another formula is useful for certain calculations:

$$V(X) = E(X^2) - \mu^2 \tag{3.10}$$

and this, too, is easily verified. In many contexts the variation relative to the mean is summarized using the *coefficient of variation*, given by

$$\text{CV}(X) = \frac{\sigma}{\mu}. \tag{3.11}$$

The *quantiles* or *percentiles* are often used in working with continuous distributions: for p a number between 0 and 1 (such as .25), the p quantile or 100pth percentile (e.g., the .25 quantile or the 25th percentile) of a distribution having cdf $F(x)$ is the value η such that $p = F(\eta)$. Thus, we write the p quantile as $\eta_p = F^{-1}(p)$, where F^{-1} is the inverse cdf.

Illustration: Exponential distribution Let us illustrate these ideas in the case of the exponential distribution, which is special because it is easy to handle and also because of its importance in applications. We provide an application in Example 3.5

A random variable X is said to have an exponential distribution with parameter λ, with $\lambda > 0$, when its pdf is

$$f(x) = \lambda e^{-\lambda x} \tag{3.12}$$

for $x > 0$, and is 0 for $x \leq 0$. We will then say that X has an *Exp*(λ) distribution and we will write $X \sim Exp(\lambda)$. The pdf of X when $X \sim Exp(1)$ is shown in Fig. 3.6. Also illustrated in that figure is computation of probabilities as areas under the pdf for the case

$$P(X > 2) = \int_2^\infty f(x)dx$$

which means we compute the area under the curve to the right of $x = 2$. For the exponential distribution this value is easy to compute using calculus. The cdf of an exponential distribution is

[8] The definitions of expectation and variance assume that the integrals are finite; there are, in fact, some important probability distributions that do not have expectations or variances because the integrals are infinite.

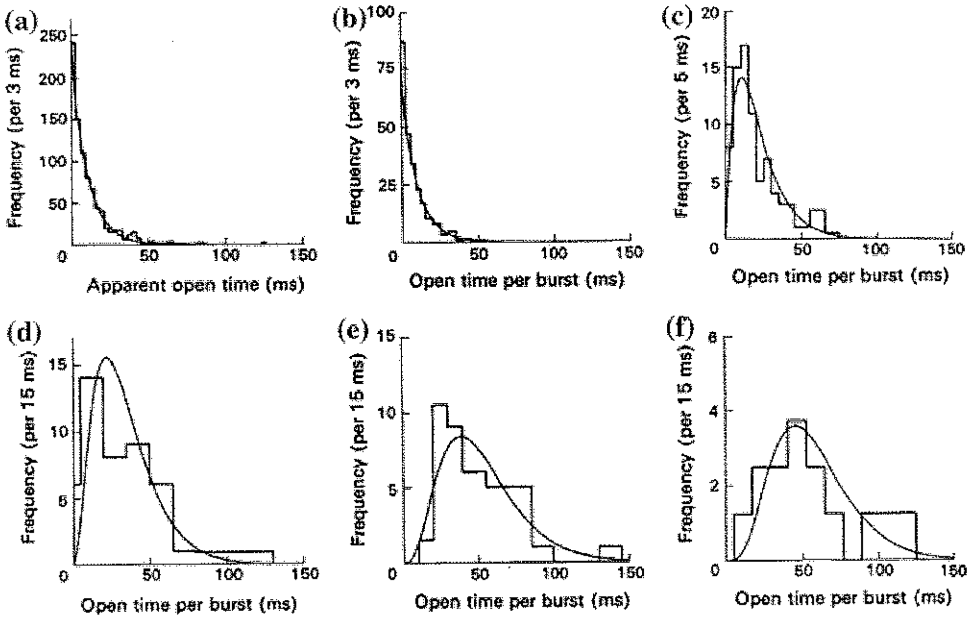


Fig. 3.8 Duration of channel openings. Panel **a** depicts the distribution of burst durations for a particular agonist. Panel **b** displays the distribution of bursts for which there was only 1 opening, with an exponential pdf overlaid. This illustrates the good fit of the exponential distribution to the durations of ion channel opening. Panel **c** displays the distribution of bursts for which there were 2 apparent openings, with a gamma pdf, with shape parameter 2, overlaid. Panel **c** again indicates good agreement. Panels **d–f** show similar results, for bursts with 3–5 openings. Adapted from Colquhoun and Sakmann (1985).

Illustration: Uniform distribution (continued from p. 52) If a continous random variable X has cdf $F(x) = x$ on the interval $(0, 1)$ we may differentiate to get the $U(0, 1)$ pdf $f(x) = 1$. On the other hand, if $X \sim U(0, 1)$ we integrate $f(x) = 1$ to get

$$F(x) = \int_0^x 1 \cdot dx = x.$$

continuous

In other words, X has a $U(0, 1)$ distribution if and only if its cdf is $F(x) = x$ on the interval $(0, 1)$. □

Illustration: Normal distribution (continued from p. 53) When X is distributed normally with mean μ and standard deviation σ it has a pdf given by Eq. 3.7. Its cdf is given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\right) dx.$$

This integral can not be evaluated in explicit form. Therefore, normal probabilities of the form $P(a \leq X \leq b)$ are obtained by numerical approximation. □

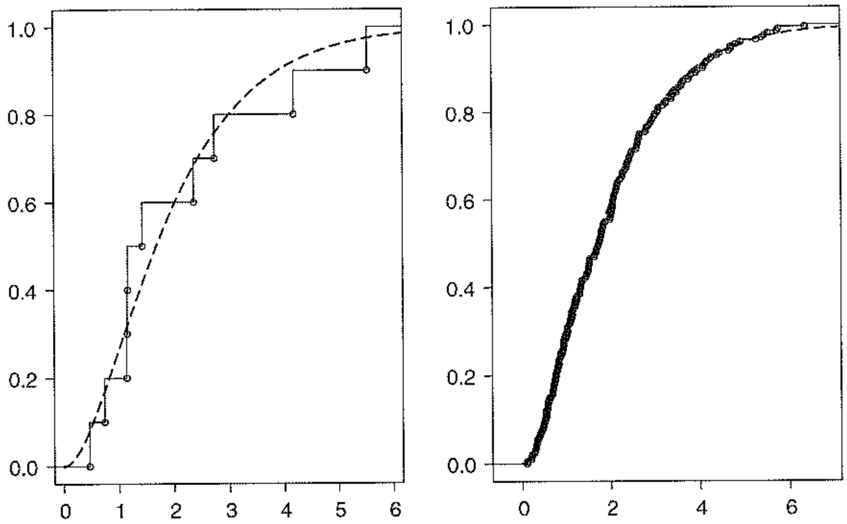


Fig. 3.9 Convergence of the empirical cdf to the theoretical cdf. The *left panel* displays the empirical cdf for a random sample of size 10 from the *gamma* distribution whose pdf is in the *top right panel* of Fig. 3.3, together with the gamma cdf (*dashed line*). The *right panel* shows the empirical cdf for a random sample of size 200, again with the gamma cdf. In the *right panel* the emprical cdf is quite close to the theoretical gamma cdf.

empirical

i

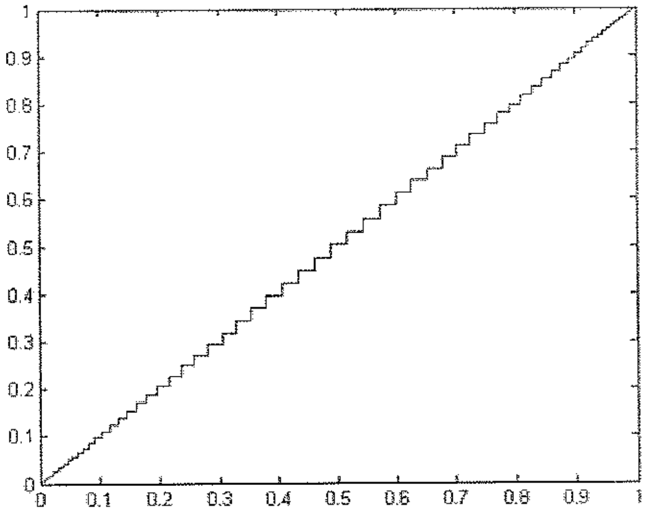


Fig. 3.10 A P–P plot of the MEG noise data from Fig. 3.4. The straightness of the plot indicates excellent agreement with the normal distribution.

that there is no unique analogue and instead one of several variants may be used. If we start from a sample of observations $x_1, x_2, \ldots, x_n$ we first put the data in ascending order according to the size of each observation: we write $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)}$ is the smallest value, $x_{(2)}$ is the second-smallest, and $x_{(n)}$ is the largest. Let us use

3.2.4 The hazard function provides the conditional probability of an event, given that it has not yet occurred. ~~of a random variable X at x is its conditional probability density, given that $X = x$.~~

fix ~~see~~ table of contents also

Another useful characterization of a probability distribution arises in specialized contexts, including the analysis of spike train data, where a random variable X represents the waiting time until some event occurs. In the case of a spiking neuron, X would be the elapsed time since the neuron last fired, and the event of interest would be next time it fires. We want a formula for the instantaneous probability that the neuron will fire at time x, i.e., that it will fire in an interval $(x, x + dx)$, given that it has not yet fired in $(0, x)$. Assuming X is a continuous random variable, the event that the neuron has not yet fired in $(0, x)$ is the same as $X > x$. Recall that if $P(B) > 0$ then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Applying this with A being the event $X \in (x, x+h)$ and B being the event that $X > x$ we have

$$P(X \in (x, x+h)|X > x) = \frac{F(x+h) - F(x)}{1 - F(x)}.$$

Passing to the limit as h vanishes gives

$$\lim_{h \to 0} \frac{P(X \in (x, x+h)|X > x)}{h} = \frac{f(x)}{1 - F(x)},$$

which we may interpret as the probability $X \in (x, x+dx)$ given $X > x$. The function

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

is called the *hazard function* of X. For example, if X is the elapsed time that an ion channel is open, so that its values are times x, then $\lambda(x)dx$ becomes the probability the ion channel will close in the interval $(x, x+dx)$, given that it has remained open up to time t. Similarly, if X is the elapsed time since a neuron last fired an action potential then $\lambda(x)dt$ becomes the probability the neuron will fire in the interval $(x, x + dx)$, given that it has not yet fired again before elapsed time x. In spike train analysis, the hazard function for a neuron becomes its theoretical firing rate (its instantaneous probability of firing per unit time), which is known in general as the *intensity* or *conditional intensity* function. See Chapter 19.

The "hazard" terminology comes from lifetime analysis, where the random variable X is the lifetime (of a lightbulb or a person, etc) in units of time t and $\lambda(t)dt$ is the probability of failure (death) in the interval $(t, t + dt)$ given that failure has not yet occurred.

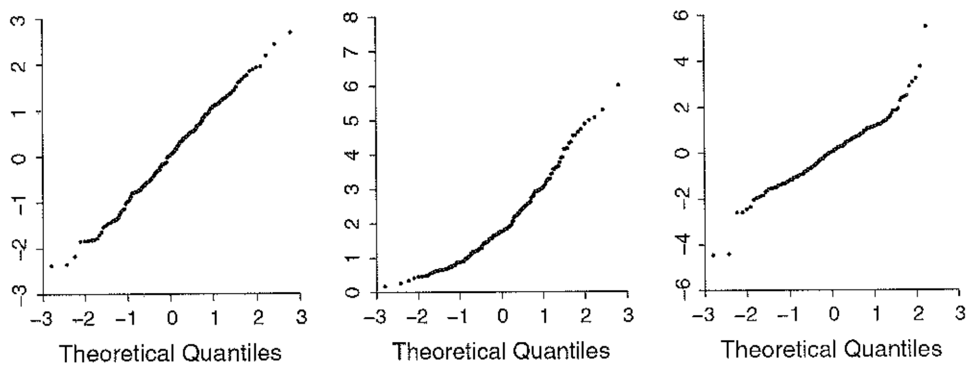


Fig. 3.11 Q–Q plots for 200 randomly-drawn observations from a three distributions. *Left*: observations from a $N(0, 1)$ distribution; *middle*: observations from a gamma distribution, whose pdf is shown in the *top right panel* of Fig. 3.3, which is skewed to toward high values; *right*: observations from a t distribution (see Section 5.4.7), which is symmetric with heavy tails. In each case the theoretical quantiles come from a normal distribution.

italic r

rth

remove hyphen

r to denote the index of ordered values, meaning that $x_{(r)}$ is the r-th smallest value. Working by analogy with the definition $\eta = F^{-1}(p)$ we could define the $\frac{r}{n}$ *sample quantile*, or the $100\frac{r}{n}$ *sample percentile*, by setting $p = \frac{r}{n}$ and replacing F with $\hat{F}_n$ to get $\hat{F}_n^{-1}(\frac{r}{n}) = x_r$. We then define

$$\eta_{(r)} = \tilde{F}^{-1}(\frac{r}{n})$$

for $r = 1, \ldots, n$ and plot the ordered data against these values. That is, we plot the points $(\eta_{(1)}, x_{(1)}), \ldots, (\eta_{(n)}, x_{(n)})$. Most software modifies the details of this procedure, but the idea remains the same.

> *Details:* A common variation is to take x_r to be the $100\frac{r-.5}{n}$ *sample percentile*. To see why this makes some sense, suppose we have $n = 7$ ordered observations. Then the 4th is the median. This divides the 7 numbers into the 3 smallest and the 3 largest and, effectively says that the 4th is part of both the smallest half of the numbers and the largest half of the numbers. It could therefore be considered the 3.5th ordered value. The reasoning behind the designation of $x_{(r)}$ as the $\frac{r-.5}{n}$ quantile is similar. Statistical software sometimes chooses alternative definitions based on expected values of $x_{(r)}$ under particular assumptions. Also, in creating a P–P plot, some software plots $\hat{F}(x_{(r)})$ against $\frac{r-.5}{n}$. □

Figure 3.11 displays three Q–Q plots, for which the theoretical quantiles are based on the normal distribution. Thus, we would make these plots in order to check whether the data could reasonably be described by a normal distribution. The three data sets were generated on the computer from three very different probability distributions.

variables the most common way to measure dependence is through their correlation, which is discussed in Section 4.2.1. We first interpret the correlation as a measure of linear dependence then, in Section 4.2.2, describe its role in the bivariate normal distribution. After we discuss conditional densities in Section 4.2.3 we re-interpret correlation using conditional expectation in Section 4.2.4. We then turn to the case of arbitrarily many random variables ($X_1, \ldots, X_n$ with $n \geq 2$), providing results in Section 4.3 that will be useful later on. We discuss general multivariate normal distributions later, in Section 5.5.

4.2.1 *The linear dependence of two random variables may be quantified by their correlation.*

When we consider X and Y simultaneously, we may characterize numerically their joint variation, meaning their tendency to be large or small together. This is most commonly done via the *covariance* of X and Y which, for continuous random variables, is

$$\begin{aligned} Cov(X, Y) &= E\left((X - \mu_X)(Y - \mu_Y)\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy \end{aligned}$$

and for discrete random variables the integrals are replaced by sums. The covariance is analagous to the variance of a single random variable. We now generalize Eq. (4.5) to the case in which the random variables may not be independent.

analogous

o

Theorem: Variance of a Sum of Random Variables For random variables X_1 and X_2 we have

$$V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2abCov(X_1, X_2).$$

More generally, for random variables $X_1, X_2, \ldots, X_n$ we have

$$V(\sum_{i=1}^{n} a_i X_i) = \left(\sum_{i=1}^{n} a_i^2 V(X_i)\right) + 2\sum_{i<j} a_i a_j Cov(X_i, X_j). \tag{4.6}$$

Proof: The proof follows from the definition by straightforward algebraic manipulations and is omitted. □

The covariance depends on the variability of X and Y individually, as well as their joint variation, and therefore depends on scaling. For instance, as is immediately verified from the definition, $Cov(3X, Y) = 3Cov(X, Y)$. To obtain a measure of joint variation that does not depend on the variance of X and Y, we standardize. The *correlation* of X and Y is

$$\beta_{X|Y} = \rho\frac{\sigma_X}{\sigma_Y} \tag{4.20}$$

so that if we combine (4.19) and (4.20) we get the following expression for the correlation:

$$\rho = \text{sign}(\beta_{Y|X})\sqrt{\beta_{Y|X}\beta_{X|Y}} \tag{4.21}$$

where $\text{sign}(\beta_{Y|X})$ is -1 if $\beta_{Y|X}$ is negative and 1 if $\beta_{Y|X}$ is positive.

Compare Eq. (4.18) to Eqs. (4.9) and (4.10). From(4.9) and (4.10) we have that the best linear predictor of Y based on X is $f(X)$ where

$$f(x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)\,. \tag{4.22}$$

In general, we may call this the *linear regression* of Y on X. In the case of bivariate normality, the regression of Y on X is equal to the *linear* regression of Y on X, i.e., the regression is linear. We derived (4.22) as the best linear predictor of Y based on X by minimizing mean squared error. More generally, if we write the regression function as $M(x) = E(Y|X = x)$. Then $M(X)$ is the best predictor of Y in the sense of minimizing mean squared error.

Prediction Theorem The function $f(x)$ that minimizes $E((Y - f(X))^2)$ is the conditional expectation $f(x) = M(x) = E(Y|X = x)$.

> *Proof details:* Note that $E(Y - M(X)) = E(Y) - E(E(Y|X))$ and by the law of total expectation (p. 85) this is zero. Now write $Y - f(X) = (Y - M(X)) + (M(X) - f(X))$ and expand $E((Y - f(X))^2)$ to get
>
> $$E((Y - f(X))^2) = E((Y - M(X))^2) + 2E((Y - M(X))\\ (M(X) - f(X))) + E((M(X) - f(X))^2). \tag{4.23}$$
>
> Applying the law of total expectation to the second term we get
>
> $$E((Y - M(X))(M(X) - f(X))) = E(E((Y - M(X))(M(X)\\ - f(X))|X)))$$
>
> but for every x we have
>
> $$E((Y - M(X))(M(X) - f(X))|X = x) = (M(x) - M(x))(M(x)\\ - f(x)) = 0$$
>
> so that the second term in (4.23) is 0. The third term $E((M(X) - f(X))^2)$ is alwaysnon-negative and it is zero when $f(x)$ is chosen

Space

$Cor(X,Y)^2 = Cor(f(X), g(Y))^2$ (see attached page)

$$I(X, Y) = -\frac{1}{2}\log(1-\rho^2). \qquad (4.29)$$

Thus, when X and Y are independent, $I(X, Y) = 0$ and as they become highly correlated (or negatively correlated) $I(X, Y)$ increases indefinitely. □

Theorem For random variables X and Y that are either discrete or jointly continuous having a positive joint pdf, mutual information satisfies (i) $I(X, Y) = I(Y, X)$, (ii) $I(X, Y) \geq 0$, (iii) $I(X, Y) = 0$ if and only if X and Y are independent, and (iv) for any one-to-one continuous transformations $f(x)$ and $g(y)$, $I(X, Y) = I(f(X), g(Y))$.

Proof: Omitted. See, e.g, Cover and Thomas (1991). □

Property (iv) makes mutual information quite different from correlation. For correlation, $Cor(X, Y) = Cor(f(X), g(Y))$ when $f(x)$ and $g(y)$ are linear functions, but when they are nonlinear the value of the correlation can change. squared

The use here of the word "information" is important. For emphasis we say, in somewhat imprecise terms, what we think is meant by this word.

> Roughly speaking, information about a random variable Y is associated with the random variable X if the uncertainty in Y is larger than the uncertainty in $Y|X$.

For example, we might interpret "uncertainty" in terms of variance. If the regression of Y on X is linear, as in (4.18) (which it is if (X, Y) is bivariate normal), we have

$$\sigma^2_{Y|X} = (1-\rho^2)\sigma^2_Y. \qquad (4.30)$$

In this case, information about Y is associated with X whenever $|\rho| > 0$ so that $1-\rho^2 < 1$. The reduction of uncertainty in Y provided by X becomes

$$\sigma^2_Y - \sigma^2_{Y|X} = \rho^2\sigma^2_Y,$$

which retains the multiplier σ^2_Y (coming from the multiplicative form of (4.31)). To remove the factor σ^2_Y we may consider the relative reduction of uncertainty,

$$\frac{\sigma^2_Y - \sigma^2_{Y|X}}{\sigma^2_Y} = \rho^2.$$

In this sense, ρ^2 becomes a measure of the information about Y supplied by X.

A different rewriting of (4.30) will help us connect it more strongly with mutual information. First, if we redefine "uncertainty" to be standard deviation rather than variance, (4.30) becomes

$$\sigma_{Y|X} = \sqrt{1-\rho^2}\sigma_Y. \qquad (4.31)$$

for page 94

$$Cor(X,Y)^2 = Cor(f(X),g(Y))^2$$

representing a transmitted message and Y is a random variable representing the received message after noise has been injected during the transmission process, then the channel capacity is

$$C = \max_X I(X, Y)$$

where the maximum is taken over all possible distributions of X. This concept, developed to characterize electronic communication channels, has also been applied to human behavior and neural activity. Because the mutual information in this context concerns discrete distributions for (X, Y), and $\log_2$ is used, the units are said to be in *bits* for "binary digits" (because, for a positive integer n, $\log_2(n)$ is the number of binary digits used to represent n in base 2). Thus, human and neural information processing capacity is usually reported in bits.

Example 4.5 The Magical Number Seven In a famous paper, George Miller reviewed several psychophysical studies that attempted to characterize the capacity of humans to process sensory input signals (Miller 1956). One study, for example, exposed subjects to audible tones of several different values of pitch (frequency) and asked them to identify the pitch (e.g., pitch 1, 2, or 3, corresponding to high, medium, or low). The question was, how many distinct values of pitch can humans reliably discriminate? It turned out that with five or more tones of different pitch, the human observers made frequent mistakes. The experimental design allowed calculation of the probability of responding with a particular answer Y based on a particular input tone X, and with this the mutual information could be calculated. By examining several different studies, of similar yet different types, Miller concluded that mutual information had an asymptotic maximum at about $C = 2.6 \pm .6$ bits, which could be interpreted as the channel capacity of a human observer. Transforming this back to numbers of discernable categories gives $2^{2.6-.6} = 4$ and $2^{2.6+.6} = 9.2$. After looking at other, related psychophysical data Miller summarized by saying there was a "magical number seven, plus or minus two," which characterized many aspects of human information processing in terms of channel capacity. □

discernible

i

Mutual information has also been used extensively to quantify the information about a stochastic stimulus Y associated with a neural response X. In that context the notation is often changed by setting $S = Y$ for "stimulus" and $R = X$ for neural "response," and the idea is to determine the amount of information about the stimulus that is associated with the neural response.

Example 4.6 Temporal coding in inferotemporal cortex In an influential paper, Optican and Richmond (1987) reported responses of single neurons in inferotemporal (IT) cortex of monkeys while the subjects were shown various checkerboard-style grating patterns as visual stimuli. Optican and Richmond computed the mutual information between the 64 randomly-chosen stimuli (the random variable Y here taking 64 equally-likely values) and the neural response (X), represented by a vector of time-varying firing rates across multiple time bins. They compared this with the mutual information between the stimuli and a single firing rate across a large time interval and concluded that there was considerably more mutual information in the

4.3.4 Bayes classifiers are optimal.

Suppose X is a random variable (or random vector) that may follow one of two possible distributions having pdf $f_1(x)$ or $f_2(x)$. If $X = x$ is observed, which distribution did it come from? This is the problem of binary *classification*. Typically, there is a random sample $X_1, \ldots, X_n$ and the problem is to classify (to one of the two distributions) each of the many observations. A *decision rule* or *classification rule* is a mapping that assigns to each possible x a classification (that is, a distribution). A classic scenario for binary classification is when patients having characteristics summarized in a vector x (for example, brain features found from PET imaging), are to be considered diseased (e.g., having Alzheimer-like amyloid deposits, see Vandenbergh et al. (2013)) or not. The problem extends to m categories, where X follows one of many alternative distributions, with pdf $f_i(x)$, for $i = 1, \ldots, m$. A classification error is made if $X \sim f_k(x)$ and the observation $X = x$ is classified as coming from $f_i(x)$ with $i \neq k$. In this section we present a remarkable result: it is, in principle, possible to define a classifier that minimizes the probability of classification error.

Let C_i refer to the case $X \sim f_i(x)$. We use the letter C to stand for "class," so that the problem is to assign to each observed x a class C_i. We assume that X is selected from class C_i with probability $P(C = C_i) = \pi_i$, for $i = 1, \ldots, m$. Often the π_i probabilities are taken to be equal, i.e., $\pi_i = 1/m$, for $i = 1, \ldots, m$ (so that the classes are *a priori* equally likely), but the theory does not require this. The *Bayes classifier* assigns to each observed value x the class having the maximal *posterior probability*

$$P(C = C_k|X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{m} f_i(x)\pi_i} \tag{4.38}$$

among all the classes C_i. Writing $f_i(x) = f_{X|C}(x|C = C_i)$, Eq. (4.38) has the same form as (4.37). The following theorem says that Bayes classifiers minimize the probability of classification error.

Theorem on Optimality of Bayes Classifiers Suppose X is drawn from a distribution having pdf $f_i(x)$, where $f_i(x) > 0$ for all x, with probability π_i, for $i = 1, \ldots, m$, where $\pi_1 + \cdots + \pi_m = 1$, and let C_i be the class $X \sim f_i(x)$. Then the probability of committing a classification error is minimized if $X = x$ is classified as arising from the distribution having pdf $f_k(x)$ for which C_k has the maximum posterior probability given by (4.38).

The proof is somewhat lengthy and appears at the end of this section.

Corollary Suppose that with equal probabilities X is drawn either from a distribution having pdf $f_1(x)$, where $f_1(x) > 0$ for all x, or from a distribution having pdf $f_2(x)$, where $f_2(x) > 0$ for all x. Then the probability of committing a classification error is minimized if $X = x$ is classified to the distribution having the higher pdf at x.

Corollary Suppose n observations $X_1, \ldots, X_n$ are drawn, independently, from a distribution havingpdf $f_i(x)$, where $f_i(x) > 0$ for all x, with probability π_i, for

Space

Some additional references concerning ideal observer analysis, and Bayesian approaches to modeling neural systems more generally, appear at the beginning of Chapter 16. Here is a different setting in which utilities and Bayes rules have been invoked.

Example 4.9 ACT-R theory of procedural memory ACT-R is a theory of human problem-solving that is implemented in a computer program (Anderson 1993, 2007). A typical domain is elementary algebra problem-solving, involving equations such as $7x+3 = 38$. The many steps involved in solving algebra problems include actions such as "subtract," which require calls to memory (e.g., to retrieve $8-3=5$). These are encoded as *production rules* which are IF-THEN statements, and are often called *procedures.* At the completion of each step ACT-R must select from memory the next production rule to use. To do so it considers a utility function based on the value V of the goal, the probability P_i of achieving the goal if production rule i is selected, and the cost D_i of rule i. Each production rule is then assigned the utility

$$U_i = P_i V - D_i.$$

ACT-R picks the production rule with the highest utility. Because the probabilities are actually posterior probabilities based on previous experience, ACT-R may be considered to be using a Bayes rule for this situation. The acronym ACT stands for "adaptive character of thought" and the R is tacked on as a nod to "rational" in the sense of optimal decision-making. □

Proof of theorem on optimality of Bayes classifiers:
We consider the binary case where $m = 2$. We also assume the two distributions are discrete and, for simplicity, we take $\pi = \frac{1}{2}$. Here, the Bayes classfier assigns class C_1 to $X = x$ whenever $f_1(x) > f_2(x)$, and assigns class C_2 when $f_2(x) \geq f_1(x)$.

classifier

Let $R = \{x : f_1(x) \leq f_2(x)\}$. We want to show that the classification rule assigning $x \rightarrow f_2(x)$ whenever $x \in R$ has a smaller probability of error than the classification rule $x \rightarrow f_2(x)$ whenever $x \in A$ for any set A that is different than R. To do this we decompose R and its complement R^c as $R = (R \cap A) \cup (R \cap A^c)$ and $R^c = (R^c \cap A) \cup (R^c \cap A^c)$. We have

$$\sum_{x \in R} f_1(x) = \sum_{x \in R \cap A} f_1(x) + \sum_{x \in R \cap A^c} f_1(x) \tag{4.39}$$

and

$$\sum_{x \in R^c} f_2(x) = \sum_{x \in R^c \cap A} f_2(x) + \sum_{x \in R^c \cap A^c} f_2(x). \tag{4.40}$$

By the definition of R we have, for every $x \in R, f_1(x) \leq f_2(x)$ and, in particular, for every $x \in R \cap A^c, f_1(x) \leq f_2(x)$. Therefore, from (4.39) we have effective medications (such as Ritalin) involve inhibition of dopamine transport. There is also evidence of involvement of the nicotine system, possibly due to its effects on dopamine receptors. Kent et al. (2001) examined genotype frequencies for the nicotinic acetylcholine receptor subunit $\alpha 4$ gene among children with ADHD and their parents. At issue was the frequency of a $T \rightarrow C$ exchange in one base in the gene sequence. In order to carry out the standard analysis the authors first examined whether the population appeared to be in equilibrium. If so, the probabilities of the allele combinations TT, CT, CC would be given by $B(2, p)$ distribution, according to the Hardy-Weinberg model. The frequencies for the 136 parents in their study were as follows:

	TT	CT	CC
Number	48	71	17
Frequency	.35	.52	.13
Hardy-Weinberg Probability	.38	.47	.15

In this case, the probabilities determined from the Hardy-Weinberg model (how we obtain these will be discussed in Chapter 7) are close to the observed allele frequencies, and there is no evidence of disequilibrium in the population (we also discuss these details later). Kent et al. went on to find no evidence of an association between this genetic polymorphism and the diagnosis of ADHD. □

In some cases the probability p is not stable across rep$\overset{e}{\text{i}}$titions. Indeed, sometimes the change in probability is the focus of the experiment, as when learning is being studied.

repetitions
e

Example 5.2 Learning impairment following NMDA antagonist injection Experiments on learning often record responses of subjects as either correct or incorrect in sequences of trials during which the subjects are given feedback as to whether their responses are correct or not. The subjects typically begin with a probability of being correct that is much less than 1, perhaps near the guessing value of .5, but after some number of trials they get good at responding and have a high probability of being correct, i.e., a probability near 1. An illustration of this paradigm comes from Smith et al. (2005), who examined data from an experiment in rats by Stefani et al. (2003) demonstrating that learning is impaired following an injection of an NMDA antagonist into the frontal lobe. In a first set of trials, the rats learned to discriminate light from dark targets, then, in a second set of trials, which were the trials of interest, they needed to discrimate smooth versus rough textures of targets. In two groups of rats a buffered salt solution with the NMDA antagonist was injected prior to the second set of trials, and in two other groups of rats the buffered salt solution without the antagonist was injected. Figure 5.1 displays the responses across 80 learning trials for set 2. It appears from the plot of the data that the groups of rats without the NMDA antagonist did learn the second task more quickly than the second group of rats, as expected.

discriminate
in

B converges to $e^{-\lambda}$; and the expression over the third underbrace defining B converges to 1. This gives (5.7). □

5.2.3 *The Poisson distribution results when the binary events are independent.*

In thinking about the binomial assumption for a random variable X one generally ponders whether it is reasonable to conceptualize X as a sum of Bernoulli trials with the independence and homogeneity assumptions. Similarly, in the Poisson case, one typically asks whether the count variable X could be considered a sum of Bernoulli trials for small p and large n. The first requirement is that the counts really are sums of binary events. This means that X results from a string of 0s and 1s, as in Fig. 5.1, p. 109. In Example 5.4, p. 111, each emission event corresponds to a state transition in the nucleus of a particular atom. It is reasonable to assume that it is impossible for two nuclei to emit particles at precisely the same time and, furthermore, that each Geiger-counter "click" corresponds to exactly one particle emission. Independence, usually the crucial assumption, here refers to the independence of the many billions of n~~e~~uclei residing within the specimen. This is an assumption, apparently well justified, within the quantum-mechanical conception of radioactive decay. It implies, for example, that any tendency for two particles to be emitted at nearly the same time would be due to chance alone: because there is no interaction among the nuclei, there is no physical "bursting" of multiple particles. Furthermore, the probability of an emission would be unlikely to change over the course of the experiment unless the specim~~a~~n were so tiny that its mass changed appreciably. To summarize, the Poisson distribution for counts of events across time makes intuitive sense when we can conceptualize the events as Bernoulli trials, which are homogeneous and independent, where the success probability p is small.

nuclei

specimen

The framework we have constructed above to discuss emission of α particles would apply equally well to quanta of light in the Hecht et al. experiment. What about the vesicles at the neuromuscular junction? Here, the quantal hypothesis is what generates the sequence of dichotomous events (release vs. no release). Is release at one vesicle independent of release at another vesicle? If neighboring vesicles tend to release in small clumps, then we would expect to see more variability in the counts than that predicted by the Poisson, while if release from one vesicle tended to inhibit release of neighbors we would expect to see more regularity, and less variability in the counts. It is reasonable to begin by assuming independence, but ultimately it is an empirical question whether this is justified. Homogeneity is suspect: the release probability at one vesicle may differ substantially from that at another vesicle. However, as del Castillo and Katz realized, homogeneity is actually not an essential assumption. We elaborate on this point when we return to the Poisson distribution, and its relationship to the Poisson process in Section 19.2.2.

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

Thus, every exponential distribution is memoryless. On the other hand, let $G(x) = 1 - F(x)$ where $F(x)$ is the distribution function of X. Memorylessness implies

$$P(X > t + h) = P(X > t)P(X > h)$$

i.e.,

$$G(t + h) = G(t)G(h)$$

for all positive t and h. But (as mentioned in Section A.4 of the Appendix), $G(x)$ can satisfy this equation for all positive t and h only if it has an exponential form $G(x) = ae^{bx}$. Because $F(x) = 1 - G(x)$ is a distribution function, it satisfies $F(x) \to 1$ as $x \to \infty$, which implies $b < 0$, and it satisfies $F(x) \to 0$ as $x \to 0$, which implies $a = 1$. Thus $F(x) = 1 - e^{-\lambda x}$ for some λ, i.e., $X \sim Exp(\lambda)$. □

An additional characterization of the exponential distribution is that it has a constant hazard function.

Theorem: A continuous random variable X satisfies $X \sim Exp(\lambda_0)$ if and only if its satisfies
hazard function is $\lambda(x) = \lambda_0$.

Proof: First suppose $X \sim Exp(\lambda_0)$. The hazard function is easy to compute from the definition

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

Substituting $f(x) = \lambda_0 e^{-\lambda_0 x}$ and $F(x) = 1 - e^{-\lambda_0 x}$ we have

$$\begin{aligned}
\lambda(x) &= \frac{\lambda_0 e^{-\lambda_0 x}}{e^{-\lambda_0 x}} \\
&= \lambda_0.
\end{aligned}$$

On the other hand, if the hazard function is $\lambda(x) = \lambda_0$ we may rewrite the definition of $\lambda(x)$ and solve for $F(x)$, probabilities in the middle of the distribution, the t_ν distribution may be considered essentially the same as $N(0, 1)$. For small ν, however, the probability of large positive and negative values becomes much greater than that for the normal. For example, if $X \sim N(0, 1)$ then $P(X > 3) = .0014$ whereas if $T \sim t_3$ then $P(T > 3) = .029$, about 20 times the magnitude. To describe this phenomenon we say that the t_3 distribution has much *heavier tails* (or *thicker tails*) than the normal.

The t distribution was first derived by William Gosset under the pen name "A. Student." It is therefore often called *Student's t* distribution.

If $X \sim \chi^2_{\nu_1}$ and $Y \sim \chi^2_{\nu_2}$, independently, then

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

is said to have an F distribution on ν_1 and ν_2 degrees of freedom, which are usually referred to as the numerator and denominator degrees of freedom. We may write this as $F \sim F_{\nu_1,\nu_2}$. This distribution arises in regression and analysis of variance, where ratios of sums of squares are computed and each sum of squares has (under suitable assumptions) a chi-squared distribution.

When $\nu_1 = 1$ the numerator is the square of a normal and $F = T^2$, where T is the ratio of a $N(0, 1)$ and the square-root of a $\chi^2_{\nu_2}$. That is, the square of a t_ν distributed random variable has an $F_{1,\nu}$ distribution. Also, analagously to the situation with the t_ν distribution, when ν_2 gets large the denominator Y/ν_2 is a random variable that takes values mostly very close to 1 and F_{ν_1,ν_2} becomes close to a $\chi^2_{\nu_1}$.

analogously

5.5 Multivariate Normal Distributions

5.5.1 *A random vector is multivariate normal if linear combinations of its components are univariate normal.*

We now generalize the bivariate normal distribution, which we discussed in Section 4.2.2. We say that an m-dimensional random vector X has an *m-dimensional multivariate normal distribution* if every nonzero linear combination of its components is normally distributed. If μ and Σ are the mean vector and variance matrix of X we write this as $X \sim N_m(\mu, \Sigma)$. Using (4.25) and (4.26) we thus characterize $X \sim N_m(\mu, \Sigma)$ by saying that for every nonzero m-dimensional vector w we have $w^T X \sim N(w^T \mu, w^T \Sigma w)$.

Notice that, just as the univariate normal distribution is completely characterized by its mean and variance, and the bivariate normal distribution is characterized by means, variances, and a correlation, the multivariate normal distribution is completely characterized by its mean vector and variance matrix. In many cases the components of a multivariate normal random vector are treated separately, with each diagonal element of the covariance matrix furnishing a variance, and the off-diagonal elements to the degenerate random variable X for which $P(X = c) = 1$. We often write this as

$$X_n \xrightarrow{P} c.$$

The notion of convergence in probability is more general than the defintion above definition indicates, but we do not need the general definition. There are also two stronger notions of convergence, convergence in quadratic mean and convergence with probability one—but again we do not need these here.

Details: In applying convergence in probability, the criterion that is used is the following.

Theorem A sequence $X_1, X_2, \ldots$ converges in probability to c if and only if for every $\epsilon > 0$, $P(|X_n - c| > \epsilon) \to 0$ as $n \to \infty$.
Proof: This involves straightforward manipulations using the definition. The details are omitted. □

6.2 The Law of Large Numbers

6.2.1 As the sample size n increases, the sample mean converges to the theoretical mean.

The LLN is an accessible result, in the sense that its statement may be understood without advanced mathematics. The proof is not especially difficult, and we include it here, but we will regard it as an inessential detail.

Theorem: The Law of Large Numbers If $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables having a distribution with mean μ_X and standard deviation σ_X, then $\bar{X}$ converges in probability to μ_X, i.e.,

$$X_n \xrightarrow{P} \mu_X.$$

The form of the LLN given here is sometimes called the "weak" law of large numbers. The strong law instead says that convergence occurs with probability 1. However, considerably more machinery is needed in order to say this in precise mathematical terms. Intuitively, "with probability 1" means that the convergence is certain to occur.

Details: The proof will require the following lemma.

theory, and also why it seems to fit, at least crudely, so many observed phenomena. It says that whenever we average a large number of small independent effects, the result will be approximately normally distributed.

> *A detail*: Another way to interpret the CLT uses entropy, as defined in Eq. (4.33). Among all distributions having mean μ and standard deviation σ, the $N(\mu, \sigma^2)$ distribution is the most disorderly possible, in the sense of having maximal entropy. The CLT says that as the sample size gets very large the distribution of the sample mean becomes as disorderly as possible. This characterization provides an alternative way to understand and prove the CLT. See Madiman and Barron (2007).

There are also versions of the CLT for non-independent variables, though they are considerably more complicated. Those results typically require the sequence to be *stationary*, as defined on p. 515 of Chapter 18, and further limit the dependence among the random variables X_i and X_j within the sequence as $j - i$ increases. See Billingsley (1995, Theorem 27.4) and also Francq and Zakoian (2005).

SEE ATTACHED PAGE

$Cor(X_{ij}, X_{kj}) = \rho_{ik}$

6.3.2 *For large n, the multivariate sample mean is approximately multivariate normal.*

The multivariate version of the CLT is analogous to the univariate CLT. We begin with a set of multidimensional samples of size n: on the first variable we have a sample $X_{11}, X_{12}, \ldots, X_{1n}$, on the second, $X_{21}, X_{22}, \ldots, X_{2n}$, and so on. In this notation, X_{ij} is the jth observation on the ith variable. Suppose there are m variables in all, and suppose further that $E(X_{ij}) = \mu_i$, $V(X_{ij}) = \sigma_i^2$, and $Cor(X_{ij}, X_{ik}) = \rho_{jk}$ for all $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $k = 1, \ldots, m$. As before, let us collect the means into a vector μ and the variances and covariances into a matrix Σ. We assume, as usual, that the variables across different samples are independent. Here this means X_{ij} and X_{hk} are independent whenever $i \neq h$. The sample means

$$\bar{X}_1 = \frac{1}{n}\sum_{j=1}^{n} X_{1j}$$

$$\bar{X}_2 = \frac{1}{n}\sum_{j=1}^{n} X_{2j}$$

$$\vdots$$

$$\bar{X}_m = \frac{1}{n}\sum_{j=1}^{n} X_{mj}$$

for page 147

$$Cor(\mathbf{X}_{ij}, \mathbf{X}_{kj}) = \rho_{ik}$$

It took roughly 50 more years to refine the early concepts to its full-fledged modern incarnation and, in fact, new variants of algorithms continue to be developed so that it may be applied to ever more complicated situations. In contexts where finitely-many parameter values completely specify[1] the statistical model, implementation of ML estimation is conceptually straightforward while, from a theoretical perspective, ML estimation is also provably unbeatable—no other method offers better performance, for large samples. ML estimation has, therefore, become the dominant approach to parameter estimation. We will review basic properties and uses of ML estimation in Chapter 8.

In Section 7.3 we discuss confidence intervals. In Chapter 1, on p. 13, we described the use of a confidence interval to assess the uncertainty associated with responses of patient P.S. when forced repeatedly to choose between pictures of burning and non-burning houses; we noted that an approximate 95 % confidence interval for her propensity to choose the non-burning house was (.64, 1.0) and we concluded it was not very likely that she was choosing them with equal probabilities (a propensity of .5); instead, she apparently saw the two complete pictures without conscious awareness of processing their left ends, which is where the fire appeared. As a data-analytic tool, confidence intervals have become straightfoward to use in many, varied situations. We treat several simple yet important problems in Section 7.3 and supplement with more general methods in Chapters 8 and 9. As one thinks harder about interpretation, the subject gets somewhat more subtle. We review the issues in Sections 7.3.8 and 7.3.9. On the other hand, confidence intervals are fundamental to statistical practice and, from a contemporary standpoint, they seem very natural. Seen in historical context, the introduction of confidence intervals by Jerzy Neyman in the 1930s was quite ingenious, and a giant leap forward.

straightforward

r

One of the ways confidence intervals are found in conjunction with maximum likelihood is to apply the *bootstrap*, which is discussed in Chapter 9. As additional motivation for the discussion in this and subsequent chapters, here is a concrete example where these methods have been used in fitting a statistical model of mental processes.

Example 7.1 A Model of Visual Attention Experiments on visual attention often study the ability of subjects to see and remember multiple objects that are exposed to them for a very short time. Following Sperling (1967), Bundesen and colleagues developed a quantitative theory of visual attention (Bundesen, 1998) according to which, objects in the visual field are compared with representations in visual memory, and if the comparison is completed prior to the end of visual exposure, the object is recognized. In this theory the time taken to process and store an object identity is a random variable. For object i call this random variable X_i. The processing is considered to begin after a latency of length t_0, so that if t is the total time an object is displayed then the ith object is recognized if $X_i \leq t - t_0$. Bundesen assumed $X_i \sim Exp(\lambda_i)$. Letting $f_i(x)$ and $F_i(x)$ be the $Exp(\lambda_i)$ pdf and cdf, for exposure of length $x = t - t_0$, $F_i(t - t_0)$ is the probability of object recognition success

[1] From the point of view of the mathematical theory, a nonparametric method does not eliminate the parameters but rather makes them infinite dimensional.

$$\sigma_T = \sqrt{\frac{p(1-p)}{n}}. \tag{7.4}$$

The formula in Eq. (7.4) quantifies the variation we can associate with the observed proportion $\hat{p} = 14/17 = .824$. However, we can not compute a numerical value for σ_T from Eq. (7.4) because we do not know what value of p to use. The obvious solution is to substitute $\hat{p}$ for p in Eq. (7.4). When we do this we obtain the *standard error* for the binomial proportion

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{7.5}$$

Applying this to the data from P.S. we get

$$SE = \sqrt{\frac{\frac{14}{17}(1-\frac{14}{17})}{17}} = .092.$$

We then typically write the estimate in the form .824 ± .092, with the ± indicating that the likely variability in the estimate is .092. When, instead, we write $\hat{p} \pm 2SE$ we get the confidence interval (.64, 1.0), reported on p. 13. □

The general procedure for computing the standard error is, in essence, the same as in the binomial case. To emphasize the substitution of the estimated parameter for the unknown paramater we define the *standard error* of an estimator T to be of the form

parameter
e

$$SE(T) = \sqrt{\hat{V}(T)} \tag{7.6}$$

with the hat on V indicating that we have estimated the variance. In fact, definition (7.6) is very general in the sense that it does not specify *how* we estimate the variance. As we will see in Chapters 8 and 9, several different methods are used to obtain variance estimates. We have used T in (7.6) to emphasize that it is a random variable, but in an alternative notation we use more often we may rewrite (7.6) as

$$SE(\hat{\theta}) = \sqrt{\hat{V}(\hat{\theta})}.$$

One note on terminology: the term "standard error" is sometimes used to refer to the standard error of the mean, as in Eq. (7.17), which is a special case of (7.6).

It is very common practice to report an estimate together with its standard error in the form

$$\hat{\theta} \pm SE(\hat{\theta}).$$

This gives a simple, rough sense of how accurate the estimate is. A more refined statement, made in terms of probability, comes from the use of a confidence interval:

have written the subscript n on T to indicate that we are examining its behavior as $n \to \infty$. Two things drove the derivation of (7.22) above. First, the CLT was invoked to produce the approximate normality of $\bar{X}$ according to (7.20) and, second, in the standard deviation $\sqrt{\frac{p(1-p)}{n}}$, p was replaced by $\hat{p}$ (which was justified by the convergence of $\bar{X}$ to p in probability). If we assume these two phenomena apply, then we obtain (7.8) according to the following theorem.

remove parentheses $\frac{T_n - \theta}{\sigma_{T_n}} \xrightarrow{D} N(0,1)$

Theorem If T_n is an asymptotically normal estimator of θ satisfying

$$\frac{(T_n - \theta)}{\sigma_{T_n}} \xrightarrow{D} N(0, 1)$$

and $\hat{\sigma}_{T_n}$ satisfies

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then we have

$$\frac{(T_n - \theta)}{\hat{\sigma}_{T_n}} \xrightarrow{D} N(0, 1).$$

Proof: This follows by Slutsky's theorem (p. 163), as in the binomial case. □

We now re-state the theorem as a "result", by putting it in a form that is less precise mathematically but more useful in practice.

Result If T_n is an asymptotically normal estimator of θ satisfying

$$\frac{(T_n - \theta)}{\sigma_{T_n}} \xrightarrow{D} N(0, 1) \tag{7.23}$$

and $\hat{\sigma}_{T_n}$ provides the standard error of T_n in the sense that

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then

$$\text{approx. } 95\,\%\text{ CI} = (T_n - 2\hat{\sigma}_{T_n}, T_n + 2\hat{\sigma}_{T_n})$$

which may also be written, equivalently, in the form (7.8), i.e.,

$$\text{approx. } 95\,\%\text{ CI} = (\hat{\theta} - 2SE(\hat{\theta}), \hat{\theta} + 2SE(\hat{\theta})).$$

The key extra assumption is that the standard error tends to decrease as $\sqrt{n}$. This holds for many estimators, including MLEs (which follows from the discussion in Section 8.4.3). Let us suppose that SE_1 is based on a sample of size n_1 and we wish to determine the sample size n_2 that would give us SE_2. Because we want the standard error SE_1 to decrease by a factor SE_1/SE_2 (e.g., if we want SE_2 to be half the size of SE_1 we want to decrease SE_1 by a factor of 2), we write

$$\frac{SE_1}{SE_2} = \sqrt{\frac{n_2}{n_1}}$$

and solve for n_2, which gives

$$n_2 = n_1 \left(\frac{SE_1}{SE_2} \right)^2 . \tag{7.25}$$

If, for instance, we wanted to decrease the standard error by a factor of 2 we would have to multiply our current sample size by a factor of 4. This is just a restatement of the $\sqrt{n}$ decrease in the standard error, with (7.25) providing the explicit formula we would use to compute n_2 in practice.

Using confidence intervals, the simple rule[3] in Eq. (7.25) is about as far as we can go. An investigator may wonder about step one, the choice of the "desired" SE_2. The selection of SE_2 must be determined by careful thinking about the scientific issues involved in the particular case at hand. The desired size of the standard error in Example 3.4, p. 164, for instance, depends on the way the information about spike counts will be used as part of the overall project. In Example 3.4 a relatively large number of trials were collected because the experiment was part of a comparative study in which relatively small differences across conditions appeared possible—yet still would have been of interest. According to the standard error on p. 164, the firing rate was determined within about ±1 spike per second. If 15 trials had been used instead of 60, according to the $\sqrt{n}$ law and (7.25) we would expect an accuracy of ~~only~~ about ±~~1~~6 spikes per second, ~~and for a mean rate of around 20 spikes per second this seems to be a rather large uncertainty unless the neural response was drastically changed under the alternative condition. On the other hand, such statistical considerations always must be balanced against experimental constraints.~~ 2

which may or may not have seemed adequate.

[3] More complicated formulas exist; however, the uncertainties involved in replicating results when collecting more data are often much larger than any extra precision one might gain from a more detailed calculation.

of the table below give the resulting possible credible intervals using .025 and .975 quantiles of the $Beta(x+1, 17-x+1)$ distribution, labeled $q_{.025}$ and $q_{.975}$.

x	$q_{.025}$	$q_{.975}$	Cover
7	.22	.64	N
8	.26	.69	N
9	.31	.74	N
10	.36	.78	N
11	.41	.83	Y
12	.47	.87	Y
13	.52	.90	Y
14	.59	.94	Y
15	.65	.96	Y
16	.73	.99	Y
17	.81	1	N

We again suppose $p = .8$. From this table we find that the Bayesian credible intervals would cover the true value of $p = .8$ when $11 \leq x \leq 16$ (again indicated by "Y" for "yes" in the last column). To find the level of confidence associated with the credible intervals we compute $P(11 \leq X \leq 16)$ when $X \sim B(17, .8)$. We find $P(11 \leq X \leq 16) = .94$, which says that these credible intervals have probability .94 of containing the true value .8. This is very nearly equal to the desired value of .95, and is much closer to .95 than the value of .87 obtained on p. 171 for the approximate CI. The discrepancy between the putative value .95 and the correct coverage probability .87 for the approximate CI is due to the small sample size ($n = 17$). As the sample size gets large, the approximate 95 % CI found from (7.22) will have very nearly probability .95 of covering the true value of p. The Bayesian method performs better in this small-sample setting. When sample sizes are relatively small it is often possible to study coverage probabilities numerically in order to determine whether they are likely to be performing according to specifications, at least approximately. □

There are many important theoretical results concerning posterior distributions. In particular, the approximate CIs given by (7.22) have a Bayesian justification for large samples (see Section 8.3.3), making valid interpretation B of Section 7.3.8, which is re-phrased above. We return to Bayesian methods in Chapter 16.

7.3.10 *For small samples it is customary to use the t distribution instead of the normal.*

When the sample size is small, the approximation (7.18) may not be accurate. An alternative is to derive an "exact" confidence interval analgous to (7.14) that corrects for the substitution of s for σ. This leads to an adjustment of the multiplier put in front of the standard error. The adjustment to the small-sample CI uses the t distribution. Recall from Chapter 5 that if $U \sim N(0,1)$ and $V \sim \chi^2_\nu$ independently then

analogous

o

therefore, $E(2(T-\mu_T)(\mu_T-\theta))=0$. Thus, we have

$$E((T-\theta)^2)=E((T-\mu_T)^2)+(E(\mu_T-\theta))^2$$

and, since $V(T)=E((T-\mu_T)^2)$, we have proven the theorem. □

This decomposition of MSE into squared bias and variance terms is used in various contexts to "tune" estimators in an attempt to decrease MSE. This typically involves some increase in one term, either the squared bias term or the variance term, in order to gain a larger decrease in the other term. Thus, reduction of MSE is often said to involve a *bias variance trade-off.* For an example, see p. 434.

Before we present an illustration of a *MSE* calculation, let us mention a property of the sample mean and sample variance. Assuming they are computed from a random sample $X_1,\dots,X_n$, we have $E(\bar{X})=\mu_X$ which may be written

$$E(\bar{X})-\mu_X=0.$$

This says that, as an estimator of the theoretical mean, the sample mean has zero bias. When an estimator has zero bias it is called *unbiased.* If an estimator T is unbiased we have *MSE(T)* $=V(T)$ so that consideration of its performance may be based on a study of its variance.

In addition to the sample mean being unbiased as an esimator of the theoretical mean, it also happens that the *sample variance*, defined by

estimator

t

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2,$$

is unbiased as an estimator of the theoretical variance:

$$E(S^2)=\sigma_X^2. \tag{8.4}$$

Details: We wish to evaluate

$$E(S^2)=E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right).$$

We write $X_i-\bar{X}=(X_i-\mu_X)+(\mu_X-\bar{X})$ and expand the square

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2&=\sum_{i=1}^{n}\left((X_i-\mu_X)+(\mu_X-\bar{X})\right)^2\\&=\sum_{i=1}^{n}(X_i-\mu_X)^2+\sum_{i=1}^{n}2(X_i-\mu_X)(\mu_X-\bar{X})\end{aligned}$$

$$+\sum_{i=1}^{n}(\mu_X-\bar{X})^2.$$

We now rewrite the three terms in the last expression above. Because $E(X_i-\mu_X)^2=\sigma_X^2$, and the expectation of a sum is the sum of the expectations, the first term has expectation

$$E\left(\sum_{i=1}^{n}(X_i-\mu_X)^2\right)=n\sigma_X^2. \qquad (8.5)$$

Next, the second term may be rewritten

$$\begin{aligned}\sum_{i=1}^{n}2(X_i-\mu_X)(\mu_X-\bar{X}) &= 2(\mu_X-\bar{X})\sum_{i=1}^{n}(X_i-\mu_X)\\ &= -2(\bar{X}-\mu_X)\sum_{i=1}^{n}(X_i-\mu_X)\\ &= -2n(\bar{X}-\mu_X)^2,\end{aligned}$$

where the last equality uses $\sum_{i=1}^{n}(X_i-\mu_X)=n(\bar{X}-\mu_X)$, and then, because $E((\bar{X}-\mu_X)^2)=V(\bar{X})=\sigma_X^2/n$, the expectation of the second term becomes

$$E\left(\sum_{i=1}^{n}2(X_i-\mu_X)(\mu_X-\bar{X})\right)=-2\sigma_X^2. \qquad (8.6)$$

Finally, because again, $E((\bar{X}-\mu_X)^2)=\sigma_X^2/n$, the expectation of the third term is

$$E\left(\sum_{i=1}^{n}(\mu_X-\bar{X})^2)\right)=\sigma_X^2 \qquad (8.7)$$

and, combining (8.5), (8.6), and (8.7) we get

$$E\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=(n-1)\sigma_X^2$$

which gives (8.4). □

unbiasedness

We use the unbiasedness of the sample mean and sample variance in the following illustration of the way two estimators may be compared theoretically.

Illustration: Poisson Spike Counts On p. 164 we considered 60 spike counts from a motor cortical neuron and found an approximate 95 % CI for the resulting firing rate using the sample mean. The justification for that approximate CI involved the CLT, and the practical implication was that as long as the sample size is fairly large, and the distribution not too far from normal, the CI would have approximately .95 probability of covering the theoretical mean. In this case, the spike counts do, indeed, appear not too far from normal. Sometimes they are assumed to be Poisson distributed. This is questionable because careful examination of spike trains almost always indicates some departure from the Poisson. On the other hand, the departure is sometimes not large enough to make a practical difference to results. In any case, for the sake of illustrating the *MSE* calculation, let us now *assume* the counts follow a Poisson distribution with mean λ. The sample mean $\bar{X}$ is a reasonable estimator of λ, but one might dream up alternatives. For example, a property of the Poisson distribution is that its variance is also equal to λ; therefore, the sample variance S^2 could also be used to estimate the theoretical variance λ. This may seem odd, and potentially inferior, on intuitive grounds because the whole point is to estimate the mean firing rate, not the variance of the firing rate. On the other hand, once we take the Poisson model seriously the theoretical mean and variance become equal and, from a statistical point of view, it is reasonable to ask whether it is better to estimate one rather than the other from their sample analogues. Our purpose here is to present a simple analysis that demonstrates the inferiority of the sample variance compared with the sample mean as an estimator of the Poisson mean λ. We are going through this exercise so that we can draw an analogy to it later on.

departure

Now, because, as we mentioned immediately before beginning this illustration, $\bar{X}$ and S^2 are unbiased for the theoretical mean and variance they are, in this case, both unbiased as estimators of λ. As a consequence, $MSE(T) = V(T)$ for both $T = \bar{X}$ and $T = S^2$. Analytical calculation of the variance of these estimators (which we omit here) gives

$$\begin{aligned} V(\bar{X}) &= \frac{\lambda}{n} \\ V(S^2) &= \frac{\lambda}{n} + \frac{2\lambda^2}{n-1} \end{aligned}$$

where n is the number of counts (the number of trials). Therefore, the *MSE* of S^2 is always larger than that of $\bar{X}$ so that S^2 tends to be further from the correct value of λ than $\bar{X}$. For example, if we take $n = 100$ trials and $\lambda = 10$, we find $V(\bar{X}) = .10$ while $V(S^2) = 2.12$. The estimator S^2 has about 21 times the variability as $\bar{X}$, so that estimating λ using S^2 would require about 2,100 trials of data to gain the same accuracy as using $\bar{X}$ with 100 trials. Figure 8.2 shows a pair of histograms of $\bar{X}$ and S^2 values calculated from 1,000 randomly-generated samples of size $n = 100$ when the true Poisson mean was $\lambda = 10$. The distribution represented by the histogram on the right is much wider. □

This illustration nicely shows how one method of estimation can be very much better than another, but it is admittedly somewhat artificial; because the distribution of real spike counts may well depart from Poisson, a careful comparison of $\bar{X}$ versus then MSE is the risk function

$$MSE(T) = E\left(L(d(X_1, \ldots, X_n), \theta)\right).$$

This terminology, viewing MSE as "risk under squared-error loss," is quite common.

8.2 Estimation in Large Samples

8.2.1 *In large samples, an estimator should be very likely to be close to its estimand.*

In the introduction to this chapter we offered the reminder that the sample mean satisfies

$$\bar{X} \xrightarrow{P} \theta$$

which is the law of large numbers. Suppose T_n is an estimator of θ. If, as $n \to \infty$, we have

$$T_n \xrightarrow{P} \theta \tag{8.14}$$

then T_n is said to be a *consistent* estimator of θ. This means that for every positive ϵ, as $n \to \infty$ we have

$$P(|T_n - \theta| > \epsilon) \to 0.$$

Note that, by (8.3), if $MSE(T_n) \to 0$ then T_n is consistent. Also, if T_n satisfies (8.1) and $\sigma_{T_n} \to 0$ then T_n is consistent.

Slutsky's S

> *Details:* Multiplying the left-hand side of (8.1) by σ_{T_n} and applying Slutzky's theorem we have $T_n - \theta \xrightarrow{P} 0$, which is equivalent to $T_n \xrightarrow{P} \theta$. □

In words, to say that an estimator is consistent is to say that, for sufficiently large samples, it will be very likely to be close to the quantity it is estimating. This is clearly a desirable property. When T_n satisfies (8.1) and $\sigma_{T_n} \to 0$ we will call T_n *consistent and asymptotically normal.*

8.2.2 *In large samples, the precision with which a parameter may be estimated is bounded by fisher information.*

cap F

Fisher

Let us consider all estimators of θ that are consistent and asymptotically normal in the sense of Section 8.2.1. For such an estimator $T = T_n$ we may say that its distribution for all n. When this happens the estimator contains all of the information about θ that is available in the data, and it is called a *sufficient statistic*. For instance, if we have a sample from a $N(\mu, \sigma^2)$ distribution with σ known, then the sample mean $\bar{X}$ is sufficient for estimating μ. Sufficiency may be characterized in many ways. If T is a sufficient statistic, then the likelihood function based on T is the same as the likelihood function based on the entire sample. For example, it is not hard to verify that the likelihood function based on a sample $(x_1, \ldots, x_n)$ from a $N(\mu, \sigma^2)$ distribution with σ known is the same as the likelihood function based on $\bar{X}$. This property is sometimes known as *Bayesian sufficiency* (see Bickel and Doksum (2001)). In addition, if θ is given a prior distribution as in Section 7.3.9, then T is sufficient when the mutual information between θ and T is equal to the mutual information between θ and the whole sample (see Cover and Thomas (1991)). Parametrized families of distributions for which it is possible to find a sufficient statistic with the same dimension as the parameter vector are called *exponential families*. See Section 14.1.6. □

A related result is the following. If we let $\psi(\theta) = E(T)$, where the expectation is based on a random sample from the distribution with pdf $f(x|\theta)$, it may be shown[6] that

$$V(T) \geq \frac{\psi'(\theta)^{2}}{I(\theta)}.$$

$\psi'(\theta)^2$

Therefore, if T is an unbiased estimator of θ based on a random sample from the distribution with pdf $f(x|\theta)$ we have $\psi'(\theta) = 1$ and

$$V(T) \geq \frac{1}{I(\theta)}. \tag{8.24}$$

This is usually called the *Cramér-Rao lower bound.* Although Eq. (8.24) is of less practical importance than the asymptotic result (8.23), authors often speak of the bound in (8.23) as a Cramér-Rao lower bound.

Fisher information also arises in theoretical neuroscience, particularly in discussion of neural decoding and optimal properties of tuning curves (see Dayan and Abbott (2001)).

8.2.3 *Estimators that minimize large-sample variance are called efficient.*

A consistent and asymptotically normal estimator T satisfies (8.1) and it also satisfies (8.22). In (8.1) we suppressed the dependence of T and σ_T on n. The information $I^T(\theta)$ also depends on n, as does $I(\theta)$. We now consider what happens as $n \to \infty$.

[6] See Bickel and Doksum (2001, Chapter 3).

In other words, instead of the expected information evaluated at $\hat{\theta}$ in (8.33), we use the negative second derivative of the loglikelihood, evaluated at $\hat{\theta}$, without[8] any expectation. Again, under certain conditions, we have

$$\sqrt{I_{OBS}(\hat{\theta})}(\hat{\theta}-\theta) \xrightarrow{D} N(0,1). \tag{8.35}$$

Details: Note that

$$-\frac{1}{n}\ell''(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f(x_i|\theta)$$

and that the expectation of the right-hand side is $I_F(\theta)$. From the LLN we therefore have

$$-\frac{1}{n}\ell''(\theta) \xrightarrow{P} I_F(\theta),$$

and it may also be shown that

$$\sqrt{\frac{I_{OBS}(\hat{\theta})}{I(\hat{\theta})}} \xrightarrow{P} 1,$$

Slutsky's

which, again by Slutzky's Theorem, gives (8.35). □

S

Equation (8.35) provides large-sample standard errors and confidence intervals based on ML estimation, given in the following result.

Result For large samples, under certain general conditions, the MLE $\hat{\theta}$ satisfies (8.35), so that its standard error is given by

$$SE = \frac{1}{\sqrt{-\ell''(\hat{\theta})}} \tag{8.36}$$

and an approximate 95 % CI for θ is given by $(\hat{\theta}-2SE, \hat{\theta}+2SE)$.

Additional insight about the observed information can be gained by returning to the derivation of (8.17) and applying it, instead, to the likelihood function based on a sample $x_1, \ldots, x_n$ from a $N(\mu, \sigma^2)$ distribution with σ known, as in Section 7.3.2. There, we found the loglikelihood function to be

[8] For the special class of models known as exponential families, which are used with the generalized linear models discussed in Chapter 14, we have $I(\hat{\theta}) = I_{OBS}(\hat{\theta})$ (see, e.g., Kass and Vos (1997)) but this is not true in general.

$$\ell(\theta) = -\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}$$

which simplified to Eq. (7.2),

$$\ell(\theta) = -\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta).$$

Differentiating this twice we get

$$\ell''(\theta) = -\frac{n}{\sigma^2},$$

so that

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{-\ell''(\theta)}}. \tag{8.37}$$

In other words, $1/\sqrt{-\ell''(\theta)}$ gives the standard error of the mean in that case.

Quite generally, for large samples, the likelihood function has an approximately normal form and there is a strong analogy with this paradgm case. Specifically, a quadratic approximation to the loglikelihood function (using a second-order Taylor expansion) produces a normal likelihood (because if $Q(\theta)$ is quadratic then $\exp(Q(\theta))$ is proportional to a normal likelihood function) and in this normal likelihood the value of the standard deviation is $1/\sqrt{-\ell''(\hat{\theta})}$. This heuristic helps explain (8.36).

paradigm

Details: The quadratic approximation to $\ell(\theta)$ at $\hat{\theta}$ is

$$Q(\theta) = \ell(\hat{\theta}) + \ell'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2.$$

Using $\ell'(\hat{\theta}) = 0$ and setting $c = \exp(\ell(\hat{\theta}))$ we have

$$\exp(Q(\theta)) = c\exp\left(-\frac{1}{2}(-\ell(\hat{\theta}))(\hat{\theta} - \theta)^2\right). \tag{8.38}$$

The function on the right-hand side of (8.38) has the form of a likelihood function based on $X \sim N(\theta, \sigma^2)$ where $\hat{\theta}$ plays the role of x and $\sigma = 1/\sqrt{-\ell''(\hat{\theta})}$. □

We now consider two simple illustrations.

Illustration: Exponential distribution Suppose $X_i \sim Exp(\lambda)$ for $i = 1, \ldots, n$, independently. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda \sum x_i} \\ &= \lambda^n e^{-\lambda n \bar{x}} \end{aligned}$$

and the loglikelihood function is

$$\ell(\lambda) = n \log \lambda - n\lambda\bar{x}.$$

Differentiating this and setting equal to zero gives

$$0 = n(\frac{1}{\lambda} - \bar{x})$$

and solving this for λ yields the MLE

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

Continuing, we compute the observed information:

$$\begin{aligned} -\ell''(\hat{\lambda}) &= \frac{n}{\hat{\lambda}^2} \\ &= n\bar{x}^2 \end{aligned}$$

which gives us the large-sample standard error

$$SE(\hat{\lambda}) = \frac{1}{\bar{x}\sqrt{n}}.$$ □

Illustration: Binomial For a $B(n, p)$ random variable it is straightfoward to obtain the observed information

$$-\ell''(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}.$$

straightforward

r

This gives

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

which is the same as the SE found in Section 7.3.5. Therefore, the approximate 95 % CI in (7.22) is an instance of that provided by ML estimation with SE given by (8.36).

□

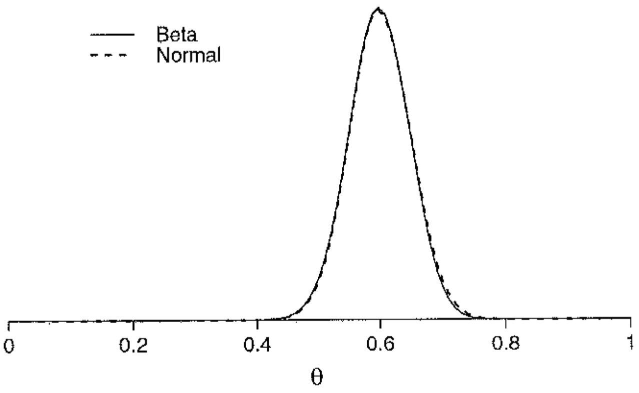


Fig. 8.8 Normal approximation $N(.6, (.049)^2)$ to beta posterior $Beta(61, 41)$.

Cap B Bayesian

8.3.3 *In large samples, ML estimation is approximately bayesian.*

In Section 7.3.9 we said that Bayes' theorem may be used to provide a form of estimation based on the posterior distribution according to (7.28), i.e.,

$$f_{\theta|x}(\theta|x) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}.$$

One of the most important results in theoretical statistics is the approximate large-sample equivalence of inference based on ML and inference using Bayes' theorem.

Result For large samples, under certain general conditions, the posterior distribution of θ is approximately normal with mean given by the MLE $\hat{\theta}$ and standard deviation given by the standard error formula (8.36).

We elaborate in Section 16.1.5 and content ourselves here with a simple illustration.

Illustration: Binomial distribution Suppose $Y \sim B(n, \theta)$ with $n = 100$ and we observe $y = 60$. As we said in Section 7.3.9, if we take the prior distribution on θ to be $U(0, 1)$, which is also the $Beta(1, 1)$ distribution, we obtain a $Beta(61, 41)$ posterior. The observed proportion is the MLE $\hat{\theta} = x/n = .6$. The usual standard error then becomes $SE = \sqrt{\hat{\theta}(1-\hat{\theta})/n} = .049$. As shown in Fig. 8.8 the normal distribution with mean $\hat{\theta}$ and standard deviation $\sqrt{\hat{\theta}(1-\hat{\theta})/n}$ is a remarkably good approximation to the posterior. □

For the data from subject P.S. in Example 1.4, which involves a relatively small sample, we already noted (see p. 174) that the approximate 95 % confidence interval (.64, 1.0) found using (8.36) (which is the same as (7.22), see p. 206) differed by

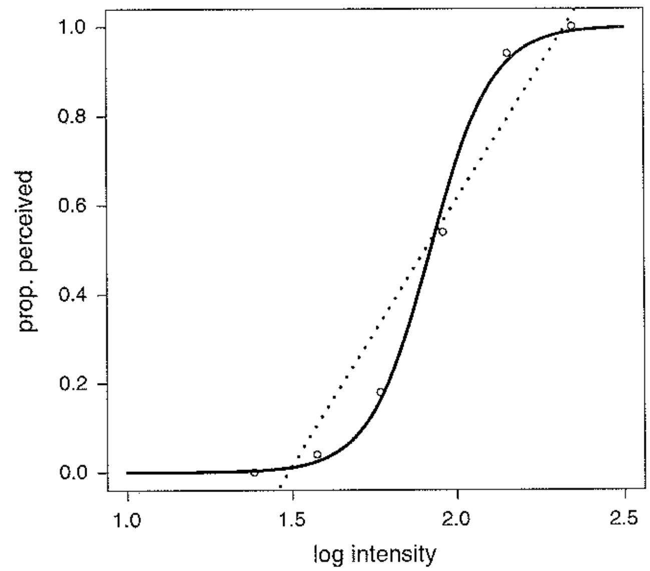


Fig. 8.9 Proportion of trials, out of 50, on which light flashes were perceived by subject S.S. as a function of $\log_{10}$ intensity, together with fits. Data from Hecht et al. (first series of trials) are shown as *circles*. *Dashed line* is the fit obtained by linear regression. *Solid curve* is the fit obtained by logistic regression.

to one as $x \to \infty$. The fitted curve in Fig. 8.9 is based on the following statistical model: for the *i*-th value of light intensity we let Y_i be the number of light flashes on which the subject perceives light and then take

ith

italic i

remove hyphen

$$Y_i \sim B(n_i, p_i) \tag{8.43}$$

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \tag{8.44}$$

This is known as the *logistic regression model*. There are many possible approaches to estimating the parameter vector $\theta = (\beta_0, \beta_1)$ but the usual solution is to apply maximum likelihood. The observed information matrix is then used to get standard errors of the coefficients. These calculations are performed by most statistical software packages. For the data in Fig. 8.9 we obtained $\hat{\beta}_0 = -20.5 \pm 2.4$ and $\hat{\beta}_1 = 10.7 \pm 1.2$. Further discussion of logistic regression, and interpretation of this result, are given in Section 14.1. □

8.4.4 *When using numerical methods to implement ML estimation, some care is needed.*

There are three issues surrounding the application of numerical maximization to ML estimation. The first is that, while loglikelihood functions are usually well behaved

4. Increment j and return to Step 2.
5. Repeat Steps 2–4 until convergence. □

A key step in formulating the EM algorithm in the mixture of two Gaussians model, above, was the introduction of the random variables W_i. In order to maximize the loglikelihood $\ell_Y(\theta)$ defined by the pdf $f_Y(y|\theta)$ we effectively introduced the loglikelihood $\ell_{(Y,Z)}(\theta)$ in (8.46) based on the augmented data pdf $f_{(Y,Z)}(y, z|\theta)$. Step 2 of the algorithm, known as *the expectation step*, is based on the expectation $E(\ell_{(Y,Z)}(\theta)|\boxed{Z = z}, \theta = \theta^{(j)})$. In Step 2 the conditional expectation in (8.49) was evaluated for $\theta = \theta^{(j)}$. In Step 3 the loglikelihood was maximized in terms of the expectations computed in Step 2.

$Y=y$

SEE ATTACHED SHEET (A)

In general, if $Y = y$ is the data vector augmented by $Z = z$ we define

$$Q(\theta, \theta^{(j)}) = E(\ell_{(Y,Z)}(\theta)|\boxed{Z = z}, \theta = \theta^{(j)}). \tag{8.50}$$

Beginning with an initial guess $\theta^{(1)}$, for each j the EM algorithm computes $Q(\theta, \theta^{(j)})$ and sets $\theta^{(j+1)}$ equal to the maximizer of $Q(\theta, \theta^{(j)})$ as a function of θ. The EM algorithm works well for problems in which some kind of data augmentation greatly simplifies the problem, so that $Q(\theta, \theta^{(j)})$ is easy to compute (as in Step 2 of the mixture of two Gaussians illustration above). In addition to models that incorporate latent variables, the EM algorithm is often applied to problems with missing data, where the missing data are treated as augmenting the observed data. (See also the related discussion of Gibbs sampling in Section 16.2.2.)

One way to see that this iterative scheme should work is to apply the formula[11]

$$\frac{d}{d\theta}Q(\theta, \theta^*)|_{\theta=\theta*} = \ell'_Y(\theta^*) \tag{8.51}$$

(see the details below). If $\theta^{(1)}, \theta^{(2)}, \ldots$ is a sequence of EM iterates that converge to a value θ^* then, because each iterate maximizes $Q(\theta, \theta^{(j)})$ its derivative is 0, i.e.,

$$\frac{d}{d\theta}Q(\theta, \theta^*)|_{\theta=\theta*} = 0.$$

From (8.51) we then have

$$\ell'_Y(\theta^*) = 0.$$

Thus, for sufficiently good initial values, when the EM algorithm converges to θ^* we get $\theta^* = \hat{\theta}$, i.e., the EM algorithm converges to the MLE $\hat{\theta}$.

Details: We derive Eq. (8.51). From (8.50) we have

$$\frac{f(y,z|\theta^*)}{f(y|\theta^*)}$$

$$Q(\theta, \theta^*) = \int \boxed{f(y, z|Y = y, \theta^*)} \log f(y, z|\theta)dz.$$

SEE ATTACHED SHEET (B)

[11] This formula was used by Fisher, in his discussion of sufficiency, to substantiate the argument mentioned in Section 8.2.2 (see p. 200 and Kass and Vos (1997, Section 2.5.1))

FOR PAGE 218 Ⓐ

$$E(\ell_{(\mathrm{Y},\mathrm{Z})}(\theta)|Y = y, \theta = \theta^{(j)})$$

FOR PAGE 218 Ⓑ

$$Q(\theta, \theta^*) \quad = \quad \int \frac{f(y, z|\theta^*)}{f(y|\theta^*)} \log f(y, z|\theta) dz$$

Result: Simulation-Based Propagation of Uncertainty in Estimation Suppose the random vector X is a consistent estimator of a parameter vector θ having an approximate distribution from which we are able to simulate observations and we wish to estimate $\phi = f(\theta)$ for some real-valued function $f(x)$. If we apply simulation-based propagation of uncertainty, with G large, then an approximate 95 % CI for ϕ is given by $(w_{.025}, w_{.975})$ where $w_{.025}$ and $w_{.975}$ are the .025 and .975 quantiles among the pseudo-data $W^{(1)}, W^{(2)}, \ldots, W^{(G)}$.

The beauty of this simulation-based method of getting approximate confidence intervals is its simplicity and practicality, as long as it is easy to generate observations from the distribution of the esimator X. If, in addition, the estimator $\hat{\phi} = f(\hat{\theta})$ is approximately normal, then we have a slightly different option. Although it will often produce essentially the same answers, it simplifies the reporting of results by producing a standard error, which is connected to the confidence interval by the 95 % rule (p. 117).

estimator

t

Result: Simulation-Based Propagation of Uncertainty in Estimation When the Estimator is Approximately Normal Suppose X is an approximately multivariate normal estimator of θ having estimated variance matrix $\hat{\Sigma}$, and we want to estimate $\phi = f(\theta)$ for some real-valued (univariate) function $f(x)$. Let us take $Y = f(X)$ to be the estimator of ϕ. We will write the observed estimate of θ as $X = \hat{\theta}$ and the observed estimate of ϕ as $Y = \hat{\phi} = f(\hat{\theta})$. If the function $f(x)$ is approximately linear near $x = \hat{\theta}$ and $f'(\hat{\theta})$ is not the zero vector (i.e., not all of its partial derivatives are zero) then

1. Y is approximately normally distributed, and
2. the standard error obtained from (9.6) by simulation-based propagation of uncertainty

$$SE(\hat{\phi}) = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(W^{(g)} - \overline{W})^2} \tag{9.7}$$

furnishes approximate inferences. In particular, an approximate 95 % CI is given by $(Y - 2SE(Y), Y + 2SE(Y))$. □

If these two methods differ, it is an indication that the distribution of $\hat{\phi}$ is noticeably non-normal and it is better to use the quantiles as they are likely to be more accurate. The second method, based on approximate normality, is justified by the theorem on p. 235 leading to (9.20).

We illustrate both methods by returning to the example involving perception of dim light.

Example 5.5 (continued from p. 221) At the beginning of the chapter we motivated propagation of uncertainty using the problem of calculating x_{50}, defined on p. 221,

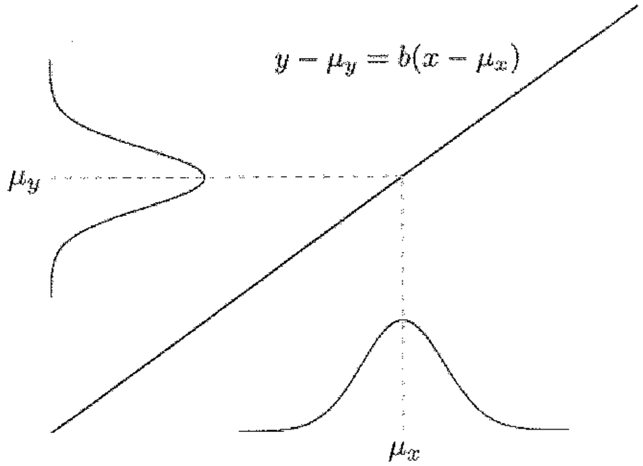


Fig. 9.1 The effect of the transformation $y = a + bx$ operating on a normally distributed random variable X having mean μ_X and standard deviation σ_X. The random variable $Y = a + bX$ is again normally distributed, with mean $\mu_Y = a + b\mu_X$ and standard deviation $\sigma_Y = |b|\sigma_X$. The normal distributions are displayed on the x and y axes; the linear transformation is displayed as a *line*, which passes through the point (μ_X, μ_Y) so that it may be written, equivalently, as $y - \mu_Y = b(x - \mu_X)$.

and the approximate mean and variance of Y is given from the approximating linear transformation, as in the theorem on p. 63.

Theorem Suppose that a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ satisfies

$$\frac{X_n - \mu}{\sigma_{X_n}} \xrightarrow{D} N(0, 1)$$

as $n \to \infty$, and that the function $f(x)$ is continuously differentiable with $f'(\mu) \neq 0$. Then

$$\frac{f(X_n) - f(\mu)}{\sigma_{Y_n}} \xrightarrow{D} N(0, 1)$$

with $\sigma_{Y_n} = |f'(\mu)|\sigma_{X_n}$.

Proof: We omit the proof, which is a consequence of Slutzky's theorem (p. 163), but give the essential idea.

Slutsky
S

First, from the theorem on transformation of a normal random variable (p. 63), if $Y = a + bX$ and $X \sim N(\mu_X, \sigma_X^2)$ then $Y \sim N(\mu_Y, \sigma_Y^2)$ with $\mu_Y = a + b\mu_X$ and $\sigma_Y = |b|\sigma_X$. A pictorial display of this situation is given in Fig. 9.1. Now, suppose that $f(x)$ is not linear, but let us assume that it is only mildly nonlinear within the "most probable" range of X. That is, $f(x)$ is mildly nonlinear within, say, $\mu_X \pm 2.5\sigma_X$, which is the range over which we are assuming X to be approximately normally distributed. Then we may approximate $f(x)$ with the best-fitting linear approximation at $x = \mu_X$:

is based on Behseta et al. (2009), which provides some details about propagation of error for the difference index. □

Additional details: We may also propagate uncertainy analytically to $x_{50} = f(\beta_0, \beta_1)$ using Eq. (9.19) with (9.10), which gives the standard error

uncertainty

t

$$SE = \sqrt{f'(\hat{\beta}_0, \hat{\beta}_1)^T \hat{\Sigma} f'(\hat{\beta}_0, \hat{\beta}_1)}$$

where the partial derivates are

$$\frac{\partial f}{\partial \beta_0} \Big|_{(\hat{\beta}_0, \hat{\beta}_1)} = -\frac{1}{\hat{\beta}_1}$$

$$\frac{\partial f}{\partial \beta_1} \Big|_{(\hat{\beta}_0, \hat{\beta}_1)} = \frac{\hat{\beta}_0}{\hat{\beta}_1^2}.$$

Plugging into the formulas above the values of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\Sigma}$, the $\log_{10}$ intensity at which subject S.S. would have perceived half the flashes is estimated to have been $\hat{x}_{50} = 1.921 \pm .019$. This agrees with the approximate 95 % CI obtained by the simulation method. □

9.2 The Bootstrap

The *bootstrap* is a very simple way to obtain standard errors and confidence intervals. It has turned out to be one of the great inventions in the field of statistics. In Section 9.2.1 we explain the essential idea, and we contrast the *parametric bootstrap* with the *nonparametric bootstrap*, elaborating on these two distinct methods in Sections 9.2.2 and 9.2.3.

9.2.1 The bootstrap is a general method of assessing uncertainty.

The algorithm for simulation-based propagation of uncertainty (p. 225) began with a random vector X having a known distribution (from which observations could be generated on the computer). In practice, applying the result on p. 226, X becomes an estimator of a parameter vector θ and its distribution is known approximately; typically it is a normal distribution. From this, uncertainty can be propagated from X to an estimator $\hat{\phi}$ of $\phi = f(\theta)$. As illustrated in Example 5.5 on p. 226, an essential input to the algorithm is the variance matrix of X (in Example 5.5 we had $X = (\hat{\beta}_0, \hat{\beta}_1)$ and used $\hat{\Sigma} = I_{OBS}(\hat{\beta}_0, \hat{\beta}_1)^{-1}$). But what if it is difficult to compute the variance matrix of X? The bootstrap instead backs up a step, using the variation in the data which is the same as (7.19). The idea of the bootstrap is analogous: we replace F_X by an estimate of it and then apply the algorithm above. If we have a parametric model and we use ML estimation to estimate the parameters, we can use the model with the fitted parameters to generate the pseudo-data $U_1^{(g)}, \ldots, U_n^{(g)}$. This scheme is called the *parametric bootstrap.* Otherwise, we replace F_X by the empirical cdf $\hat{F}_n$ and draw the pseudo-data $U_1^{(g)}, \ldots, U_n^{(g)}$ from $\hat{F}_n$. This is the *nonparametric bootstrap.* Both methods extend to cases in which we replace scalar estimates (e.g., $\hat{\beta}_1$) by vectors of estimated quantities (e.g., $(\hat{\beta}_0, \hat{\beta}_1)$).

The parametric bootstrap and nonparametric bootstrap both begin, conceptually, by estimating the data distribution F_X. The parametric bootstrap uses a specific assumption, such as normality of the data. The nonparametric bootstrap does not require any specific data distributional assumption, and this is the sense in which it is "nonparametric." The nonparametric bootstrap is also usually easier to implement. Its disadvantage is that it requires i.i.d. random variables to represent the variation in the data. There are many cases where the data are not modeled as i.i.d., such as in regression, time series, and point processes. Sometimes a clever transformation makes the nonparametric bootstrap applicable (see Davison and Hinkely 1997, for examples), but in other cases the parametric bootstrap is either the only available approach or at least a more straightforward methodology to apply. Both forms of bootstrap use propagation of uncertainty.

Hinkley

9.2.2 *The parametric bootstrap draws pseudo-data from an estimated parametric distribution.*

Suppose we assume that a set of data $x_1, x_2, \ldots, x_n$ is a random sample from a distribution with pdf $f(x_i|\theta)$, and we estimate θ with the MLE $\hat{\theta}$. If we assume for the moment that the parameter θ is a scalar then, according to the scheme in Section 9.2.1, we may obtain the standard error of $\hat{\theta}$ as $SE_{sim}(\hat{\theta})$ by generating pseudo-samples $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from the distribution with pdf $f(x_i|\theta)$. Because we do not know the value of θ we plug in the MLE $\hat{\theta}$ and instead generate pseudo-samples from the distribution with pdf $f(x_i|\hat{\theta})$. This is a *parametric bootstrap*, and the resulting value of $SE_{sim}(\hat{\theta})$ is a *parametric bootstrap* standard error.

Algorithm: Parametric bootstrap estimate of standard error To obtain the standard error $SE(\hat{\theta})$ we proceed as follows:

```
1. For $g = 1$ to $G$
   Generate a random sample $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ from the distribution having pdf $f(x_i|\hat{\theta})$.
   Find the MLE $\hat{\theta}^{(g)}$ based on $U_1^{(g)}, U_2^{(g)}, \ldots, U_n^{(g)}$ and set $W^{(g)} = \hat{\theta}^{(g)}$.
2. Compute $\overline{W} = \frac{1}{G}\sum_{i=1}^{G} W^{(g)}$ and then
```

$$\hat{V} = \hat{V}(f_1(\hat{\theta}), f_2(\hat{\theta}), \ldots, f_k(\hat{\theta})),$$

by following step 1, above, for each of $f_1(\theta), f_2(\theta), \ldots, f_k(\theta)$ to get

$$W_j^{(g)} = f_j(\hat{\theta}^{(g)})$$

for $j = 1, \ldots, k$, and then setting $\hat{V}$ equal to the sample variance matrix (see p. 90) of the k-dimensional vectors $W^{(g)} = (W_1^{(g)}, \ldots, W_k^{(g)})$. □

Example 8.2 (continued from p. 193) In discussing the way previous seizures affect the relationship between spike width and preceding inter-spike interval length we displayed results based on change-point models. The statistical model assumed that, on average, Y decreases quadratically with x for $x < \tau$ but remains constant for $x \geq \tau$, with τ being the change point. In Fig. 8.7 we displayed fitted change-points together with standard errors, which led to the conclusion that the seizure group reset to baseline average spike widths earlier than the control group. We said that the standard errors shown in Fig. 8.7 were based on a parametric bootstrap. The specifics of computing the bootstrap standard errors followed the steps given above: based on the fitted $\hat{\tau}$, together with the fitted parameters for the quadratic relationship when $x < \tau$ and the constant relationship when $x \geq \tau$ (see p. 408), pseudo-data samples were generated and for the gth such sample a value $\hat{\tau}^{(g)}$ was calculated following the same procedure that had been used with the real data; then formula (9.25) was applied. □

There are modifications of the bootstrap confidence interval procedure that offer improvements. These are reviewed by DiCiccio and Efron (1996). Particularly effective[3] are the *bias-corrected and accelerated* (or BC_a) intervals, which are often used as defaults in bootstrap software.

9.2.3 *The nonparametric bootstrap draws pseudo-data from the empirical cdf.*

In Section 9.2.2 we showed how the parametric bootstrap is used to get standard errors and confidence intervals. The key theoretical point was captured by Eq. (9.24), which says that, for large samples, the distribution of the pseudo-data based on the MLE plug-in estimate will be close to the distribution of the data. The idea of the nonparametric bootstrap is to generate pseudo-data, instead, from the empirical cdf

[3] The bootstrap approximate 95 % CI based on percentiles in Eq. (9.26) has the property that as $n \to \infty$ the probability of coverage is $.95 + \eta_n$ where η_n vanishes at the rate of $1/\sqrt{n}$. The BC_a intervals have the analgous property with η_n vanishing at the rate $1/n$, which means the theoretical coverage probability should be closer to .95.

analogous

o

The nonparametric bootstrap has been studied extensively, and has been shown to work well in a variety of theoretical and empirical senses. For more information about the bootstrap, see Efron and Tibshirani (1993) and Davison and Hinkley(1997).

An important caveat is that arbitrary shuffles of the data do not necessarily produce bootstrap samples. The key assumption is *independent and identically distributed* sampling of $X_1, \ldots, X_n$, so that the key result (9.27) applies. Many problems may be put in this form, but the nonparametric bootstrap only applies once they are.

9.3 Discussion of Alternative Methods

At the beginning of this chapter we considered the data on perception of dim light to illustrate propagation of uncertainty according to the diagram in (9.4). We went on to discuss analytical propagation of uncertainty, simulation-based propagation of uncertainty, and then both the parametric and non-parametric bootstrap methods of obtaining uncertainty about the target estimand, in this case x_{50}, the intensity at which a flash of light is perceived 50 % of the time.

The choice among these methods is largely a matter of convenience. It is often easy to obtain the variance matrix of the parameter MLEs and then simulation-based propagation of uncertainty is easy to implement. Sometimes it is also easy to get the derivatives analytically, and the analytical approach becomes an option. The percentile method of getting confidence intervals from simulation becomes more accurate than that based on $\pm 2SE$ when the nonlinearity in the target estimand as a function of the parameters is pronounced (relative to the uncertainty in the parameters, as explained in Section9.1.2). With i.i.d. data the nonparametric bootstrap is very easy to apply, and is often the preferred method. But many examples involve non-i.i.d. data. In regression or time series contexts, for instance, nonparametric bootstrap methods require modification and may be difficult or impossible to apply (this is the case for some point process models of neural spike train data). In such settings the parametric bootstrap is often used.

These methods can produce valid 95 % confidence intervals, which cover the estimand 95 % of the time, when the statistical model is correct and the sample size is sufficiently large. The statistical model used with the nonparametric bootstrap, in the form we have presented, assumes i.i.d. sampling but is otherwise very general. All of the methods aim to provide an appropriate spread of the confidence interval about the estimate, which is what leads to the correct coverage probability. The bias in the estimator is ignored because, for sufficiently large samples, it becomes vanishingly small. Furthermore, as we noted in Chapter8, the bias squared often becomes vanishingly small faster than the variance becomes vanishingly small, so that the MSE is dominated by the variance. In practice, however, it is worth remembering that nontrivial bias in the estimator can greatly diminish the coverage probability of a putatively 95 % confidence interval. If a statistical model isgrossly incorrect

space

Result: Suppose $X_1, \ldots, X_n$ has joint pdf $f(x_1, \ldots x_n|\theta)$, with θ a scalar, and suppose further that T_n is an asymptotically normal estimator of θ with standard error $SE(T_n) = \hat{\sigma}_{T_n}$. Then the null hypothesis H_0: $\theta = \theta_0$ may be tested by using the statistic

$$z_{obs} = \frac{T_n - \theta_0}{SE(T_n)}, \tag{10.12}$$

with large values of $|z_{obs}|$ indicating evidence against H_0. If the sample size is large, an approximate p-value may be obtained from

$$p = P(|Z| \geq |z_{obs}|) \tag{10.13}$$

where $Z \sim N(0, 1)$.

This result follows from the theorem in Section 7.3.5, which said that if $\hat{\sigma}_{T_n}$ is the standard error of T_n in the sense that

$$\frac{\hat{\sigma}_{T_n}}{\sigma_{T_n}} \xrightarrow{P} 1$$

then

$$\frac{(T_n - \theta)}{\hat{\sigma}_{T_n}} \xrightarrow{D} N(0, 1).$$

If $\theta = \theta_0$ then the random variable

$$Z = \frac{T_n - \theta_0}{SE(T_n)}$$

follows, approximately, for large n, a $N(0, 1)$ distribution and the p-value based on $Z \sim N(0, 1)$ will be approximately correct. Because Z is a common notation for a $N(0, 1)$ random variable, the value z_{obs} in (10.12) is often called a *z-score* and the procedure in (10.12) and (10.13) is a *z-test.*

Example 1.4 (continued from p. 257) Suppose $X \sim B(n, \theta)$ and we wish to test $H_0 : \theta = \theta_0$. The usual formula for SE is $SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$. It is customary to find SE under the null hypothesis, $\theta_0 = .5$, i.e., we replace[4] $\hat{\theta}$ with $\theta_0 = .5$ in the calculation of SE. In the case of the data from P.S. we had $n = 17$ so we get $SE = \sqrt{(.5)(.5)/17} = .121$, and $z_{obs} = (.824 - .5)/.121 = 2.68$. This gives us a

[4] The logic of the procedure does not demand that we use θ_0 in place of $\hat{\theta}$. The justification of the large-sample significance test, the Theorem in Section 7.3.5 that says Z is approximately $N(0, 1)$, ~~and~~ is not refined enough to distinguish between the two alternative choices for $SE(T_n)$ (both would satisfy the theorem). However, because we are doing the calculation under the assumption that $\theta = \theta_0$, it makes some sense to use the value $\theta = \theta_0$ in computing the standard error.

10.4 Interpretation and Properties of Tests

We now turn to some theoretical aspects of significance tests. In practice, new situations arise where no standard test is available. Researchers then invent significance tests, and sometimes they are not valid. What do we mean by this? The key property is Eq. (10.24). For an evaluation of statistical significance to be correct, theoretically, (10.24) must be satisfied.

Let $F_Q(x)$ be the cdf of Q under the statistical model specified by H_0 and let us assume that Q follows a continuous distribution. We then have $P(Q \leq q) = 1 - P(Q \geq q)$ and we obtain from (10.24) the equivalent form

$$p = 1 - F_Q(q_{obs}). \tag{10.26}$$

This will help below. Sometimes (10.24) does not hold exactly, but it does hold approximately, as in the case of chi-squared tests. In Section 10.4.1 we derive two consequences that allow us to check whether (10.24) is approximately true. That section describes the behavior of a valid significance test when H_0 is true. In Section 10.4.3 we consider what happens when H_0 is false.

10.4.1 Statistical tests should have the correct probability of falsely rejecting H_0, at least approximately.

The criteria for determining statistical significance, usually taken to be .05 or .01, are called *significance levels*. Fisher suggested[9] that research workers might routinely use $p < .05$ as a "convenient convention" to summari~~s~~e the evidence against H_0. Indeed, this became standard practice. Neyman and Pearson then considered, formally, the behavior of such a procedure. They began by saying one might *reject* H_0 for sufficiently large values of the test statistic Q. If we let c be the cut-off value for which H_0 is rejected whenever $Q \geq c$, then c is called the *critical value* and

summarize

3

$$\alpha = P(Q \geq c)$$

is called the *level* of the test for the critical value c. Now, for the t-test on p. 265 based on $Q = |T|$ and $q_{obs} = t_{obs}$ defined in (10.19), at a particular level, such as $\alpha = .05$, we may reverse the process and, for any α, we can find a critical value c_α such that

$$\alpha = P(Q \geq c_\alpha). \tag{10.27}$$

[9] See pages 114 and 128 of the fourteenth (1970) edition of Fisher (1925).

as 1–6 is typically used; from many repeated trials, with a 6 point scale, 5 points are obtained along an empirical ROC curve. (For the lowest confidence value there is never a perception of response at all, so it is considered to correspond to $c = \infty$.) The common terminology used in SDT replaces the y-axis label of power with "hit rate," or "hits," and the x-axis label of level with "false alarm rate," or simply "false alarms." As in the case of statistical tests, the null and signal-plus-noise distributions are often assumed to be normal, but that is not essential to the logic of the method.

lower case

Example 10.6 Dual-Process Theory of Memory One method of studying memory has involved recall of words taken from a list that was previously studied. A variant of this uses a list of words consisting of some previously studied (or "old") words together with some new words; then, for each word taken from the composite list the experimental subject is asked to say whether the word is new or old. This produces a series of binary judgments to which SDT may be applied: the old words define the signal-plus-noise condition, while the new words define the null condition. According to certain dual-process theories of memory, there is a distinction between remembering based on some set of details or related events, and remembering without such corresponding details being available and, instead, there is only a sense of "familiarity." Yonelinas (2001) reported an experiment in which 19 subjects were each given a list of 58 words to study, and then were tested on a composite list of 75 words. Half of the old words were studied under "full attention" and half were studied under "divided attention." In the full attention condition subjects saw each word for 1.5 s and were instructed to try to remember it. In the divided attention condition the subjects also had to judge the magnitude of a number, presented on the same screen as the word. The composite list of 75 test words consisted of 25 old words studied under full attention, 25 studied under divided attention, and 25 new words. The subjects were required to judge whether each test word was new or old using a 6 point scale (ranging from "sure it was new" to "sure it was old") and then, after the judgment had been made (and the word was no longer visible), they were also required to indicate whether they could remember details about the word, such as what it looked like or sounded like, and whether they would be able to report such details. Words were considered to be recognized based on familiarity when no details could be recalled.

According to the dual-process model of Yonelinas, ROC curves for familiar objects should be similar to those obtained from a pair of displaced normal distributions, as in Fig. 10.3, whereas words recollected with details would have a constant probability of memory retrieval once a minimal confidence threshold was exceeded. A pair of ROC curves for the familiarity words, in both the full attention and divided attention conditions, are shown in the left-hand part of Fig. 10.4. As support for the dual-process theory, Yonelinas also presented ROC curves for the words recognized with detailed recollection, and these curves were quite flat, with an apparent threshold at which recollection occurred. These are in the right-hand part of Fig. 10.4. □

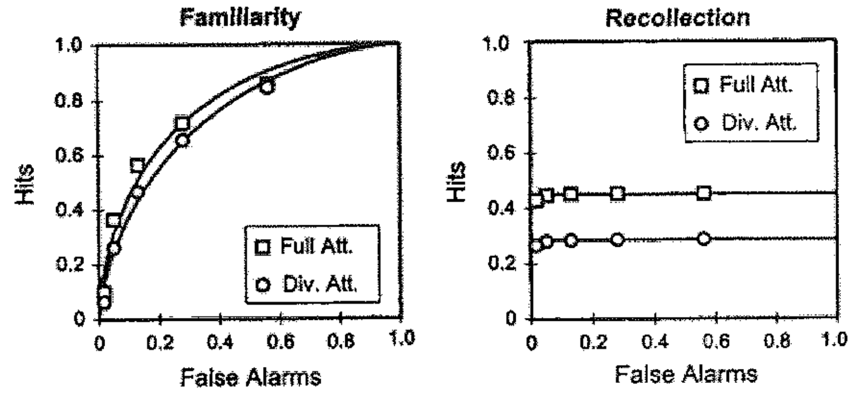


Fig. 10.4 ROC curves, adapted from Yonelinas (2001). On the *Left* are curves from words recognized based only on familiarity, and on the *Right* are curves from words for which recognition was based on detailed recollection. Curves for full attention and divided attention words are plotted separately.

10.4.5 The p-value is not the probability that H_0 is true.

The p-value is commonly misinterpreted as the probability that the null hypothesis is true. ~~This is wrong.~~ A correct statement is necessarily rather cumbersome. Let us continue to write a generic test statistic as Q and the value it takes when calculated from data as q_{obs}. In the case of the chi-squared tests we used $Q = X \sim \chi^2_\nu$ with $x_{obs} = \chi^2_{obs}$ and for the two-sided t test (10.19) we used $Q = |T|$ with $q_{obs} = |t_{obs}|$. We chose the notation q_{obs} so that we can clearly distinguish the observed value from the theoretical random variable Q. The p-value is then given by Eq. (10.24). In words, p is the probability that one *would observe* a value of the test statistic as discrepant from the null hypothesis as the one observed from the data, *if the null hypothesis were true*. Or, again, in slightly different words: if the null hypothesis were true, the test statistic Q would have a probability distribution; the p-value is the resulting probability that Q would be as discrepant from the null hypothesis as the value q_{obs} actually observed. There is no substantially simpler way to say this. The important point about the correct interpretation is its subjunctive nature: the p-value is a probability based on what *might* have happened if a random sample had been drawn under H_0.

Because the logic behind p-values is somewhat convoluted, they are very often misinterpreted to mean something much simpler and more direct, namely the probability that H_0 is true based on the data. That is, a value $p = .05$ is often misinterpreted as meaning that .05 is the probability that H_0 is true, which we would write as $P(H_0|data) = .05$. This is sometimes called the *p-value fallacy* (Goodman et al. 1999a, b). There is no denying how nice it would be to have $P(H_0|data)$. In principle, that probability may be obtained, instead, from Bayes' Theorem:

$$Y_i \sim N(\beta_0, \sigma^2),$$

independently, for $i = 1, \ldots, n$ is nested within the simple linear regression model

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2),$$

independently, for $i = 1, \ldots, n$. Note that LR_{obs} satisfies $LR_{obs} \leq 1$: if

$$L(\hat{\omega}, \hat{\theta}) = \max_{(\omega,\theta)} L(\omega, \theta)$$

and

$$L(\hat{\omega}_0, \theta_0) = \max_{\omega} L(\omega, \theta_0),$$

as in (11.5), then, by definition of the maximum, $L(\hat{\omega}, \hat{\theta}) \geq L(\omega, \theta)$ for any other value of (ω, θ), including $(\hat{\omega}_0, \theta_0)$. Therefore, we have

$$L(\hat{\omega}, \hat{\theta}) \geq L(\hat{\omega}_0, \theta_0). \tag{11.8}$$

The likelihood ratio test accounts for this necessity, and judges the degree to which $L(\hat{\omega}, \hat{\theta})$ exceeds $L(\hat{\omega}_0, \theta_0)$ according to (11.7).

When two models are to be compared and neither is a reduced special case of the other the models are called *non-nested.* For non-nested models the likelihood ratio test no longer applies. How should non-nested models be compared? If the two models have the same parameter dimensionality it is possible to compare their maximized loglikelihood functions. However, because of (11.8), when non-nested models of different dimensionality are to be compared, some adjustment for dimensionality of the parameter vectors must be made. The most common methods introduce a criterion that starts with the maximized loglikelihood and then subtracts a penalty for dimensionality. By convention, to match the usual form of the loglikelihood ratio statistic, these criteria are often defined to include a multiplier of -2 so that they may be written as

penalty

$$\text{criterion} = -2 \cdot \text{max loglikelihood} + \text{penalty}.$$

The most widely used criteria are the *Akaike information criterion*, or AIC (Akaike 1974), and the *Bayesian information criterion*, or BIC (Schwarz 1978), for which the penalties are

$$\text{AIC penalty} = 2m$$

where m is the number of parameters in the model, and

$$\text{BIC penalty} = m \log n,$$

where n is the sample size. Thus, for a random vector X following a model M having an m-dimensional parameter vector θ and pdf $f(x|\theta)$ we have

$$\text{AIC(M)} = -2\log f(x|\hat{\theta}) + 2m$$

and

$$\text{BIC(M)} = -2\log f(x|\hat{\theta}) + m\log n.$$

Many variants on these two model selection criteria have also been proposed; they begin with the same idea, and have more or less the same general form. Note that according to the definition we have just given of AIC(M) and BIC(M), smaller values indicate better models. Alternative equivalent forms, such as that obtained by omitting the multiplier -2 (so that larger values indicate better models), are also used frequently in the literature.

Example 11.1 Interspike interval distribution in resting retinal ganglion cells In Section 5.4.6 we introduced the inverse Gaussian distribution as the distribution of interspike intervals for a theoretical integrate-and-fire neuron. Brown et al. (2003), following Iyengar and Liao (1997), analyzed interspike intervals from a resting retinal ganglion neuron recorded *in vitro*, and compared the fits of exponential, gamma, and inverse Gaussian distributions. They obtained AIC $=$ 8,598, 8,567, 8,174 for these three models, respectively, indicating a much better fit for the inverse Gaussian distribution than for either of the other distributions. Plots of fitted pdfs overlaid on the interspike interval histogram were consistent with this evaluation. □

divergence defined

divergence

The motivation for AIC begins with the Kullback-Liebler ~~discrepancy~~ defined on p. 92. Suppose we let $f(x)$ be the true pdf and we wish to obtain a model with pdf $g(x)$ that is a close as possible to $f(x)$ in the sense of minimzing $D_{KL}(f,g)$. When we minimize over $g(x)$ we are maximizing $E_f(\log(g(X)))$. Consider the special case of trying to determine the value of a single scalar parameter θ, where the true value is θ_0, based on data x. Then we are trying to find the closest pdf $g(x|\theta)$ to $f(x) = g(x|\theta_0)$. It is not too hard to show that the expectation $E_f(\log g(X|\theta))$ is maximized by $\theta = \theta_0$. Because θ_0 is unknown we might use the loglikelihood $\log g(x|\theta)$ as an estimate of $E_f(\log g(X|\theta)$, and thus might maximize to get the maximized loglikelihood $\log g(x|\hat{\theta})$. But this is, in general, a biased estimate of $E_f(\log g(X|\theta)$. Akaike proposed to subtract off an estimate of the bias, and then showed that the bias is, in general, approximately equal to the dimensionality of θ. (See Konishi and Kitagawa (2007) for full details.) Multiplying the maximized loglikelihood by -2 gives the form of AIC above.

minimizing

i

BIC begins, instead, with the Bayesian formulation of choosing between models M_1 and M_2 based on posterior probability:

$$P(M_1|x) = \frac{f_1(x|M_1)P(M_1)}{f_1(x|M_1)P(M_1) + f_2(x|M_2)P(M_2)} \tag{11.9}$$

Example 7.2 (continued from p. 300) Applying the bootstrap procedure based on the statistic (11.14) we obtained $p = .0022$. □

11.3 Multiple Tests

11.3.1 When multiple independent data sets are used to test the same hypothesis, the p-values are easily combined.

Sometimes results for each of several subjects, or several experimental units (such as neurons), are equivocal yet all lean in the same direction. Intuitively, such consistency seems to provide additional evidence of a possible effect. Fisher (1925) suggested a simple method of combining multiple independent p-values.

Example 11.2 Precisely repeated intracellular synaptic patterns It has been suggested that precisely timed patterns of synchronous neural activity may propagate across a cortical circuit and, indeed, that such propagation is a crucial mode of information transmission in the brain (see Abeles (2009)). Experimental evidence aimed at supporting this idea, which is controversial, was provided by Ikegaya et al. (2004), who recorded spontaneous intracellular activity in vitro from slices of mouse primary visual cortex and in vivo from cat primary visual cortex. Ikegaya et al. (2008) conducted additional experiments and reanalyzed the original data. The in vitro recordings produced relatively long traces of post-synaptic currents which the authors examined for repeated precise patterns. To judge whether observed patterns might be explained by chance, in one of their analyses they performed a kind of permutation test. Because the computations were very time consuming they used only 50 permutations and, when they found their observed test statistic to exceed the values obtained from all 50 sets of pseudo-data they thus achived statistical significance $p < .02$. This was repeated across 5 neurons. In other words, for each of 5 neurons they achieved $p < .02$, which would seem to be strong statistical evidence that their null hypothesis should be rejected.[3] □

achieved

e

Suppose we have p-values from n independent tests. Fisher observed that under H_0 the p-value for test i would be a uniformly distributed random variable P_i, with $i = 1, \ldots, n$ (see p. 273) and, therefore, the random variable

$$X = -2\sum_{i=1}^{n} \log P_i \tag{11.15}$$

[3] Some care is required to state correctly the null hypothesis, but roughly speaking it corresponds to time intervals between post-synaptic currents being i.i.d., which they would not be if there were repeated patterns.

Example 11.3 Adaptation in fMRI activity among autistic and control subjects Autism is characterized by difficulty in social interaction and communication. One proposal is that autism may involve a defect in the mirror neuron system, which is active in response to observation of activity by other subjects (thus the idea that an individual subject's brain may "mirror" the activity of the other subject). Several studies found the human mirror system to contain subpopulations of neurons that adapt when hand movements are observed or executed repeatedly.[4] Specifically, fMRI responses to observed or executed movements decreased when the movement occurred for a second time. Dinstein et al. (2010) studied brain response adaptation using fMRI, and found that adaptation occurred among autistic subjects as well as controls across multiple regions of interest. The authors considered this to be evidence against mirror system dysfunction in autism.

A crucial step in their argument involved the definition of each region of interest (ROI). For this they combined anatomical and functional characterizations: for each ROI they included every voxel that was both (i) located within 15 mm of an anatomically-defined region and (ii) significantly active based on a t-test of experimental condition versus baseline. Across their ROIs, however, there were thousands of voxels to be examined. In other words, the authors had to perform thousands of tests, of thousands of null hypotheses. This is very common in fMRI studies. □

To see that multiple tests require an additional calculation consider what happens when 100 tests are made. It might be tempting to declare any of the tests signficant when $p < .05$. However, if each of the 100 null hypotheses were true, then we would expect about $(.05)(100) = 5$ of the p-values to satisfy $p < .05$, indicating statistical significance. Thus, we would expect several such tests (about 5) to yield spurious (false) results of evidence against the null. An additional calculation makes the situation even more worrisome. Let us suppose that we have 100 random variables T_i representing test statistics for null hypotheses $H_{0,i}$ with[5]

significant

$$P(|T_i| > c_\alpha | H_{0,i}) = \alpha. \tag{11.19}$$

This implies

$$P(|T_i| \leq c_\alpha | H_{0,i}) = 1 - \alpha$$

for $i = 1, 2, \ldots, 100$. If all the tests are independent then we have

$$P(|T_i| \leq c_\alpha \text{ for all } i | H_{0,i} \text{ for all } i) = (1 - \alpha)^{100}$$

and, therefore,

[4] This is important to the logic of the mirror neuron argument. See Dinstein (2008).

[5] We use the absolute value form $|T_i| > c_\alpha$ for consistency with the two-sided tests emphasized in Chapter 10 but the logic is the same for all significance tests.

$$Y \longleftarrow \begin{cases} \text{noise} \\ f(x_1, \ldots, x_p). \end{cases} \tag{12.4}$$

This diagram is supposed to indicate a variety of generalizations of linear regression which, together, form the class of methods known as *modern regression.*

In this chapter we provide a concise introduction to linear regression. In Sections 12.1–12.4 we treat the *simple linear regression model* given by

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{12.5}$$

for $i = 1, \ldots, n$, where ϵ_i is a random variable. The adjective "simple" refers to the single x variable on the right-hand side of (12.5). When there are two or more x variables on the right-hand side the terminology *multiple regression* is used instead. We go over some of the most fundamental aspects of multiple regression in Section 12.5. That section also lays the groundwork for modern regression. Generalizations are described in Chapters 14 and 15.

12.1 The Linear Regression Model

To help fix ideas, as we proceed we will refer to several examples.

Example 12.1 Neural correlates of reward in parietal cortex Platt and Glimcher (1999) suggested that cortical areas involved in sensory-motor processing may encode not only features of sensation and action but also key inputs to decision making. To support their claim they recorded neurons from the lateral intraparietal (LIP) region of monkeys during an eye movement task, and used linear regression to summarize the increasing trend in firing rate of intraparietal neurons with increasing expected gain in reward (volume of juice received) for successful completion of a task. Figure 12.1 shows plots of firing rate versus reward volume for a particular LIP neuron following onset of a visual cue. □

Example 2.1 (continued from p. 24) In their analysis of saccadic reaction time in hemispatial neglect, Behrmann et al. (2002) used linear regression in examining the modulation of saccadic reaction time as a function of angle to target by eye, head, or trunk orientation. We refer to this study in Section 12.5. □

In Chapter 1 we used Example 1.5 on neural conduction velocity to illustrate linear regression. Another plot of the neural conduction velocity data is provided again in Fig. 12.2.

Before we begin our discussion of statistical inference in linear regression, let us recall some of the things we said in Chapter 1 and provide a few basic formulas.

Given ~~data~~ n data pairs (x_i, y_i), least squares finds $\hat{\beta}_0$ and $\hat{\beta}_1$ that satisfy

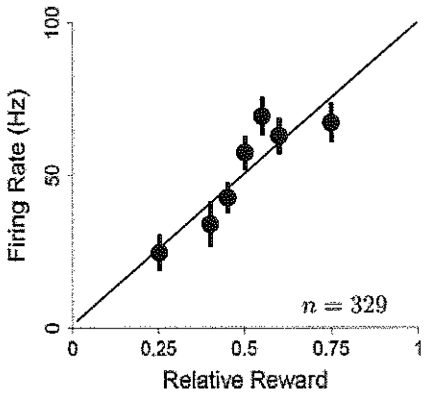


Fig. 12.1 Plots of firing rate (in spikes per second) versus reward volume (as fraction of the maximal possible reward volume). The plot represents firing rates during 200 ms following onset of a visual cue across 329 trials recorded from an LIP neuron. The 329 pairs of values have been reduced to 7 pairs, corresponding to seven distinct levels of the reward volume. Each of the y_i values in the figure is a mean (among the trials with x_i as the reward volume), and error bars representing standard errors of each mean are also visible. A least-squares regression line is overlaid on the plot. Adapted from Platt and Glimcher (1999).

space

$$\sum_{i=1}^{n}\left(y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)\right)^2=\min_{\beta_0^*,\beta_1^*}\sum_{i=1}^{n}\left(y_i-(\beta_0^*+\beta_1^*x_i)\right)^2 \tag{12.6}$$

where we use β_0^* and β_1^* as generic possible estimates of β_0 and β_1. The least-squares estimates (obtained by calculus) are

$$\hat{\beta}_1=\frac{\sum_i(y_i-\bar{y})(x_i-\bar{x})}{\sum_i(x_i-\bar{x})^2} \tag{12.7}$$

and

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}. \tag{12.8}$$

The resulting fitted line

$$y=\hat{\beta}_0+\hat{\beta}_1x \tag{12.9}$$

is the *linear regression* line (and often "linear" is dropped).

> *Details:* To be clear what we mean when we say that the least-squares estimates may be found by calculus, let us write
>
> $$g(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))^2.$$
>
> The formulas (12.8) and (12.7) may be obtained by computing the partial derivatives of $g(\beta_0,\beta_1)$ and then solving the equations

Confidence interval for ρ
Suppose we have a random sample from a bivariate normal distribution with correlation ρ and $R_n = r$ is the sample correlation. Then an approximate 95 % confidence interval for ρ is given by (L, U) where L and U are defined by (12.43), (12.42), and (12.40).

The result (12.41) may also be used to test $H_0{:}\rho = 0$, which holds if and only if $H_0{:}\ \beta_1 = 0$. The procedure is to apply the z-test in Section 10.3.2 using

$$z_{obs} = \sqrt{n-3}z_r,$$

which is z_r divided by its large-sample standard deviation $1/\sqrt{n-3}$, and is thus a z-ratio.

12.4.4 When noise is added to two variables, their correlation diminishes.

When measurements are corrupted by noise, the magnitude of their correlation de~~s~~creases. The precise statement is given in the theorem below, where we begin with two random variables U and W and then add noise to each, in the form of variables ϵ and δ. The noise-corrupted variables are then $X = U + \epsilon$ and $Y = W + \delta$.

Theorem: Attenuation of Correlation Suppose U and W are random variables having correlation ρ_{UW} and ϵ and δ are independent random variables that are also independent of U and V. Define $X = U + \epsilon$ and $Y = W + \delta$, and let ρ_{XY} be the correlation between X and Y. If $\rho_{UW} > 0$ then

$$0 < \rho_{XY} < \rho_{UW}.$$

If $\rho_{UW} < 0$ then

$$\rho_{UW} < \rho_{XY} < 0.$$

Proof details: We assume that $V(\epsilon) > 0$ and $V(\delta) > 0$ and we begin by writing

$$\begin{aligned} Cov(X, Y) &= Cov(U + \epsilon, W + \delta) \\ &= Cov(U, W) + Cov(U, \delta) + Cov(W, \epsilon) + Cov(\epsilon, \delta). \end{aligned}$$

Because of independence the last 3 terms above are 0. Therefore, $Cov(X, Y) = Cov(U, W)$, which shows that ρ_{XY} and ρ_{UW} have the same sign. Suppose $\rho_{UW} > 0$, so that $Cov(U, W) > 0$. Then we have

Example 12.4 Neural correlates of developmental change in working memory from fMRI Many studies have documented the way visuo-spatial working memory (VSWM) changes during development. Kwon et al. (2002) used fMRI to examine neural correlates of these changes. These authors studied 34 children and young adults, ranging in age from 7 to 22. Each subject was given a VSWM task while being imaged. The task consisted of 12 alternating 36-s working memory (WM) and control epochs during which subjects viewed items on a screen. During both the WM and control versions of the task the subjects viewed the letter "O" once every 2 s at one of nine distinct locations on the screen. In the WM task the subjects responded when the current location was the same as it was when the symbol was presented two stimuli back. This required the subjects to engage their working memory. In the control condition the subjects responded when the "O" was in the center of the screen.

visuospatial

One of the y variables used in this study was the maximal BOLD activation (as a difference between WM and control) among voxels within the right prefrontal cortex. They were interested in the relationship of this variable with age (x_1). However, it is possible that Y would increase due to better performance of the task, and that this would increase with age. Therefore, in principle, the authors wanted to "hold fixed" the performance of task while age varied. This is, of course, impossible. What they did instead was to introduce two measures of task performance: the subjects' accuracy in performing the task (x_2) and their mean reaction time (x_3). □

Example 12.1 (continued, see p. 310) The firing rates in Fig. 12.1 appear clearly to increase with size of reward, and the analysis the authors reported (see p. 326) substantiated this impression. Platt and Glimcher also considered whether other variables might be contributing to firing rate by fitting a multiple regression model using, in addition to the normalized reward size, amplitude of each eye saccade, average velocity of saccade, and latency of saccade. This allowed them to check whether firing rate tended to increase with normalized reward size after accounting for these eye saccade variables. □

Equation (12.6) defined the least squares fit of a line. Let us rewrite it in the form

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min_{\beta^*}\sum_{i=1}^{n}(y_i - y_i^*)^2 \tag{12.45}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $y_i^* = \beta_0^* + \beta_1^* x_i$ and $\beta^* = (\beta_0^*, \beta_1^*)$. If we now re-define y_i^* as

$$y_i^* = \beta_0^* + \beta_1^* x_i + \cdots + \beta_p^* x_{pi}$$

x_{1i}

↑

subscript 1 missing

with $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$, Eq. (12.45) defines the least-squares multiple regression problem. We write the solution in vector form as

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p), \tag{12.46}$$

where the components satisfy (12.45) with the fitted values being

x_{1i}

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_p x_{pi}. \quad (12.47)$$

Subscript 1 missing

We interpret the multiple regression equation in Section 12.5.1 and discuss the decomposition of sums of squares in Section 12.5.2. In Section 12.5.3 we show how the multiple regression model may be written in matrix form, which helps in demonstrating how it includes ANOVA models as special cases, and in Section 12.5.4 we show that multiple regression also may be used to analyze certain nonlinear relationships. In Section 12.5.5 we issue an important caveat concerning correlated explanatory variables; in Section 12.5.6 we describe the way interaction effects are fitted by multiple regression; and in Section 12.5.7 we provide a brief overview of the way multiple regression is used when there are substantial numbers of alternative explanatory variables. We close our discussion of multiple regression in Section 12.5.8 with a few words of warning.

12.5.1 Multiple regression estimates the linear relationship of the response with each explanatory variable, while adjusting for the other explanatory variables.

To demonstrate multiple regression in action we consider a simple example.

Example 12.5 Toxicity as a function of dose and weight In many studies of toxicity, including neurotoxicity (Makris et al. 2009) a drug or other agent is given to an animal and toxicity is examined as a function of dose and animal weight. A relatively early example was the study of sodium arsenate (arsenic) in silkworm larvae (Bliss 1936). We reanalyzed data reported there. The response variable (y) was $\log(w/1{,}000)$ where w was minutes survived, and the two predictive variables were log weight, in log grams, and log dose, given in 1.5 plus log milligrams. A plot of log survival versus log dose is given in Fig. 12.8. Because there were two potential outliers that might affect the slope of the line fitted to the plotted data we have provided in the plot the fitted regression lines with and without those two data pairs. The results we discuss were based on the complete set of data.

The linear regression of log survival on log dose gave the fitted line

$$\text{log survival} = .140(\pm .057) - .7\overset{0}{\cancel{3}}4(\pm .0\overset{7}{\cancel{5}}8)\text{log dose}$$

.704(±.078)

which says that survival decreased roughly .704(±.078) log 1,000 min for every log milligram increase in dose. The regression was very highly significant ($p = 10^{-12}$), consistently with the obvious downward trend.

The linear regression of log survival on both log dose and log weight gave the fitted line

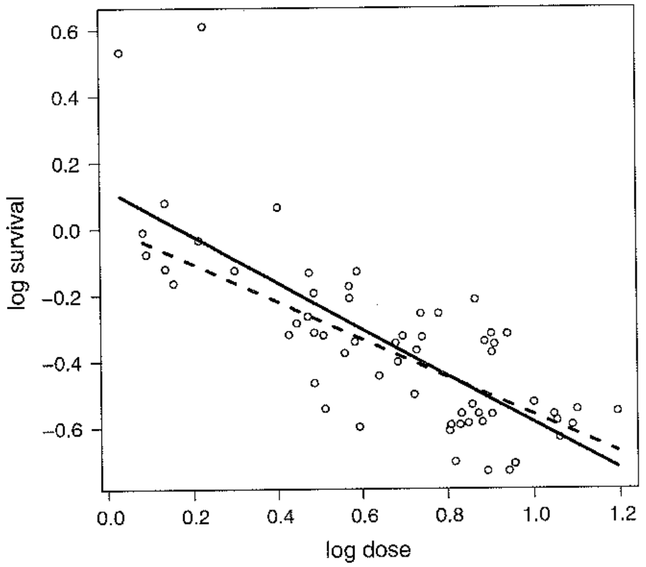


Fig. 12.8 Plot of log survival time ($\log(w/1{,}000)$ where w was minutes survived) versus log dose (1.5 plus log milligrams) of sodium asenate in silkworm larvae; data from Bliss (1936). Lines are fits based on linear regression: *solid line* used the original data shown in plot; *dashed line* after removing the two high values of survival at low dose.

arsenate

r

$$\text{log survival} = .140(\pm.057) - .734(\pm.058)\text{log dose} + 1.07(\pm.16)\text{log weight}.$$

In this case, including weight in the regression does not change very much the relationship between dose and survival: the slope is nearly the same in both cases. □

12.5.2 Response variation may be decomposed into signal and noise sums of squares.

As in simple linear regression we define the sums of squares *SSE* and *SSR*, again using (12.22) and (12.28) except that now $\hat{y}_i$ is defined by (12.47). If we continue to define the total sum of squares as in (12.24) we may again decompose it as

$$SST = SSR + SSE$$

and we may again define R^2 as in (12.25) or, equivalently, (12.27). In the multiple regression context R^2 is interpreted as a measure of the strength of the linear relationship between y and the multiple explanatory variables.

Table 12.1 Simple linear regression results for Example 12.5.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	.120	.057	2.1	.038
Log dose	−.704	.078	−9.1	10^{-12}

Table 12.2 Multiple regression results for Example 12.5.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	−.140	.057	−2.49	.017
Log dose	−.734	.058	−12.6	2×10^{-16}
Log weight	1.07	.16	6.8	6×10^{-9}

Example 12.5 (continued) Returning to the toxicity data, the results for the regression of log survival on log dose are given in Table 12.1. We also obtained $s = .17$ and $R^2 = .59$. The F-statistic was $F = 82$ on 1 and 58 degrees of freedom, with $p = 10^{-12}$ in agreement with the p-value for the t-test in Table 12.1. The results for the regression of log survival on both log dose and log weight are in Table 12.2 and here $s = .13$ and $R^2 = .77$, which is a much better fit. The F-statistic was $F = 97$ on 2 and 57 degrees of freedom, with $p = 2 \times 10^{-16}$.

We would interpret the t ratios and F-statistics as follows: there is very strong evidence of a linear relationship between log survival and a linear combination of log dose and log weight ($F = 97, p << 10^{-5}$); given that log weight is included in the regression model, there is very strong evidence ($t = -12.6, p << 10^{-5}$) that log survival has a decreasing linear trend with log dose; similarly, given that log dose is in the model, there is very strong evidence ($t = 6.8, p << 10^{-5}$) that survival has an increasing linear trend with log weight. □

Example 12.4 Neural correlates of developmental change in working memory from fMRI (continued from p. 333) Recall that in one of their analyses Kwon et al. defined Y to be the maximal BOLD activation (as a difference between WM and control) among voxels within the right prefrontal cortex, and they considered its linear relationship with age (X_1), accuracy (X_2) and reaction time (X_3). They then performed multiple linear regression and found $R^2 = .53$ with $\beta_1 = .75(\pm.20)$, $p < .001$, $\beta_2 = -.21(\pm.19), p = .28$, and $\beta_3 = -.15(\pm.17), p = .37$. They interpreted the results as showing that the right PFC tends to become much more strongly activated in the VSWM task as the subjects' age increases, and that this is not due solely to improvement in performance of the task. □

Example 12.1 (continued from p. 333) Platt and Glimcher fit a multiple regression model to the firing rate data using as explanatory variablesnormalized reward size,

Space

Theorem: Asymptotic normality of least squares estimators For the linear regression model (12.53) suppose conditions (i)–(iv) hold and let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of design matrices such that

$$\frac{1}{n}X^T X \rightarrow C \tag{12.62}$$

for some positive definite matrix C, as $n \rightarrow \infty$. Then the least-squares estimator defined by (12.56) satisfies

$$\frac{1}{s}(X_n^T X_n)^{1/2}(\hat{\beta} - \beta) \xrightarrow{D} N_{p+1}(0, I_{p+1}). \tag{12.63}$$

Proof: See Wu (1981) for references. □

A Detail: It is also possible to use the bootstrap in regression, but this requires some care because under the assumptions (i)–(iv) the random variables Y_i have distinct expected values,

$$E(Y_i) = (1, x_{i1}, \ldots, x_{ip})\beta$$

and so are not i.i.d. The usual approach is to resample the studentized residuals (see p. 319), which are approximately i.i.d. See Davison and Hinkley (1997, page 275). Alternatively, when each vector $x_i = (x_{i1}, \ldots, x_{ip})$ is observed, rather than chosen by the experimenter, it is possible to treat x_i as an observation from an unknown multivariate probability distribution, and thus (x_i, y_i) becomes an observation from unknown distribution, and the data vectors $((x_1, y_1), \ldots, (x_n, y_n))$ may be resampled.[14] This was the bootstrap procedure mentioned in Example 8.2 on p. 241. For additional discussion see Davison and Hinkley (1997). □

an unknown

an

There are many conveniences of the matrix formulation of multiple regression in (12.53) together with (12.54). One is that the independence and homogeneity assumptions in (12.54) may be replaced. Those assumptions imply

$$V(\epsilon) = \sigma^2 I_n,$$

as in (12.54). The analysis remains straightforward if we instead assume

$$V(\epsilon) = R \tag{12.64}$$

[14] Here, Eq. (9.27) becomes

$$\hat{F}_n(x, y) \xrightarrow{P} F_{(X,Y)}(x, y)$$

where $\hat{F}_n$ is the empirical cdf computed from the random vectors $((X_1, Y_1), \ldots, (X_n, Y_n))$.

Table 12.3 Quadratic regression results for the artificial data in the illustration.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	−2.4	2.5	−.95	.37
x	1.86	1.04	1.8	.12
x^2	−.067	.092	−.73	.487

where $u_i \sim N(0, 4)$. We then defined w_1 and w_2 using (12.65) and (12.66) and regressed $y = (y_1, \ldots, y_n)$ on both w_1 (representing x) and w_2 (representing x^2). We obtained the results shown in Table 12.3, with $R^2 = .77$, $s = 2.1$ and $F = 11.9$ on 2 and 7 degrees of freedom, yielding $p = .0056$. From Table 12.3 alone this regression might appear to provide no evidence that y was linearly related to either x nor x^2. However, regressing y on either x or x^2 alone produces a highly significant linear regression. Furthermore, the F-statistic from the regression on both variables together is highly significant. These potentially puzzling results come from the high correlation of explanatory variables: the correlation between x and x^2 is $r = .975$. Keep in mind that the t-statistic for x^2 in Table 12.3 reflects the contribution of x^2 *after* the variable x has been used to explain y and likewise the t-statistic for x reflects the contribution of x after the variable x^2 has been used to explain y. □

or

Let us consider this phenomenon further. Suppose we want to use linear regression to say something about the degree to which a particular variable, say x_1, explains y (meaning the degree to which the variation in y is matched by the variation in the fit of x to y) but we are also considering other variables $x_2, \ldots, x_p$. We can regress y on x_1 by itself. Let us denote the resulting regression coefficient by b. Alternatively we can regress y on $x_1, \ldots, x_p$ and, after applying Eq. (12.56), the relevant regression coefficient would be $\hat{\beta}_1$, the first component of $\hat{\beta}$. When the explanatory variables are correlated, it is not generally true that $b = \hat{\beta}_1$ and, similarly, the quantities that determine the proportion of variability explained by x_1, the squared magnitudes of the fitted vectors, are not generally equal. Thus, when the explanatory variables are correlated, as is usually the case, it is impossible to supply a unique notion of the extent to which a particular variable explains the response—one must instead be careful to say which other variables were also included in the linear regression.

This lack of uniqueness in explanatory power of a particular variable may be considered a consequence of the geometry of least squares.

Details: Let us return to the geometry depicted in Fig. 12.9. As in that figure we take V to be the linear subspace spanned by the columns of X. Because the columns of X are vectors, let us write them in the form $v_1, \ldots, v_p$, and let us ignore the intercept (effectively assuming it to be zero, as we did when we related the SST decomposition to the Pythagorean theorem). The observations on the first explanatory variable x_1 then make up the vector v_1. The extent to which x_1 "explains" the response vector y now becomes the proportion of y that

remove hyphens

jth ith th jth ith ith

where Y_{ij} is the j-th observation in the i-th group, $\mu + \alpha_i$ is the mean for the i-th group and ϵ_{ij} is the error for the j-th observation in the i-th group (the discrepancy between Y_{ij} and $\mu + \alpha_i$). Here, μ is the overall mean (the "grand mean") and α_i is the increment added to that overall mean in obtaining the mean for the ~~i-th~~ group, so that

$$\frac{1}{I}\sum_{i=1}^{I} \mu + \alpha_i = \mu$$

and this implies

$$\sum_{i=1}^{I} \alpha_i = 0. \tag{13.2}$$

We take the number of groups to be I, so that $i = 1, 2, \ldots, I$, and write the number of observations in group i as n_i. In some places we also write the ith group mean as

$$\mu_i = \mu + \alpha_i.$$

The one-way ANOVA assumptions are

(i) the ANOVA model (13.1) holds;
(ii) the errors satisfy $E(\epsilon_i) = 0$ for all i;
(iii) the errors ϵ_i are independent of each other;
(iv) $V(\epsilon_i) = \sigma^2$ for all i (homogeneity of error variances), and
(v) $\epsilon_i \sim N(0, \sigma^2)$ (normality of the errors).

Note that these are the same assumptions as those used in linear regression (apart from the replacement of (12.5) with (13.1); see p. 315). As a result, residual analysis may be used in very much the same way as in regression. Indeed, mathematically, analysis of variance may be considered a special case of linear regression. We return to this in Section 13.2.

The purpose of this model is to provide a basis for statistical comparison of the group means $\mu + \alpha_i$. That is, we ask whether there is evidence that the means are different and, if so, we can estimate how different they are. Formally, we want to test the null hypothesis that the groups means are equal:

$$\mu + \alpha_1 = \mu + \alpha_2 = \cdots = \mu + \alpha_I.$$

The usual way the hypothesis is stated is as follows:

$$\mathrm{H}_0 : \alpha_i = 0 \tag{13.3}$$

for all i, which implies that the group means are equal. It also satisfies the condition that the grand mean μ remains the expectation of Y_{ij} under H_0.

13.1.2 One-way ANOVA decomposes total variability into average group variability and average individual variability, which would be roughly equal under the null hypothesis.

At the beginning of Section 12.5.2 we wrote the basic signal and noise decomposition for regression,

$$SST = SSR + SSE.$$

In ANOVA we decompose the variability in the data similarly into two pieces, replacing SSR with a treatment or "group" sum of squares SS_{group}. To test H_0 defined by (13.3) we compute a measure of the *average* amount of variability due to the groups, and an *average* amount of variability due to error, then compare these. Under the null hypothesis that the group means are equal, there should be no systematic variability due to groups, so that the variability we see in our "average variability due to groups" is the result of background variability in the measurements themselves, that is, the error variability. In other words, the average variability due to groups should be about the same size as the average variability due to error. Thus, to test H_0 we use a ratio of these measures of average variability and when the ratio is much larger than 1 there is evidence against H_0, in favor of there being differences among the groups. We first specify and illustrate the procedure and then indicate its motivation as a likelihood ratio test.

We begin with the total sum of squares

$$SST = \sum_{i,j} (y_{ij} - \bar{y}_{..})^2$$

where the double dots in the subscript on $y_{..}$ indicate that the mean is being taken over all the values of y, averaging across both rows and columns. In the infant exercise example we average across all 24 values. We also define the error (residual) sum of squares to be

$$SSE = \sum_{i,j} (y_{ij} - \bar{y}_{i.})^2$$

ith

remove hyphen

where the single dot in the subscript on $y_{i.}$ indicates that the mean is being taken *within* the i-th group. In the infant exercise example there would be 4 means $\bar{y}_{i.}$ for $i = 1, 2, 3, 4$ and each would be an average across all 6 values in the appropriate column. The group sum of squares is then

$$SS_{group} = SST - SSE.$$

We next obtain averages of the group and error sums of squares by dividing by their respective degrees of freedom, df_{group} and df_{error}. Because of the constraint (13.2) we have $df_{group} = I - 1$ and, with n being the total number of observations, this leaves $n - 1 - (I - 1) = n - I$ degrees of freedom for error, i.e., $df_{error} = n - I$.

analogously with Eq. (10.19). Here, however, we are using *all* the data from the 4 groups to compute s, rather than only the data from two groups we are currently comparing. Therefore, we have 20 degrees of freedom going into s and thus 20 degrees of freedom for the t-test (rather than 10 degrees of freedom if we were using only the 2 groups). We obtain $p = .017$.

An alternative analysis compares the active exercise group with the other three groups, all of which could be considered controls. In this case, we would combine the data from the 3 control groups and thereby end up with two groups: the active excercise group and a single control group, the latter now having 18 observations. We would then use the "two-sample t" analysis, as in (10.21). Carrying this out, we obtain (i) a test of the null hypothesis that the means for these two groups are equal, which we may write as H_0: $\mu_{active} - \mu_{controls} = 0$, and (ii) a 95 % CI for the difference between the means $\mu_{active} - \mu_{controls}$.

exercise

First, we find the two means and standard errors to be 10.12 ± 0.59 and 11.81 ± .34, which gives a t-ratio of 2.46 on 22 degrees of freedom and $p = .022$. Second, applying the formula for the 95 % CI in Eq. (7.31) we find our 95 % CI for the decrease in mean age of walking for the active group compared with controls to be (.26, 3.1) months.

The conclusions from this analysis are different from those on p. 366, based on the F-test. We summarize on p. 374. □

13.1.4 *Two-way ANOVA assesses the effects of one factor while adjusting for the other factor.*

On p. 363 we described the distinction between one-way and two-way tables by contrasting Examples 13.1 and 13.2. To introduce the two-way analysis let us first look further at the data in Example 13.2.

Example 13.2 (continued from p. 363) Figure 13.2 displays the tapping rates for the three drugs across the four subjects. We can see that the subjects have very different tapping rates, but for all four of them the placebo rate is noticeably lower than that obtained with theobromine or caffeine. Also, the comparison of rates for theobromine and caffeine is inconsistent across subjects. The quantitative analysis, below, will support these qualitative observations. □

The two-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where Y_{ij} is the observation for the i-th treatment on the j-th subject, $\mu + \alpha_i + \beta_j$ is its mean, and ϵ_{ij} is the error for the i-th treatment and j-th subject. Here, α_i is the increment added to the overall mean μ in obtaining the mean for the i-th treatment while β_j is the increment added to overall mean in obtaining the mean for the j-th subject. We say that α_i is the *effect* for the ith treatment and β_j the effect for the

ith

jth

ith

jth

ith

jth

remove hyphens

source of variation. In this case their effects may be modeled as random variables. This generates *random-effects models* and they too require specialized techniques. We discuss random-effects models briefly in Chapter 16.

13.1.7 *Additional analyses, involving multiple comparisons, may require adjustments to p-values.*

Because ANOVA involves comparison of several means, many possible hypotheses may be of interest.

Example 13.1 (continued from p. 368) We have already looked at the data on development of motor control in two different ways. On p. 366 we used ANOVA to test the hypothesis of no differences among the mean age of walking, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Then, on p. 368, we reported two further analyses. The first used a t-test to test the null hypothesis of no difference between the active exercise group and the eight-week control group mean ages of walking, H_0: $\mu_1 = \mu_4$ with a t-test. The second used a t-test to test the null hypothesis of no difference between the mean age of walking in the active exercise group and that in the three control groups combined, H_0: $\mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$. We also could have singled out the other control groups and tested H_0: $\mu_1 = \mu_2$ and H_0: $\mu_1 = \mu_3$. Furthermore, because the p-value quantifies the rarity, or surprise, of the results, we ought to ask what other results *might have been* as surprising as those we actually observed. What if the passive exercise group had produced apparent earlier walking, similar to the active exercise group, by comparison with the eight-week control group? Wouldn't that have been a result we would have found interesting? Once we admit that this, too, would have been reported as a finding, then we realize that we were, effectively, testing many possible null hypotheses. The problem of testing multiple hypotheses was discussed in Section 11.3. □

As illustrated in Example 13.1, above, ANOVA often generates many plausible null hypotheses and, in this context, the problem of multiple hypothesis testing is also called the problem of *multiple comparisons*. In Section 11.3 we presented the Bonferroni correction, which can be applied when the number of comparisons (null hypotheses) is easily enumerated. We commented that the Bonferroni method is conservative, in the sense of yielding adjusted p-values that sometimes seem unnecessarily large, making it relatively difficult to obtain statistically significant results. This has spawned a large literature on multiple comparison procedures, most of which aim to provide smaller p-values under specific circumstances, so that it becomes easier to declare statistical signficance. For example, a method due to Dunnett assumes there is a single control group with mean μ_c and considers all null hypotheses of the form H_0: $\mu_i = \mu_c$, for $i \neq c$. When there are I means, there are $I - 1$ such null hypotheses and, under the standard ANOVA assumptions it is possible to find an exact p-value for this case. Similarly, when there is no single control group, a method due to Tukey examines all pairs of means, i.e., all null hypotheses of the form H_0: $\mu_i = \mu_j$ for

Significance

comparisons. Under the standard assumptions, it may be shown that the F-test is significant at level α if and only if there exists a linear contrast for which a test of H_0 defined by (13.9) is significant at level α according to the Scheffé test. □

summarize
3

Example 13.1 (continued from p. 372) Where does all this leave us in this example? We may summarise by saying that there is some evidence, but not strong evidence, that the active group mean age of walking is a bit younger than that for the control groups. The marginal nature of this evidence becomes clear when we ignore the special feature that the latter three groups are all controls and look for differences among all four groups: we find no evidence for this, according to the F-test. Given that it may be difficult to determine exactly when a given child walks, and it is not clear that the parents made this determination in the absence of knowledge about what to expect based on the experimental hypothesis, some skepticism would seem appropriate.[4] □

13.2 ANOVA as Regression

13.2.1 The general linear model includes both regression and ANOVA models.

We now return to the matrix formulation of multiple regression, discussed in Section 12.5.3, and show how linear regression may be used to solve problems of analysis of variance. The points are, first, it can be helpful conceptually to re-frame ANOVA as regression and, second, statistical software typically does this.

ANOVA concerns the comparison of means among several groups, corresponding to experimental conditions. Let us consider two simple examples. Suppose X is the $n \times 1$ vector of 1s

$$X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

We then compute $X^T X = n$ and $X^T Y = \sum y_i$ and find

$$(X^T X)^{-1} X^T y = \bar{y}.$$

Therefore, the sample mean may be found by applying regression with this very special version of the design matrix X.

[4] On the other hand, the paper by Zelazo et al. presented an additional measure where the results were more striking. On this subject, see Adolph (2002).

Next, consider two groups of m values $y_{11}, \dots, y_{1m}$ and $y_{21}, \dots, y_{2m}$, corresponding to two experimental conditions, having sample means $\bar{y}_1$ and $\bar{y}_2$. We define

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1m} \\ y_{21} \\ \vdots \\ y_{2m} \end{pmatrix} \tag{13.10}$$

and

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \tag{13.11}$$

where the first column contains m rows of 1s followed by m rows of 0s and the second column contains m rows of 0s followed by m rows of 1s. The first column of X is an *indicator variable*, indicating membership in the first group, i.e., the ith element of the first column of X is 1 if the ith element of y is in the first group and is 0 otherwise. The second column of X is an indicator variable indicating membership in the second group. We compute

$$X^T X = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}$$

$$X^T y = \begin{pmatrix} \sum y_{1i} \\ \sum y_{2i} \end{pmatrix}$$

and

$$(X^T X)^{-1} X^T y = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix}.$$

experimental

Thus, the sample means are obtained from multiple regression based on the design matrix in (13.11). In a similar manner we may use linear regression to compute means across several expe‸rmental conditions: for each condition we introduce an additional indicator variable as an additional column of the design matrix. The ANOVA from this regression becomes the same as the ANOVA table used in 1-way ANOVA. In

i

Table 13.6 Data from Frezza et al. (1990) on first-pass alcohol metabolism.

Alcoholic Women	Non-alcoholic Women	Alcoholic Men	Non-alcoholic Men
0.6	0.4	1.5	0.3
0.6	0.1	1.9	2.5
1.5	0.2	2.7	2.7
	0.3	3.0	3.0
	0.3	3.7	4.0
	0.4		4.5
	1.0		6.1
	1.1		9.5
	1.2		12.3
	1.3		
	1.6		
	1.8		
	2.0		
	2.5		
	2.9		

13.3.1 Distribution-free nonparametric tests may be obtained by replacing data values with their ranks.

To describe rank-based ANOVA we begin with an example.

Example 13.5 Alcohol metabolism among men and women Women seem to have a lower tolerance for alcohol than men, and are more prone to develop alcohol-related diseases. When men and women of the same size and history of drinking consume equal amounts of alcohol, the alcohol in the bloodstream of the women tends to be higher. In research by Frezza et al. (1990), the "first-pass" metabolism of alcohol in the stomach was studied. The data shown in Table 13.6 come from 18 women and 14 men who volunteered to be studied. Each subject was given two doses of .3 g ethanol per kilogram of body weight, one orally and one intraven~~e~~ously on two different days. The difference in concentrations of alcohol in the blood (at some fixed time after administration), between the intraven~~e~~ous dose and the oral dose, provides a measure of first-pass metabolism in the digestive system and liver; this defines the response variable in the table, with units in mmols per liter per hour. If first-pass metabolism were more effective in men than women, the difference in levels following intraven~~e~~ous and oral administration would tend to be higher among men.

intravenously

intravenous

We begin by ignoring the distinction between alcoholic and non-alcoholic subjects. This reduces the data to two groups: women and men. The data in Table 13.6 are strikingly skewed toward high values. One possibility would be transform the data and apply the usual t-test. Instead, we describe a rank-based analysis.

13.3.2 Permutation and bootstrap tests may be used to test ANOVA hypotheses.

In Section 11.2 we described how permutation and bootstrap tests may be used as alternatives to the t-distribution for computing a p-value in order to test H_0: $\mu_1 = \mu_2$ based on data involving sample sizes n_1 and n_2. The essential method was to (i) merge the data, then (ii) repeatedly resample the $n_1 + n_2$ data values, putting them arbitrarily into groups of size n_1 and n_2 to create pseudo-data, (iii) to each pseudo-data pair of samples apply the t-statistic, and finally (iv) see what proportion of the pseudo-data give t-statistic values greater than that observed in the real data. When the sampling is done without replacement the method is a permutation test, and with replacement it becomes a bootstrap test.

For one-way ANOVA the procedure is exactly analogous. For instance, with 3 conditions we would have data with sample sizes n_1, n_2, and n_3; we would follow step (i) then in (ii) resample the $n_1 + n_2 + n_3$ data values and put them into groups of sizes n_1, n_2, n_3; in (iii) we would get the F-statistic, and likewise in (iv) we would see what proportion of the pseudo-data F values exceed the F obtained for the real data.

Two-way ANOVA is more complicated because the two-way structure must be respected, but the concept is the same. See Manly (2007).

13.4 Causation, Randomization, and Observational Studies

13.4.1 Randomization eliminates effects of confounding factors.

Most studies aim to provide causal explanations of observed phenomena. To claim causality, investigators must argue that alternative explanations of an observed relationship are implausible.

Example 13.6 IQ and breast milk Lucas et al. (1992) obtained IQ test scores from 300 children who had been premature infants and initially fed milk by a tube. The children were 8 years old when they took the IQ test. The milk they had been fed by tube was either breast milk or prepared formula, or some combination of the two. Of interest was the relationship between IQ test scores and the proportion of milk the infants received that was breast milk. The amount of breast milk a baby had drunk was determined by whether or not the mother wished to feed the infant by breast milk, and how much milk the mother was able to express. □

In Example 13.6, immediately we must be aware of possible *confounding factors.* The decision to administer the treatment, i.e., to use breast milk or not, was the mother's; whatever might determine that decision *and also be related to susequent IQ* would affect the observed relationship between IQ and consumption of breast subsequent

b

intensity could be collected into a proportion out of 50 that resulted in perception. We fit the data by applying maximum likelihood estimation to the logistic regression model in (8.43) and (8.44). This[2] is known as *logistic regression*. □

Example 2.1 (continued from p. 378) In Section 13.2.2 we discussed ANOVA interactions in the context of the study by Behrmann et al. (2002) on hemispatial neglect, where the response was saccadic reaction time and one of the explanatory variables was angle of the starting fixation point of the eyes away from "straight ahead." A second response variable of interest in that study was saccadic error, i.e., whether the patient failed to execute the saccade within a given time window. Errors may be coded as 0 and successful execution as 1. Behrmann et al. (2002) used logistic regression to analyze the error rate as a function of the same explanatory variables. They found, for example, that the probability of error increased as eyes fixated further to the right. □

From (14.1) and (14.2) together with normality, for a single explanatory variable x, in linear regression we assume

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

There are three problems in applying ordinary linear regression with binary responses to obtain fitted probabilities: (i) a line won't be constrained to (0, 1), (ii) the variances are not equal, and (iii) the responses are not normal (unless we have proportions among large samples, in which case the proportions would be binomial for large n and thus would be approximately normal, as in Section 5.2.2). The first problem, illustrated in Fig. 8.9, is that the linear regression may not make sense beyond a limited range of x values: if $y = a + bx$ and $b > 0$ then y must become infinitely large, or small, as x does. In many data sets with dichotomous or proportional responses there is a clear sigmoidal shape to the relationship with x. The second problem was discussed in the simpler context of estimating a mean, in Section 8.1.3. There we derived the best set of weights to be used for that problem, and showed that an esimator that omits weights can be very much more variable, effectively throwing away a substantial portion of the data. Much more generally it is also possible to solve

estimator

t

[2] The analysis of Hecht et al. (1942) was different, but related. They wished to obtain the minimum number of quanta, n, that would produce perception. Because quanta are considered to follow a Poisson distribution, in the notation we used above, they took $W \sim P(\lambda)$ and $c = n$, with λ, the mean number of quanta falling on the retina, being proportional to the intensity. This latter statement may be rewritten in the form $\log \lambda = \beta_0 + x$, with x again being the log intensity. Then $Y = 1$ (light is perceived) if $W \geq n$ which occurs with probability $p = 1 - P(W \leq n - 1) = 1 - F(n - 1|\lambda)$, where F is the Poisson cdf. This is a latent-variable model for the proportional data (similar to but different than the one on p. 399). It could be fitted by finding the MLE of β_0, though Hecht et al. apparently did the fitting by eye. Hecht et al. then determined the value of n that provided the best fit. They concluded that a very small number of quanta sufficed to produce perception, but see also Teich et al. (1982).

factor of 2 the log odds thus increase by 3.22 ± .72 (where 3.22 = (.301)(10.7) and .72 = (.301)(2.4)). This gives an approximate 95 % CI for the factor by which the odds increase, when the intensity doubles, of exp(3.22 ± .72) = (12.2, 51.4).

We can go somewhat further by converting odds to the probability scale by inverting

$$\text{odds} = \frac{p}{1-p}$$

to get

$$p = \frac{\text{odds}}{1+\text{odds}}.$$

Let us pick $p = .5$, so that the odds are 1. If we increase the odds by a factor ranging from 12.2 to 51.4 then the probability would go from .5 to somewhere between .92 and .98 (where $.92 = 12.2/(1+12.2)$ and $.98 = 51.4/(1+51.4)$). Thus, if we begin at the x_{50} intensity (where $p = .5$) and then double the intensity, we would obtain a probability of perception between .92 and .98, with 95 % confidence. This kind of calculation may help indicate what the fitted model implies. □

Logistic regression extends immediately to multiple explanatory variables: for m variables $x_1, \ldots, x_m$ we write

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi}.$$

The multiple logistic regression model may be written in the form

$$\begin{aligned} Y_i &\sim B(n_i, p_i) \\ \log \frac{p_i}{1-p_i} &= x_i\beta \end{aligned} \tag{14.6}$$

where β is the coefficient vector and x_i is the $1 \times (m+1)$ vector of values of the several explanatory variables corresponding the ith unit under study.

14.1.2 In logistic regression, ML is used to estimate the regression coefficients and the likelihood ratio test is used to assess evidence of a logistic-linear trend with x.

g

It is not hard to write down the likelihood function for logistic regression. The responses Y_i are independent observations from $B(n_i, p_i)$ distributions, so each pdf has the form $\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}$ and the likelihood function is

regression

g

Table 14.1 Linear regression results for data from subject S.S. in Example 5.5.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	−1.78	.30	−5.9	.0042
Intensity	1.20	.16	7.5	.0017

Table 14.2 Logistic regression results for data from subject S.S. in Example 5.5.

Variable	Coefficients	SE	t_{obs}	p-value
Intercept	−20.5	2.4	−8.6	$p < 10^{-6}$
Intensity	10.7	1.2	8.~~6~~ 9	$p < 10^{-6}$

8.9

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{n_i-y_i}$$

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

where the second equation is substituted into the first. Standard statistical software may be used to maximize this likelihood. The standard errors are obtained from the observed information matrix, as described in Section 8.3.2.

For a single explanatory variable, the likelihood ratio test of Section 11.1.3 may be used to test $H_0 : \beta_1 = 0$. More generally, if there are variables $x_1, \ldots, x_p$ in model 1 and additional variable $x_{p+1}, \ldots, x_{p+m}$ in model 2, then the likelihood ratio test may again be applied to test $H_0 : \beta_{p+1} = \cdots = \beta_{p+m} = 0$. The log likelihood ratio has the form

$$-2 \log LR = -2[\log(\hat{L}_1) - \log(\hat{L}_2)]$$

where $\hat{L}_i$ is the maximum value of the likelihood under model i. For large samples, under H_0, $-2 \log LR$ follows the χ^2 distribution with m degrees of freedom.

In some software, the results are given in terms of "deviance." The *deviance* for a given model is $-2\log(\hat{L})$. The *null deviance* is the deviance for the "intercept-only" model, and we denote it by $-2\log \hat{L}(0)$. Often, the deviance from the full fitted model is called the *residual deviance*. In this terminology, the usual test of $H_0 : \beta_1 = 0$ is based on the difference between the null deviance and the residual deviance.

Example 5.5 (continued) The output from least-squares regression software is given in Table 14.1. The F statistic in this case is the square of t_{obs} and gives the $p = .0017$, as in Table 14.1. The results for logistic regression are given in Table 14.2. The null deviance was 257.3 on 5 degrees of freedom and the residual deviance was 2.9 on 4 degrees of freedom. The difference in deviance is

$$\text{null deviance} - \text{residual deviance} = 257.3 - 2.9 = 256.4$$

spike counts deviated from that predicted by a Poisson distribution, the deviation was small (Ventura et al. 2002). Here we will use the data to illustrate a version of ANOVA based on Poisson regression. Note that in Table 14.5 there are a total of 58 spike counts, from 58 trials. □

The problem of fitting counts is analogous to, though less extreme than, that of fitting proportions. For proportions, the (0,1) range could make linear regression clearly inappropriate. Counts have a range of $(0, \infty)$. Because the ordinary regression line is not constrained, it will eventually go negative. The simple solution is to use a log transformation of the underlying mean. The usual Poisson regression model is

$$Y_i \sim P(\lambda_i) \tag{14.7}$$

$$\lambda_i = \exp(\beta_0 + \beta_1 x_i). \tag{14.8}$$

To interpret the model we use the log transformation:

$$\log \lambda_i = \beta_0 + \beta_1 x_i.$$

For example, in the SEF data of Example 14.1 ~~$\log \lambda_i$~~ Y_i is the spike count and x_i is the experimental condition (up, down, left, right) for the ith trial. The advantage of viewing ANOVA as a special case of regression is apparent: we immediately generalize Poisson ANOVA by applying our generalization of linear regression to the Poisson regression model above.

Y_i is the

14.1.5 *In Poisson regression, ML is used to estimate coefficients and the likelihood ratio test is used to examine trends.*

As in logistic regression we use ML estimation and the likelihood ratio test ("analysis of deviance").

Example 14.1 (continued) We perform Poisson regression using indicator variables as described in Section 13.2.1 to achieve an ANOVA-like model. Specifically, we concatenate the data in Table 14.5 so that the counts form a 58×1 vector and define a variable *left* to be 1 for all data corresponding to the left saccade direction and 0 otherwise, and similarly define vectors *up* and *right*. The results from ordinary least-squares regression are shown in Table 14.6. The F-statistic was 18.76 on 3 and 54 degrees of freedom, giving $p < 10^{-6}$. The Poisson regression output, shown in Table 14.7 is similar in structure. Here the null Deviance was 149.8 on 57 degrees of freedom and the residual Deviance was 92.5 on 54 degrees of freedom. The difference in deviances is

$$\text{null deviance - residual deviance} = 149.8 - 92.5 = 57.3$$

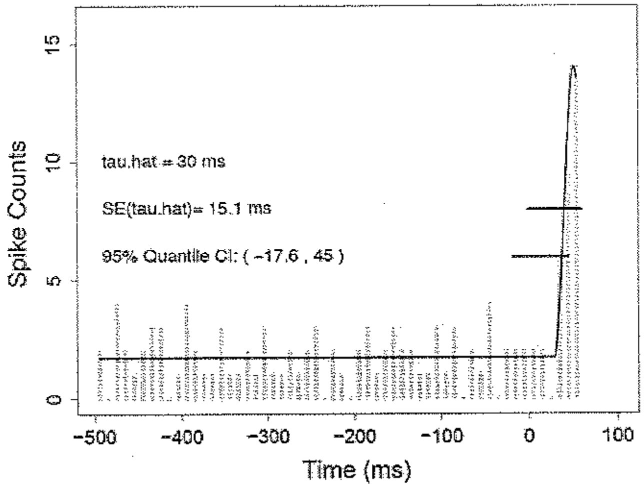


Fig. 14.3 Initiation of firing in a neuron from the basal ganglia: change-point and bootstrap confidence intervals when a quadratic model is used for the post-change-point firing rate. Two forms of approximate 95 % confidence intervals are shown. The first is the usual estimate $\pm 2SE$ interval. The second is the interval formed by the .025 and .975 quantiles among the bootstrap samples. The latter typically performs somewhat better, in the sense of having coverage probability closer to .95.

See Section 9.2.2.

14.2.2 Generalized nonlinear models may be fitted using maximum likelihood.

Nonlinear relationships also arise in the presence of non-normal noise. We use the term *generalized nonlinear model* to refer to a model in which the linear function $g(\mu_i)$ in (14.13) is replaced by a nonlinear function. We give two examples of nonlinear Poisson regression. The first involves determination of a change-point, and is similar to Example 8.2 in Section 14.2.1.

Example 14.4 Onset latency in a basal ganglia neuron An unfortunate symptom of Parkinson's disease (PD) is muscular rigidity. This has been associated with increased gain and inappropriate timing of the long latency component of the stretch reflex, which is a muscular response to sudden perturbations of limb position. One of the important components of the stretch reflex is mediated by a trans-cortical reflex, probably via corticospinal neurons in primary motor cortex that are sensitive to kinesthetic input. To investigate the neural correlates of degradation in stretch reflex, Dr. Robert Turner and colleagues at the University of Pittsburgh have recorded neurons in primary motor cortex of monkeys before and after experimental production of PD-like symptoms. One part of this line of work aims at characterizing neuronal response latency following a limb perturbation (see Turner and DeLong (2000)). Figure 14.3 displays a PSTH from one neuron prior to drug-induced PD symptoms. The statistical problem is to identify the time at which the neuron begins

corticospinal

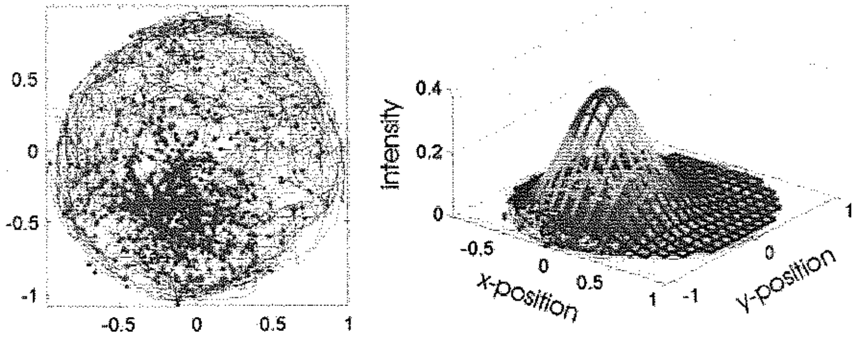


Fig. 14.4 Spiking activity of a rat Hippocampal place cell during a free-foraging task in a circular environment. *Left* Visualization of animal's path and locations of spikes. *Right* Place field model for this neuron, with parameters fit by the method of maximum likelihood.

based on a 2-dimensional bell-shaped curve. For this purpose of specifying the dependence of spiking activity on location a normal pdf may be used. Let us take $Y_t \sim P(\lambda_t)$, with t signifying time, and then define

$$\lambda_t = \exp\left\{\alpha - \frac{1}{2}\left(x(t) - \mu_x \; y(t) - \mu_y\right)\begin{pmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{pmatrix}^{-1}\begin{pmatrix}x(t) - \mu_x\\ y(t) - \mu_y\end{pmatrix}\right\}. \quad (14.21)$$

The explanatory variables in this model are $x(t)$ and $y(t)$, the animal's x and y-position. The model parameters are $(\alpha, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy})$, where (μ_x, μ_y) is the center of the place field, $\exp \alpha$ is the maximum firing intensity at that point, and σ_x^2, σ_y^2, and σ_{xy} express how the intensity drops off away from the center. Note that it is the shape of the place field that is assumed normal, not the distribution of the spiking activity. The right panel of Fig. 14.4 displays a fit of the place field to the data in the left panel. We will discuss models of this sort when we discuss point processes in Chapter 19. □

14.2.3 *In solving nonlinear optimization problems, good starting values are important, and it can be helpful to reparameterize.*

As in maximization of any likelihood, use of the numerical procedures require s care. Two important issues are the choice of initial values, and of parameterization. Both of these may be illustrated with the exponential model (14.15).

requires

Illustration: Exponential regression To fit the exponential model (14.15) a first step is to reparameterized from θ to ω using $\omega_1 = \log(\theta_1)$ and $\omega_2 = \theta_2$ so that the expected values have the form

Chapter 15
Nonparametric Regression

At the beginning of Chapter 14 we said that modern regression applies models displayed in Eqs. (14.3) and (14.4):

$$Y_i \sim f_{Y_i}(y_i|\theta_i)$$
$$\theta_i = f(x_{1i}, \ldots, x_{pi})$$

in (14.6) or (14.15).

where $f_{Y_i}(y|\theta)$ is some family of pdfs that depend on a parameter θ, which is related to $x_1, \ldots, x_p$ according to a function $f(x_1, \ldots, x_p)$. In Section 14.1 we discussed the replacement of the normal assumption in (14.3) with binomial, Poisson, or other exponential-family assumptions. In Section 14.2 we showed how the linear assumption for $f(x_1, \ldots, x_p)$ in (14.4) may be replaced with a specified nonlinear modeling assumption. What if we are unable or unwilling to specify the form of the function $f(x_1, \ldots, x_p)$? In this chapter we consider fitting general functions, which are chosen to provide flexibility for fitting purposes. This is the subject of *nonparametric regression*. The terminology "nonparametric" refers to the absence of a specified parametric form, such as in (14.~~3~~). We focus almost exclusively on the simplest case of a single explanatory variable x, and thus consider functions $f(x)$. Here is an example.

6 or (14.15)

Example 15.1 Peak minus trough differences in response of an IT neuron Some neurons in the inferotemporal cortex (IT) of the macaque monkey respond to visual stimuli by firing action potentials in a series of sharply defined bursts. Rollenhagen and Olson (2005) found that displaying an object image in the presence of a different, already-visible "flanker" image could enhance the strength of the oscillatory bursts. Figure 15.1 displays data (in the form of PSTHs) from an IT neuron under two conditions: in the first, a black patterned object was displayed as the stimulus for 600 ms; in the second condition, prior to the display of the stimulus a pair of blue rectangles appeared (as a flanker image) and these remained illuminated while the patterned-object stimulus was displayed. Overlaid on the PSTHs are fits obtained by the nonparametric regression method BARS, which will be explained briefly in Section 15.2.6. In part b of Fig. 15.1 the BARS fits are displayed together, to highlight

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_15,

and the hat matrix is $H = X(X^TX)^{-1}X^T$. In the case of linear regression we are able to propagate uncertainty using the distribution of $\hat{\beta}$ (as we did, similarly, for logistic regression in Chapter 9), but we could instead propagate the uncertainty from the distributions of the fitted values $X\hat{\beta}$: we simply need the variance

$$\begin{aligned} V((\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T) &= HV(Y)H^T \\ &= \sigma^2 HH^T. \end{aligned} \tag{15.2}$$

In the case of linear regression this simplifies because (as is easily checked) $H^T = H$ and $HH^T = H$ so that

$$V((\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n))^T) = \sigma^2 H.$$

For linear smoothers more generally, $H \neq HH^T$ but, in the case of data for which $V(Y_i) = \sigma^2$ with the Y_is being independent of each other, the variance formula (15.2) continues to hold, and it remains easy to apply propagation of uncertainty. In other words, even though we do not have an estimated parameter vector, such as $\hat{\beta}$, from which to compute quantities of interest and their SEs, we can often compute quantities of interest directly from the fitted values, as in the peak minus trough example above, and can then obtain SEs from the variance formula (15.2) together with the large-sample result that the fitted values are approximately normally distributed. Similarly, when linear smoothing methods extend to logistic or Poisson regression it again remains easy to propagate uncertainty.

15.2 Basis Functions

Suppose $f(x)$ is a continuous function on an interval $[a, b]$. A famous theorem in mathematical analysis, the Weierstrass Approximation Theorem, says that $f(x)$ may be approximated arbitrarily well by a polynomial of sufficiently high order. One might therefore think that polynomials could be effective for curve fitting. That is, we could try to fit an unknown function $y = f(x)$ by instead fitting a p-th order polynomial pth

$$y = b_0 + b_1x + b_2x^2 + \cdots + b_px^p,$$

which we can do using least squares, as described in Section 12.5.4. It turns out that polynomials do not perform as well as the theoretical result might suggest. As illustrated in Fig. 15.2, even a twentieth-order polynomial can fail to represent adequately a relatively well-behaved function in the presence of minimal noise. The idea of replacing $f(x)$ with a set of simple functions, however, is very powerful. In the case of polynomials, for data $(x_1, y_1), \ldots, (x_n, y_n)$ we could fit a quadratic using (12.65) and (12.66) and regressing $y = (y_1, \ldots, y_n)$ on w_1 and w_2, and we could similarly define higher-order terms up to

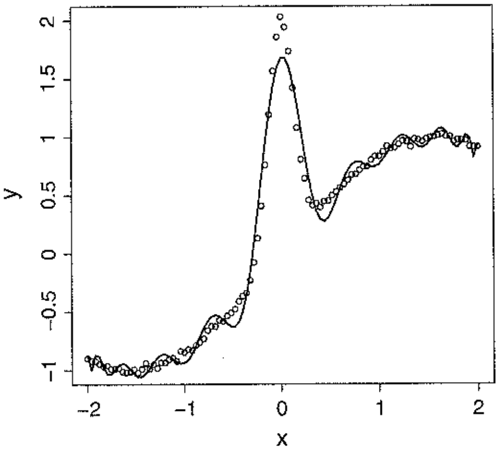


Fig. 15.2 Data simulated from function $f(x) = \sin(x) + 2\exp(-30x^2)$ together with twentieth-order polynomial fit (shown as line). Note that the polynomial is over-fitting (under-smoothing) in the relatively smooth regions of $f(x)$, and under-fitting (over-smoothing) in the peak. (In the data shown here, the noise standard deviation is 1/50 times the standard deviation of the function values.)

$$w_p = \begin{pmatrix} x_1^p \\ x_2^p \\ \vdots \\ x_n^p \end{pmatrix} \tag{15.3}$$

and could regress $y = (y_1, \ldots, y_n)$ on $w_1, w_2, \ldots, w_p$. This is an example of regression using basis functions.

The "basis function" terminology comes from the conception that the theoretical functions $f(x)$ that are, in principle, to be fitted make up an infinite-dimensional vector space for which the chosen simple functions (such as polynomials), form[1] a *basis* (see Section A.9 of the Appendix). In practice we use data $(x_1, y_1), \ldots, (x_n, y_n)$ to fit only the values $(f(x_1), f(x_2), \ldots, f(x_n))$ and thus we have an n-dimensional

[1] In Section A.9 of the Appendix we give the definition of a basis for R^n, which is an n-dimensional vector space. The basis function terminology refers to an extension of this idea to infinitely many dimensions: the functions $f(x)$ on an interval $[a, b]$ that satisfy

$$\int_a^b f(x)dx < \infty$$

(here the Lebesgue integral is used) form an infinite-dimensional vector space and if the functions $B_j(x)$ form a basis then every $f(x)$ may be written as

$$f(x) = \sum_{j=1}^{\infty} B_j(x).$$

lower case "c"

c_j

$$\sum_{j=1}^{\infty} c_j B_j(x)$$

remove hyphens

vector space for which n vectors, defined by the simple functions (such as those in (12.65), (12.66), up through (15.3) with $p = n$), form a basis.

pth

pth

A pth order polynomial regression will work well for functions $y = f(x)$ that look a lot like pth order polynomials. The inability of a 20th-order polynomial to fit the function in Fig. 15.2 is an indication that the function is different than a 20th-order polynomial. The challenge of nonparametric regression using basis functions is to find simple alternatives to polynomials that are flexible enough to fit a variety of functions with relatively few terms.

15.2.1 Splines may be used to represent complicated functions.

The problem in Fig. 15.2 is that the function $f(x)$ is not very close to being a low-order polynomial. In particular, it has a different form near $x = 0$ than it does as the magnitude of x increases. A possible solution here, and in other problems, is to glue together several pieces of polynomials. If the pieces are joined in such a way that the resulting function remains smooth, then it is called a *spline*. We will discuss cubic splines. Let $[a, b]$ be an interval and suppose we have values $\xi_1, \xi_2, \ldots, \xi_p$, where $a < \xi_1 < \xi_2 < \cdots < \xi_p < b$. There are then $p + 2$ subintervals $[a, \xi_1], [\xi_1, \xi_2], \ldots, [\xi_{p-1}, \xi_p], [\xi_p, b]$. A function $f(x)$ on $[a, b]$ is a *cubic spline* with *knots* $\xi_1, \xi_2, \ldots, \xi_p$ if $f(x)$ is a cubic polynomial on each of the $p+2$ subintervals defined by the knots such that $f(x)$ is continuous and its first two derivatives $f'(x)$, and $f''(x)$ are also continuous. This restriction of continuity, and continuity of derivative, applies at the knots; in between the knots, each cubic polynomial is already continuous with continuous derivatives. A cubic spline is shown in Fig. 15.3, and the result of fitting a cubic spline to the data of Fig. 15.2 is shown in Fig. 15.4. In contrast to the 20th order polynomial in Fig. 15.2, the cubic spline in Fig. 15.4 fits the data remarkably well.

15.2.2 Splines may be fit to data using linear models.

It is easy to define a cubic spline having knots at $\xi_1, \xi_2, \ldots, \xi_p$. Let $(x - \xi_j)_+$ be equal to $x - \xi_j$ for $x \geq \xi_j$ and 0 otherwise. Then the function

$$\begin{aligned} f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\ + \beta_4 (x - \xi_1)_+^3 + \beta_5 (x - \xi_2)_+^3 + \cdots + \beta_{p+3} (x - \xi_p)_+^3 \end{aligned} \tag{15.4}$$

is twice continuously differentiable, and is a cubic polynomial on each segment $[\xi_j, \xi_{j+1}]$. Furthermore, with $f(x)$ defined by (15.4),

$$Y_i = f(x_i) + \varepsilon_i$$

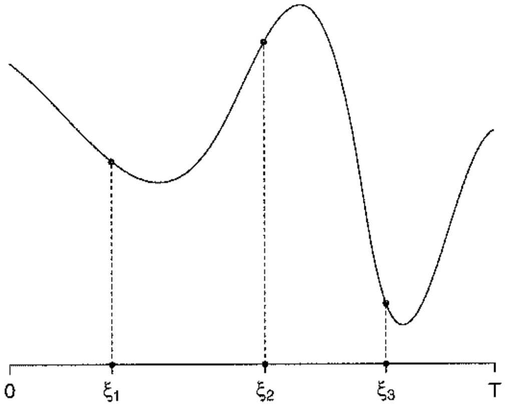


Fig. 15.3 A cubic spline with three knots, on an interval $[0, T]$. The function $f(x)$ depicted here is made up of distinct cubic polynomials (cubic polynomials with different coefficients) on each sub-interval $[0, \xi_1]$, $[\xi_1, \xi_2]$, $[\xi_2, \xi_3]$, $[\xi_3, T]$.

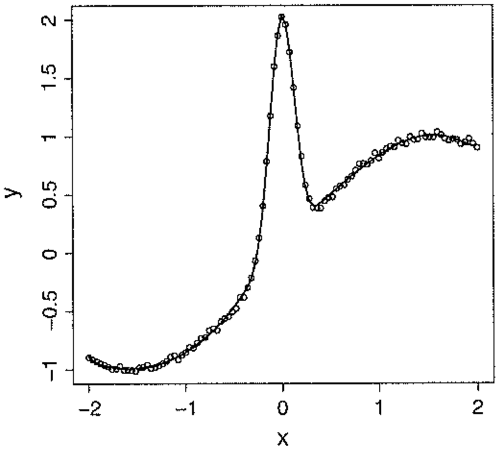


Fig. 15.4 A cubic spline fit to the data from Fig. 15.2. The spline has knots $(\xi_1, \xi_2, \ldots, \xi_7) = (-1.8, -.4, -.2, 0, .2, .4, 1.8)$.

becomes an instance of the usual linear regression model (assuming $\epsilon_i \sim N(0, \sigma^2)$, independently), so that regression software may be used to obtain spline-based curve fitting. Specifically, we define $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, $x_4 = (x - \xi_1)^3_+$, $\ldots$, $x_{p+3} = (x - \xi_p)^3_+$ and then regress Y on $x_1, x_2, \ldots, x_{p+3}$. To be concrete, let us take a simple special case. Suppose we have 7 data values $y_1, \ldots, y_7$ observed at 7 x values ~~$(x_1, \ldots, x_6) =$~~ $(-3, -2, -1, 0, 1, 2, 3)$ and we want to fit a spline with knots at $\xi_1 = -1$ and $\xi_2 = 1$. Then we define $y = (y_1, \ldots, y_7)^T$, $x_1 = (-3, -2, -1, 0, 1, 2, 3)^T$, $x_2 = (9, 4, 1, 0, 1, 4, 9)^T$, $x_3 = (-27, -8, -1, 0, 1, 8, 27)^T$. The variables x_1, x_2, x_3 represent x, x^2, x^3. We continue by defining $x_4 = (0, 0, 0, 1, 8, 27, 64)^T$ and $x_5 = (0, 0, 0, 0, 0, 1, 8)^T$, which represent $(x - \xi_1)^3_+$ (which takes the value 0 for $x \leq -1$) and $(x - \xi_2)^3_+$ (which takes the value 0 for $x \leq 1$). Having defined these variables we regress y on x_1, x_2, x_3, x_4, x_5.

values (-3,2,1,0,1,2,3)

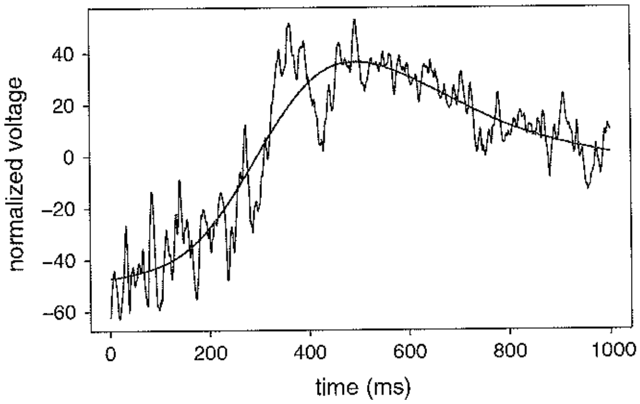


Fig. 15.5 LFP and smoothed version representing slowly-varying trend. A 1 s sample of data is shown together with a smooth fit using natural splines.

power basis and B-spline basis each have $p + 4$ free parameters. Due to the additional constraints at each end of the range of x, the natural spline basis has $p + 2$ free parameters.

Example 15.2 Local field potential in primary visual cortex Kelly et al. (2010) examined the activity of multiple, simultaneously-recorded neurons in primary visual cortex in response to visual stimuli under anaesthesia. As we noted in Example 2.2 under anesthesia the EEG displays strong delta range (1–4 Hz) wave-like activity. It is also common to see even lower frequency activity (less than 1 Hz), often called "slow waves," the effects of which are visible in Fig. 2.2. This activity appears in local field potential (LFP) recordings as well. In the data analyzed by Kelly et al., waves of firing activity were observed across the population of recorded neurons, and these were correlated with the waves of activity in the LFP. A short snippet of LFP is displayed in Fig. 15.5. In Chapter 18 we will examine the oscillatory content of this sample of the LFP. A preliminary step, discussed on p. 517, is to remove any slow trends in the data. Spline-based regression is useful for this purpose. A fit based on the natural-spline basis using knots at time points 200, 400, 600, 800 is shown in Fig. 15.5. □

anesthesia

15.2.3 *Splines are also easy to use in generalized linear models.*

Splines may also be used with logistic regression or Poisson regression, or other generalized regression models. When splines are used in regression models, they are often called *regression splines*. Standard statistical software usually includes options for using regression splines in generalized linear models.

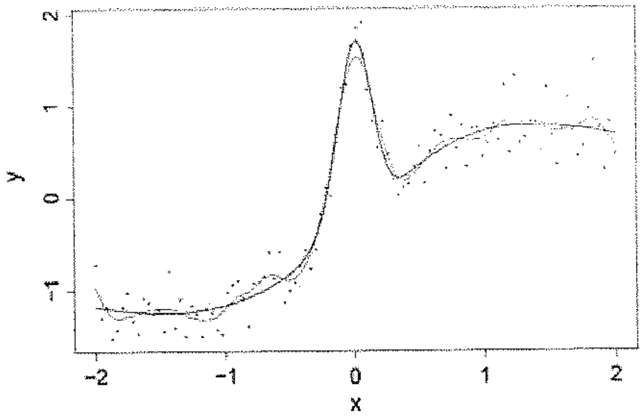


Fig. 15.7 Data from the test function of Fig. 15.2, but with more noise, as in Fig. 15.6, together with smoothing spline fit (*dotted line*) and BARS fit (*solid line*).

or intervals for quantities of interest. Figure 15.7 compares BARS and smoothing spline fits to the data from Fig. 15.6.

15.2.7 Spline smoothing may be used with multiple explanatory variables.

At the beginning of this chapter we recalled Eqs. (14.3) and (14.4), which we had used to define modern regression. In Section 15.2.2 we showed how splines are used to define a function $f(x)$ in ordinary linear regression and in Section 15.2.3 we gave the extension to binomial and Poisson regression. Those sections involved a single explanatory variable x. With p variables $x_1, \ldots, x_p$ it is too difficult to fit a function $f(x_1, \ldots, x_p)$ in full generality: there are too many possible ways that the variables may interact in defining $f(x_1, \ldots, x_p)$. However, a useful way to proceed is to make the strong assumption of an additive form:

regression

$$f(x_1, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j). \tag{15.6}$$

With this restriction, spline smoothing (or alternative smoothing methods) may be applied to each variable successively in order to fit the model

$$Y_i = \sum_{j=1}^{p} f_j(x_j) + \epsilon_i \tag{15.7}$$

under the usual assumptions for linear regression. More specifically, an iterative algorithm may be used[3] to find the least-squares fit when a spline basis represents each function $f_j(x_j)$.

Example 15.3 Decoding natural images from V1 fMRI Kay et al. (2008) showed that natural images could be identified with above-chance accuracy from V1 activity picked up in fMRI responses. Vu et al. (2011) re-analyzed the data and showed how decoding accuracy could be improved by 30 % when additive models of the general form (15.7) were used. Kay et al. had applied a model of fMRI activity in a V1 voxel based on *Gabor wavelet filters*. Briefly, as shown in Fig. 15.8, a Gabor wavelet is a product of a sinusoidal factor and a factor based on a Gaussian (normal) pdf (see Section 15.2.8). The Gaussian factor is similar to that used in the hippocampal place cell model in (14.21). It has the effect of producing a response, for a particular voxel, based only on a small region in the visual image. The sinusoidal factor produces a central peak together with neighboring troughs that represent lateral inhibition, as is characteristic of the response of V1 neurons. The response due to each filter also has a particular orientation. The activity of each voxel in response to a particular image was regressed on filtered representations of the image. A set of 48 Gabor filters at 8 orientations and 6 spatial scales, as shown in Fig. 15.8, was used. Each image in the stimulus set produced a set of magnitudes $x_j(v)$, with $j = 1, \ldots, 48$, corresponding to the 48 filters, for each voxel v. These were the explanatory variables in the regression model, while the fMRI voxel activity was the response. Due to visible nonlinearities, Kay et al. performed a version of least squares based on $\sqrt{x_j(v)}$. Vu et al. found substantial nonlinearity in the residuals from the model of Kay et al., see Fig. 15.9. They then applied a model of the form (15.7) based on splines having 9 knots placed at the 10th, 20th, ..., 90th percentiles of each explanatory variable. Because they had relatively large numbers of regression variables for each voxel, they applied a version of L1 penalized regression (see p. 358). The resulting additive model greatly improved the residual plots, see[4] Fig. 15.10. Vu et al. also showed that the additive model is more sensitive to weak stimuli, and this has the effect of brodening voxel tuning in space, frequency, and contrast. This, presumably, was the main source of improved performance. □

broadening
a

Equation (14.13) may be generalized to

$$
\begin{aligned}
Y_i &\sim f_{Y_i}(y_i|\eta_i) \\
g(\mu_i) &= \sum_{j=1}^{p} f_j(x_j)
\end{aligned}
\tag{15.8}
$$

[3] One method, known as *backfitting*, cycles through the variables x_j, using smoothing (here, spline smoothing) to fit the residuals from a regression on all other variables.

[4] There remain upward trends in the residual plots. This is due to the penalized fitting, which induces correlation of residuals and fitted values.

15.3.3 Theoretical considerations lead to bandwidth recommendations for linear smoothers.

Recall, from Section 8.1.1, that $MSE = \text{Bias}^2 + \text{Variance}$. A minimal requirement of an estimator, in large samples, is that its bias and variance vanish (as $n \to \infty$). Consider estimation of $f(x)$ at the single point x. A linear smoother is, at x, a linear combination of the data response values y_i, so that the estimator may be written in the form

$$\hat{f}(x) = \sum_{i=1}^{n} w_i(x) y_i$$

where $w_i(x)$ emphasizes that the weights are determined for each x. We want

$$E(\hat{f}(x)) \to f(x) \tag{15.15}$$

and

$$V(\hat{f}(x)) \to 0. \tag{15.16}$$

Because $E(Y_i) = f(x_i)$ we also have

$$E\hat{f}(x) = \sum_{i=1}^{n} w_i(x) f(x_i)$$

so that the bias vanishes, as stated in (15.15), if the weights $w_i(x)$ become concentrated near x and the function $f(x)$ is smooth. For the weights to become concentrated it is sufficient that

$$\sum_{i=1}^{n} (\overset{x_i}{\boxed{i}} - x)^2 w_i(x) \to 0. \qquad \leftarrow \boxed{(x_i - x)^2 w_i(x)}$$

Assuming $V(Y_i) = \sigma^2$ (or, at least, that the variances do not vary rapidly), the variance vanishes if

$$\sum_{i=1}^{n} w_i(x)^2 \to 0.$$

Conditions like these on the weights, to guarantee (15.15) and (15.16), need to be assumed by any large-sample theoretical justification of a linear smoothing method. An explicit expression for the MSE of kernel estimators was given by Gasser and Muller (1984). This allows a theoretical bias versus variance trade-off, i.e., a formula for bandwidth selection as a function of n.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{16.10}$$

Details: Taking logs in Eq. (16.6) and inserting $\sigma_{\bar{x}}$ according to (16.10), we have

$$\log f(\theta|x) = -\frac{1}{2}\left(\frac{\bar{x}-\theta}{\sigma\bar{x}}\right)^2 - \frac{1}{2}\left(\frac{\theta-\mu_\pi}{\sigma_\pi}\right)^2 + \text{constant}$$

where "constant" refers to all terms that are constant in θ. We then note that

$$\frac{1}{\sigma_{\bar{x}}^2}(\theta-\bar{x})^2 + \frac{1}{\sigma_\pi^2}(\theta-\mu_\pi)^2 = (\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_\pi^2})(\theta-\mu_{post})^2 + \text{constant}$$

where μ_{post} is given by (16.8). Therefore, we have

$$\log f(\theta|x) = (\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_\pi^2})(\theta-\mu_{post})^2 + \text{constant}$$

and, exponentiating,

$$f(\theta|x) \propto \exp\left((\frac{1}{\sigma_{\bar{x}}^2} + \frac{1}{\sigma_\pi^2})(\theta-\mu_{post})^2\right)$$

which shows the posterior is normal with mean μ_{post} and variance σ_{post}^2 given by (16.9). □

Equation (16.8) has a deeply useful interpretation. Let us rewrite it in the form

$$\mu_{post} = w\bar{x} + (1-w)\mu \tag{16.11}$$

superscript 2 missing

$\sigma_{\bar{x}}^{2}$

where

$$w = \frac{\sigma_\pi^2}{\sigma_{\bar{x}} + \sigma_\pi^2}.$$

In the special case $n = 1$ we write $x = x_1$ and get

$$\mu_{post} = wx + (1-w)\mu. \tag{16.12}$$

Equations (16.11) and (16.12) say that the posterior mean is a weighted combination of the MLE and the prior mean, with the weights determined by the relative precision (the inverse of the variance) of data and prior. In (16.11), as the precision in the data increases relative to the prior (i.e., as $\sigma_\pi^2/\sigma_{\bar{x}}^2$ increases), w increases, more weight is placed on $\bar{x}$, and the posterior mean becomes nearly the same as $\bar{x}$. Intuitively, when the weight w is large, the data contribute more knowledge than the prior, and so the posterior is centered near the data value $\bar{x}$. When the data are imprecise relative to the prior (i.e., when $\sigma_\pi^2/\sigma_{\bar{x}}^2$ is small), more weight is placed on the prior mean, so that the posterior mean is pulled away from $\bar{x}$ and toward the prior mean. The posterior mean is often said to *shrink* the value $\bar{x}$ toward μ, particularly when $\mu = 0$ (so that the magnitude of μ_{post} is smaller than that of $\bar{x}$). In this terminology, the amount of *shrinkage* is determined by $1 - w$. In Section 16.2.3 we discuss the connection between the use of "shrinkage" in this context and in regression (see p. 357).

magnitude

Example 16.1 Sensorimotor learning Körding and Wolpert (2004) designed an experiment in which visual input could be combined with a learned prior distribution in order to produce a finger movement. Subjects moved their index finger from a starting location toward a target, which was represented on a computer monitor. Halfway through the finger movement they were given visual feedback as to where their finger was at that moment (a cursor was shown briefly on the monitor) but, relative to a straight path between starting location and target, it was (a) corrupted by noise and (b) displaced to the right. The noisy location was indicated by a cloud of points drawn from a spherical bivariate normal distribution with one of 4 possible values of standard deviation (the standard deviation here refers to the standard deviation of each marginal distribution determined by the bivariate normal). This standard deviation would correspond to $\sigma_{\bar{x}}$ in Eqs. (16.8) and (16.9), and the center of the displayed cloud of points would correspond to $\bar{x}$. The size of the displacement varied with each trial, and was drawn from another normal distribution. The mean and standard deviation of this displacement distribution would correspond to μ_π and σ_π in Eqs. (16.8) and (16.9). In other words, the displacement distribution formed a prior and the center of the cloud of points (together with the standard deviation) became the subject's input data for each trial.

corrupted

Subjects were given 1,000 training trials during which they could learn the prior displacement distribution. When queried afterwards they had no awareness of the displacement. The authors used an additional 1,000 trials to collect experimental data about the final location of each subject's finger. The authors showed that the displacement of the final location from the target was predicted well by Eqs. (16.8) and (16.9). In other words, in attempting to reach the target, subjects combined the visual input information with their prior knowledge of the displacement, at least approximately, as if their nervous system were computing a posterior mean according to Eqs. (16.8) and (16.9). □

Formulas analogous to Eq. (16.8) also hold for other exponential families with conjugate priors. For example, in the binomial setting let us reparameterize the *Beta*(α, β) distribution by defining

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{16.13}$$

$$\nu = \alpha + \beta. \tag{16.14}$$

e

so that (16.20) holds due to (16.19) (by an application of Slutszky's theorem). Slutsky's
Because the observed information increases as $n \to \infty$, the loglikelihood function becomes more highly peaked about its maximum as $n \to \infty$. Furthermore, the likelihood function may be approximated by the normal pdf found by exponentiating the right-hand side of (16.16), which has standard deviation $SE = I_{OBS}(\hat{\theta})^{-1/2}$ (the standard error associated with the MLE). From Eq. (16.20), this standard deviation is decreasing as $1/\sqrt{n}$. Therefore the width of the peak in the likelihood function is decreasing at the rate of $1/\sqrt{n}$. We can write the values of θ for which the likelihood is substantial in the form

$$(a_n, b_n) = (\hat{\theta} - c \cdot SE, \hat{\theta} + c \cdot SE) \tag{16.21}$$

where we take c to be a positive constant, such as $c = 4$.

Finally, we consider the contribution of the prior to the posterior as $n \to \infty$. As in (16.19), the posterior pdf is a normalized product of the likelihood function and prior pdf. While the likelihood function becomes increasingly close to the form of a normal pdf, with standard deviation decreasing as $1/\sqrt{n}$, the prior pdf $\pi(\theta)$ is a fixed function that does not change with n. The intervals in (16.21) will have lengths $b_n - a_n$ that decrease as $1/\sqrt{n}$ and, for θ in these intervals, when n is large the value of $\theta - \hat{\theta}$ will be small so that we get

$$\pi(\theta) \approx \pi(\hat{\theta}). \tag{16.22}$$

In other words, for values "within the peak" of the loglikelihood function, the prior is approximately constant. Therefore, (16.16) gives us (16.17) and the posterior becomes concentrated near θ with an approximately normal form, as in (16.18). We have sketched the argument in the scalar case, but the steps are the same when θ is a vector.

The approximate $N(\hat{\theta}, I_{OBS}(\hat{\theta})^{-1})$ distribution of the posterior not only gives an easy way to compute posterior probabilities, for large samples, but it also provides a very nice mathematical expression of one of the guiding principles of science: any two investigators who start with differing beliefs (in the form of two different priors $\pi_1(\theta)$ and $\pi_2(\theta)$), will, with sufficiently much data, come to agreement (their posterior distributions will be essentially the same).

16.1.6 *Powerful methods exist for computing posterior distributions.*

In our introduction to this chapter we reviewed briefly the conceptual appeal of Bayesian inference. Bayesian methods have become indispensible in the analysis of indispensable neural data mainly because (i) inferences agree reasonably well with those based on ML estimation (due to the result outlined in Section 16.1.5), (ii) sometimes there

a

is available structure that can be formalized as part of a prior specification (see Section 16.2), and (iii) in many complicated statistical models there are general computational tools for computing posteriors. In this section we sketch the essential ideas behind the main such computational tool, *posterior simulation*.

We have already, in Chapter 9, described the great utility of simulation methods in statistical inference. In posterior simulation a sequence $\theta^{(1)}, \ldots, \theta^{(G)}$ of observations from the posterior distribution is generated, and inference is based on the methods outlined in our discussion of simulation-based propagation of uncertainty (p. 225). For example, to compute the probability in

$$P(a < \theta < b|x) = \int_a^b f(\theta|x)d\theta \tag{16.23}$$

we could use

$$P(a < \theta < b|x) \approx \frac{N_1}{G}$$

where N_1 is the number of $\theta^{(g)}$ such that $a < \theta^{(g)} < b$. Similarly, if $\phi = f(\theta)$ for some function $f(x)$ we could compute probabilities involving ϕ (again, as on p. 225), as

$$P(c < \phi < d|x) \approx \frac{N_2}{G}$$

where we let $W^{(g)} = f(\theta^{(g)})$ and N_2 is the number of $W^{(g)}$ such that $c < W^{(g)} < d$. This kind of computation is used in the following example.

Example 16.2 Methylphenidate-Induced Emergence from General Anesthesia When general anesthesia is administered for surgery, or for an invasive diagnostic procedure, patients recover by resting until the anesthesia's effects wear off. As an alternative, Solt et al. (2011) considered the possiblity that methylphenidate might induce emergence from general anesthesia. Methylphenidate (Ritalin) is widely used to treat Attention Deficit Hyperactivity Disorder (ADHD), and acts primarily by inhibiting dopamine and norepinephrine reuptake. But dopamine and norepinephrine can also promote arousal. The authors applied isoflurane anesthesia to rats at a dose sufficient to maintain them in a supine position (lying down) for 40 min. Five minutes after establishing their anesthetized state (from an equilibration procedure) the animals were given one of three doses of methylphenidate intraveneously ranging from a maximum of 5 mg/kg to a minimum of .05 mg/kg. At the maximum dose, 12 out of 12 rats regained their upright position and made purposeful movements within 30 s of drug administration. At the minimum dose, 0 out of 6 regained their upright position within 30 min. Apparently, 5 mg/kg of methylphenidate is sufficient to remove the immobilizing effects of isoflurane-induced anesthesia in rats. (At the intermediate dose of .5 mg/kg 11 out of 12 regained their upright position.)

possibility

i

intravenously

To evaluate the strength of this evidence, 12/12 versus 0/6, the authors considered the binomial model $X_1 \sim B(12, p_1)$ and $X_2 \sim B(6, p_2)$, introduced independent uniform priors on p_1 and p_2 as in Example 1.4 on p. 174, and computed the posterior probability $P(p_1 > p_2|X_1 = 12, X_2 = 0)$. This may be done very easily by posterior simulation: the two posterior distributions on p_1 and p_2 are $Beta(13, 13)$ and $Beta(1, 7)$, and they are independent. (It is easy to check that if X_1 and X_2 are independent, and the prior distributions on p_1 and p_2 are independent, then the posterior distributions on p_1 and p_2 are independent.) We therefore do the following:

1. Draw $G = 10,000$ observations from a $Beta(13, 13)$ distribution and put them in a vector A.
2. Draw $G = 10,000$ observations from a $Beta(1, 7)$ distribution and put them in a vector B.
3. Compute the number of components i for which $A[i] > B[i]$, and divide by G. This is, approximately, the desired posterior probability.

Performing the calculation gives $P(p_1 > p_2|X_1 = 12, X_2 = 0) = .986$. The authors concluded that methylphenidate actively induces emergence from isoflurane anesthesia. We re-evaluate this evidence using Bayes factors on p. 478. □

In Section 16.1.2 we noted that posterior probabilities may be computed easily when conjugate priors are used, and Example 16.2 made use of posterior simulation with conjugate beta posterior distributions. As soon as we leave conjugacy, numerical difficulties become apparent. Even in the simple case of estimating a normal mean θ from a sample $X_1, \ldots, X_n$, with $X_i \sim N(\theta, \sigma^2)$ and σ known, if we take the prior to be a non-normal probability distribution on $(-\infty, \infty)$, the posterior pdf becomes intractable, in the sense that ~~the product~~ $L(\theta)\pi(\theta)$ in Eq. (16.1) ~~(see p. 440)~~ form, and we can not evaluate analytically the integrals needed to compute posterior probabilities such as that in (16.23). The usual approach to solving this problem is to apply posterior simulation based on *Markov chain Monte Carlo* (*MCMC*).

has a non-normal form

has a non-normal

The nomenclature is descriptive of the idea behind MCMC: "Monte Carlo" refers to[7] simulation methods, and "Markov chain Monte Carlo" indicates that the approach is based on Markov chains. To explain, we begin by returning to an example.

Example 3.5 (Continued, see p. 58) In our discussion of Colquhoun and Sakman's results on ion channel openings we noted from Fig. 3.8, panel B, the good fit of an exponential distribution to the histogram of open durations, when there was only one opening in an activation burst. The major purpose of the paper was to demonstrate the existence of activation bursts. Let us, however, ignore bursts and imagine an ideal ion channel that opens and closes with open and closed durations governed by exponential distributions. The defining property of exponential distributions is that they are memoryless (see the theorem on p. 120). Now consider an ion channel that is observed to be either open or closed for a sequence of discrete time values, e.g., every ms for 10 min, and let $X_t = 1$ if it is open at time t and $X_t = 0$ if it is closed at time t. We refer to the channel's *state* at time t as the value of X_t, with 1 or 0

[7] When computer-based simulation methods were first being used, Monte Carlo was the site of a famous gambling establishment, which was frequented by the uncle of one of the developers of these methods. See Metropolis (1987).

signifying either open or closed. If we assume[8] the ion channel is memoryless, then its state at time $t + 1$ will depend on its state at time t, but not on any of the preceding states prior to time t. There are then four possiblities: the channel can be closed at time t and stay closed at $t + 1$, it can be closed at t and be open at $t + 1$, it can be open at t and close at $t + 1$, or it can be open at t and stay open at $t + 1$. The four possibilities have conditional probabilities given by $P(X_{t+1} = j|X_t = i)$ where i and j can take values 0 or 1. □

possibilities

Abstracting from this example, suppose we have a sequence of random variables $X_1, X_2, \ldots, X_t, \ldots$, which take values 0 and 1, and suppose further that

P(X_{t+1}|X_t)

$$P(X_{t+1}|X_1, X_2, \ldots, X_t) = P(X_{t+1} \cancel{= j}|X_t \cancel{= i}) \tag{16.24}$$

and that these conditional probabilities are time-invariant in the sense that

$$P(X_{t+1} = j|X_t = i) = P(X_{s+1} = j|X_s = i)$$

for all $s, t = 0, 1, 2, \ldots$. Then the sequence $X_1, X_2, \ldots, X_t, \ldots$ is said to form a two-state *Markov chain* having *transition probabilities* $P(X_{t+1} = j|X_t = i)$, which we write as

$$P_{ij} = P(X_{t+1} = j|X_t = i). \tag{16.25}$$

Let us note that (16.25) implies

$$P(X_{t+1} = 0) = P(X_t = 0)P_{00} + P(X_t = 1)P_{10} \tag{16.26}$$

$$P(X_{t+1} = 1) = P(X_t = 0)P_{01} + P(X_t = 1)P_{11}. \tag{16.27}$$

The definition extends immediately to the case of m states, for m an integer with $m \geq 2$.

The key property of a Markov chain is its lack of memory: the probability of being in state j at time $t + 1$ depends only on the state of the chain at time t. Under some mild conditions[9] it is possible to say something about the long-run behavior of the chain. In the case of the ideal ion channel considered above, we may ask for the probability that the channel is open at time t = 600,000, corresponding to 10 min after the commencement of observation. In principle this probability depends on the initial condition, whether the channel was open at time $t = 1$. However, because the state at time t = 600,000 is the result of 599,999 random draws from the distributions given by the transition probabilities (16.25) the influence of the intial

initial

[8] Because we are assuming discrete time the memoryless distribution of durations becomes geometric rather than exponential, as we noted on p. 120.

[9] The chain must be *irreducible* (if the chain is in state i at time t it is possible for it to get to state j in the future), *aperiodic* (the chain does not cycle deterministically through the states), and *recurrent* (if the chain is in state i at time t it will eventually return to state i in the future), see for example, Ross (1996, Theorem 4.3.3).

(because the Q_{ji} that appear on the right-hand side of (16.31) are simply those that appear on the left-hand side when the inequality is reversed). We then solve for the values of α_{ij} that produce equality in (16.31). That is, we write

$$P_\infty(i)Q_{ij}\alpha_{ij} = P_\infty(j)Q_{ji}$$

and solve for α_{ij}:

$$\alpha_{ij} = \frac{P_\infty(j)}{P_\infty(i)}\frac{Q_{ji}}{Q_{ij}}. \qquad (16.32)$$

Note that $\alpha_{ij} < 1$ if and only if (16.31) holds.

We have now produced a specification of the transition probabilities required for a chain having the target distribution P_∞ as stationary distribution: if (16.31) holds, set $P_{ij} = Q_{ij}\alpha_{ij}$, where α_{ij} are defined by (16.32), otherwise set $P_{ij} = Q_{ij}$. To create a chain with these transition probabilities is easy. Suppose the chain is in state i. If we accept the candidate (which is generated from the chain having transition probabilities $\{Q_{ij}\}$) with probability α_{ij}, then the probability of moving to state j will be $Q_{ij}\alpha_{ij}$. Thus, we use the following scheme:

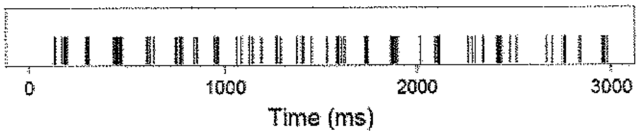
```
define $\alpha_{ij}$by(16.32)
if  $\alpha_{ij} < 1$  then accept the candididate with probability  $\alpha_{ij}$  candidate
otherwise   accept the candidate.
```

This is the Metropolis-Hastings algorithm. □

16.2 Latent Variables

When we introduced the concept of random variable (on p. 46) we were careful to distinguish the mathematical object from the data: we said that random variables and their probability distributions live in the theoretical world of mathematics while data live in the real world of observations. Random variables that are theoretical counterparts of observed data are sometimes called *observable*. But it is also possible to insert into a statistical model random variables that affect the distribution of the observable random variables without themselves representing data; instead they represent *unobserved*, hypothetical quantities. Such unobservable random variables are called[13] *latent variables*. Models that incorporate latent variables can be powerful

[13] The noise random variable ϵ_i in the regression model (12.1) is unobservable, but would not typically be called latent. To exclude such cases a random variable could be called latent only if it can not be written in terms of observable random variables. Thus, under this definition, because (12.1) implies $\epsilon_i = Y_i - f(x_i)$, ϵ_i would not be a latent variable. See Bollen (2002).



α
α

Fig. 16.1 First 3 s of spikes recorded over about 30 s in vitro, from a goldfish retinal ganglion ~~cell~~ neuron. Data from Levine (1991), furnished by Satish Iyengar; see Iyengar and Liao (1997). ~~See~~ These data are discussed in Example 19.1.

and intuitive ways to describe variation in the data. We discussed the mixture-of-two-Gaussians model on p. 216, and we will return to mixtures of Gaussians in Section 17.4.3. Here is another example.

Example 16.3 Burst Detection from spike trains In many contexts neurons exhibit burstiness, meaning that action potentials (also called "spikes," see p. 3), appear across time in small clusters, or bursts. For instance, burstiness of dopamine neurons in the midbrain is believed to be a functionally relevant signal indicating reward and goal-directed behavior (see Grace et al. (2007)). In the analysis of bursting neural spike trains, the epochs during which the neuron is bursting must somehow be inferred from the data. In Fig. 16.1, for example, due to the inherently erratic nature of the spiking, it is not always obvious whether the neuron is in a burst or not, or where a burst begins and ends.

To provide an algorithm together with statistical inferences, Tokdar et al. (2010) described bursty neurons by introducing a latent binary variable, which was 1 when the neuron was bursting and 0 when it was not bursting. Let us assume the recording time to occur in discrete steps $t = 1, 2, \ldots, T$, and define the random variable Y_t be 1 if a spike occurs at time t and 0 otherwise. Tokdar et al. discussed several alternative models. The simplest uses latent variables X_t that take the value 1 if the neuron is bursting at time t and 0 if non-bursting, and assumes that Y_t has a Bernoulli pdf with mean θ_1 if $X_t = 1$ and with mean θ_0 if $X_t = 0$. Here, θ_1 and θ_0 represent the firing rates of the neuron when bursting and when not bursting, and if θ_1 is much larger than θ_0 the neuron will tend to fire in rapid succession when $X_t = 1$, compared with its slower rate when $X_t = 0$. This describes the tendency to produce bursts. By introducing probability distributions for the latent variables X_t, and then estimating the value of each X_i, it is possible to infer where in time the bursts occurred.[14] □

In most statistical models the distribution of the random variables representing the data depends on some unknown parameters. In the model cited in Example 16.3 the distributions of the random variables representing the data depended on unknown parameters, but they also depended on the latent variables. The point is that the latent variables themselves followed probability distributions. In other words, one set of probability distributions—those describing the variation in the data—depended on random variables following another set of probability distributions, which described

[14] To speed computation Tokdar et al. chose to work with the inter-spike intervals instead of the variables Y_t we have defined here.

These data are discussed in Example 19.1.

variation among certain theoretically interesting but unobserved quantities, namely the bursting or non-bursting status of the neuron within each ISI.

The parameters in statistical models are usually fixed coefficients (though they are typically unknown, and therefore estimated from the data). In Section 16.2.1 we describe models in which the parameters become random variables, and thus latent variables. We then briefly re-interpret penalized regression in Section 16.2.3 and return to the general structure underlying Example 16.3 in Section 16.2.4.

16.2.1 *Hierarchical models produce estimates of related quantities that are pulled toward each other.*

Nearly all the statistical models we have considered[15] begin with a parameterized family of probability densities $f(x|\theta)$, and the first statistical problem is to determine from the data x the likely values of the parameter θ. Sometimes there is an obvious source of variability among values of the parameter θ, as when θ could vary from subject-to-subject, or neuron-to-neuron, etc. In such cases we may introduce a second layer into the statistical model by considering a family of densities $f(\theta|\lambda)$. For generality, we will refer to individual subjects or individual neurons, etc., as *units*. In other words, we will say that we are interested in the variation of some parameter θ across units. In neuroimaging, for example, we might have task-related effects at particular voxels whose magnitude varies across subjects, and these could be assumed to follow some probability distribution. In analyzing neural responses, the way a particular measure of neural activity varies across neurons may be of interest, and might be assumed to follow a given probability distribution. In these situations we introduce both a probability density $f(x|\theta)$ for the data given a parameter vector θ and a probability density $f(\theta|\lambda)$ for θ that itself depends on a parameter λ. Such a specification is called a two-stage[16] *hierarchical model.*

Example 12.3 (continued from p. 331) as described previously, Behseta et al. considered spike counts from 54 neurons during performanace of a serial-order eye-movement task, and the authors computed a rank order selectivity index

performance

$$I_{\text{rank}} = \frac{(f_3 - f_1)}{(f_3 + f_1)}$$

where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively, the mean being taken across trials. As part of the analysis, the rank selectivity indices across neurons were considered to follow a normal distribution. Let X_i represent I_{rank} for neuron i. Behseta et al. assumed a model of the

[15] Nonparametric methods (Section 13.3) are based on statistical models of a more general form that do not depend on a finite-dimensional parameter vector.

[16] In principle this process can continue, with λ distributed according to a family of densities, and so on, but they do not arise very often in practice.

added that Behseta et al. developed a method to correct for the attenuation and when they applied it to these data the new estimate of correlation was .82, which was more reasonable. We now provide some details about the method.

On p. 459 we let X_i be the random variable representing I_{rank} for the ith neuron and we said that Behseta et al. used the normal hierarchical model (16.33) and (16.34). Let us reformulate (16.33) by writing

$$\begin{aligned} X_i &= \theta_i + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma_{\epsilon_i}^2) \end{aligned}$$

and then let Y_i represent the value of I_{reward} for the ith neuron and write

$$\begin{aligned} Y_i &= \xi_i + \delta_i \\ \delta_i &\sim N(0, \sigma_{\delta_i}^2). \end{aligned}$$

Here θ_i and ξ_i represent the theoretical values of I_{rank} and I_{reward} for neuron i that would be obtained from noiseless measurements (or from infinitely many trials). The quantities σ_{ϵ_i} and σ_{δ_i} are the standard errors associated with x_i and y_i (they were obtained by propagation of uncertainty from the spike count means). Taking ϵ_i and δ_i to be independent we may combine the assumptions on X_i and Y_i by saying these random variables are bivariate normal according to

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N(m_i, V_i), \tag{16.43}$$

Space

where

$$m_i = \begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} \text{and} V_i = \begin{pmatrix} \sigma_{\epsilon_i}^2 & 0 \\ 0 & \sigma_{\delta_i}^2 \end{pmatrix}.$$

Equation (16.43) is the first stage of a bivariate normal hierarchical model. Behseta et al. wrote the second stage in the form

$$\begin{pmatrix} \theta_i \\ \xi_i \end{pmatrix} \sim N(\mu, \Sigma), \tag{16.44}$$

where

$$\mu = \begin{pmatrix} \mu_\theta \\ \mu_\xi \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_\theta^2 & \rho_{\theta\xi}\sigma_\theta\sigma_\xi \\ \rho_{\theta\xi}\sigma_\theta\sigma_\xi & \sigma_\xi^2 \end{pmatrix}$$

with $\mu_\theta, \mu_\xi, \sigma_\theta^2$, and σ_ξ^2 being the means and the variances of θ_i and ξ_i respectively. The quantity of interest is $\rho_{\theta\xi}$, which represents the correlation between the theoretical values θ_i and ξ_i. Let us refer back to the theorem on attenuation of the correlation on p. 330. In the notation used here, that theorem says that if $\rho_{\theta\xi} > 0$ then

$$\rho_{XY} < \rho_{\theta\xi}.$$

where $P(H_A) = 1 - P(H_0)$. When this formula is applied to statistical models under H_0 and H_A, as in Section 11.1.6, $P(data|H_0)$ becomes a discrete or continuous pdf for a random vector X, so we substitute

Space

$$f_0(x) = P(data|H_0) \text{ and} f_1(x) = P(data|H_A),$$

where we are using the subscript 1 to signify the alternative, and we write

$$P(H_0|x) = \frac{f_0(x)P(H_0)}{f_0(x)P(H_0) + f_1(x)P(H_A)}. \tag{16.62}$$

As we pointed out in Section 11.1.6, p. 297, the prior probability $P(H_0)$ may be removed by considering instead the *Bayes factor*, which is the ratio of posterior odds to prior odds,

$$BF_{01} = \frac{P(H_0|x)}{P(H_A|x)} \div \frac{P(H_0)}{P(H_A)} \tag{16.63}$$

and from

$$\frac{P(H_0|x)}{P(H_A|x)} = \frac{f_0(x)P(H_0)}{f_1(x)P(H_A)}$$

we have

$$BF_{01} = \frac{f_0(x)}{f_1(x)}. \tag{16.64}$$

The subscript on BF_{01} indicates that we are considering the Bayes factor in favor of H_0. Its reciprocal, BF_{10}, would be the Bayes factor in favor of H_A. In Section 16.3.1 we describe the way the Bayes factor quantifies evidence in favor of a hypothesis, in Section 16.3.2 we review briefly the contribution of Bayes factors and posterior probabilities to epistemology, in Section 16.3.3 we issue a note of caution concerning the strong dependence of Bayes factors on prior distributions, and in Section 16.3.4 we discuss their use in calibrating p-values.

Bayes factors were first discussed by Harold Jeffreys, who saw them as a way of evaluating the strength of evidence in favor of a new scientific theory (represented by H_A) that might replace an old one (H_0). A modern view was provided by Kass and Raftery (1995). Jeffreys (1961, Appendix B) suggested interpreting BF_{10} (the evidence in favor of the new theory) in half units on the $\log_{10}$ scale. Although probability itself provides a meaningful scale, as do the odds, Jeffreys felt it was useful to provide a rough statement about standards of evidence in scientific practice. Table 16.2 is a mildly modified version of his interpretive categories (taken from Kass and Raftery (1995)). Interpretation may depend on context, but these categories remain useful. They are stated in terms of BF_{10} because weighing evidence *against* a null hypothesis is more familiar, but Bayes factors can equally well provide evidence *in favor of* a null hypothesis. Indeed, this is one of the strengths of the Bayesian approach. We illustrate by returning to Example 16.4 in Section 16.3.1.

$H_0 : \theta_i = 0$. The results in Table 16.1 suggested that strain 12 might satisfy this hypothesis, i.e.,

$$H_0 : \theta_{12} = 0. \tag{16.68}$$

We now consider the evidence in favor of H_0 defined by (16.68), presenting results reported in Kass and Raftery (1995).

Under H_0 the data random variable X_{12} follows a normal distribution with mean 0 and known variance σ_{12}^2, and so the numerator of BF_{01} has the form of the numerator in (16.67). Under H_A we may assume $X_{12} \sim N(\theta_{12}, \sigma_{12}^2)$ and we now must choose $\pi_1(\theta_{12})$, which appears in the denominator of (16.67). Because, under H_A, strain 12 would be judged similar to all the other strains, we may use the second-stage normal distribution that appeared in the hierarchical model considered previously (p. 462), i.e., the pdf takes the form

$$\pi_1(\theta_{12}) = n(\theta_{12}; \mu, \tau^2) \tag{16.69}$$

where μ and τ are found from the data involving strains $j \neq 12$ (using ML estimation, as discussed in Section 16.2.1). Kass and Raftery reported that when this was done, the Bayes factor was

$$B_{01} = 15$$

indicating that these data produced[29] strong evidence in favor of H_0. □

Example 16.2 (continued) On p. 451 we described the way Solt et al. (2011) used the posterior probability $P(p_1 > p_2 | X_1 = 12, X_2 = 0)$ to judge their result that 12 out of 12 rats regained their upright position following a substantial dose of methylphenidate whereas 0 out of 6 did following a negligible dose. We may instead use Bayes factors.

We begin by considering the hypotheses to be tested. The data 0 out of 6 confirmed that the very small dose of methylphenidate left the rats unable to regain their upright position. If p is the probability of regaining upright position we might want to take $H_0 : p = 0$ and $H_A : p \neq 0$. Under H_0 the outcome 0 out of 6 has probability 1. Under H_A we may introduce a uniform prior on [0, 1] for the unknown value of p. Using $\binom{6}{0} = 1$, the Bayes factor in (16.65) becomes

Eq. (5.16)

6

$$BF_{01} = \frac{1}{\int_0^1 p^0(1-p)^6 dp}$$

and, from Eq. (5.1$\overset{9}{5}$), the denominator integral is equal to $6!/7! = 1/7$ and we get

[29] A possible issue is the extent to which strain 12 was selected *post hoc*, after the data had been examined. It is possible to correct the Bayes factor for such *post hoc* selection, analogously to (though differently than) they way p-values may be adjusted (see Section 11.3). The investigators repeated the experiment on strain 12 and found similar results, which provided strong confirmation of H_0.

$$BF_{01} = .00026$$

which conforms with the intuition that the evidence is overwhelmingly in favor of an effect of methylphenidate in enabling rats to regain an upright position. □

16.3.2 *Bayes factors provide an interpretation of scientific progress.*

At the end of Section 16.1.5 we said that the approximate $N(\hat{\theta}, I_{OBS}(\hat{\theta})^{-1})$ distribution of the posterior provides an expression of one of the guiding principles of science, namely that investigators with different knowledge or opinions will eventually come to agreement after taking into account sufficiently much data. This concerns the value of a parameter θ. An analogous statement can be made concerning the scientific law that describes a particular phenomenon. Following Eq. (11.12) we noted that BIC is a consistent model selection procedure in the sense that, for sufficiently large samples the probability of BIC choosing the correct model will get arbitrarily close to one. By virtue of (11.12) the same is true of Bayes factors.[30] To re-phrase this fundamental result in terms of posterior probability, suppose we have a set of m candidate models M_k, with m being as large as we like, and suppose further that we place positive prior probabilities $P(M_k)$ on them. For sufficiently much data, the posterior probability on the correct model will get arbitrarily close to one. This means that investigators having different opinions about the merits of competing scientific laws (represented as statistical models) will eventually come to agreement after taking into account sufficiently much data.

The result was recognized by Jeffreys and Wrinch (1921), and was a primary motivation for Jeffreys' monumental treatise *Theory of Probability*. In the preface to the first edition of his book (in 1939) he wrote:

> In opposition to the statistical school, [physicists] and some other scientists are liable to say that a hypothesis is definitely proved by observation, which is certainly a logical fallacy; most statisticians appear to regard observations as a basis for possibly rejecting hypotheses, but in no case for supporting them. The latter attitude, if adopted consistently, would reduce all inductive inference to guesswork; the former, if adopted consistently, would make it impossible ever to alter the hypotheses, however badly they agreed with new evidence.... In the present book I ... maintain that the ordinary common-sense notion of probability is capable of precise and consistent treatment when once an adequate language is provided for it. It leads to the results that a precisely stated hypothesis may attain either a high or a negligible probability as a result of observational data.

impossible

[30] Mathematically the situation is reversed: an elegant theorem due to Doob establishes the consistency of the posterior distribution, and thus of Bayes factors, under weak conditions. Equation (11.12) then provides consistency of BIC. For precise statements see Schervish (1995, Section 7.2.1) and the references in Kass and Raftery (1995, Section4.1.3).

16.3.4 Bayes factors can be used to calibrate p-values.

On p. 282 and 476 we distinguished between the p-value and the quantity $P(H_0|data)$, which is computed from Bayes' theorem. In order to compute $P(H_0|data)$ based on the data $X = x$, according to (16.62), we need $f_0(x)$ and $f_1(x)$. The pdf $f_0(x)$ is used in calculating[31] the p-value. If $f_1(x)$ is either known or assumed known, as in (16.69), the Bayes factor may be computed and if we further take $P(H_0) = P(H_A) = \frac{1}{2}$ then Eq. (16.62) gives

$$P(H_0|x) = \frac{BF_{01}}{1 + BF_{01}}. \tag{16.70}$$

By making assumptions about $f_1(x)$ it therefore becomes possible to compare the p-value with $P(H_0|x)$. This was done by Jeffreys (Jeffreys (1961, pp. 373–374)), and subsequently by Edwards et al. (1963) and others. See Selke et al. (2001) for a thorough discussion. The approach taken by Edwards et al. (1963) and by Selke et al. (2001) was to assume that the pdf f_1 lies in some family of distributions, and for data x such that a given p-value occurred (such as $p = .05$) they then minimized BF_{01} over all possible members of that family. This minimum represents the strongest possible evidence against H_0 that the p-value could provide, under the given assumptions. Under assumptions considered reasonable[32] by Sellke, Bayarri, and Berger, the value $p = .05$ corresponds to a minimum of $BF_{01} = .41$. In other words, under those assumptions, using (16.70), the value $p = .05$ corresponds to $P(H_0|data) \geq .41/1.41 = .29$. Calculations of this sort lead to the general conclusion that $p = .05$ is relatively weak evidence against H_0.

Sellke

l

16.4 Derivations of Results on Latent Variables

Derivation of the results for the normal hierarchical model:
Let us use the notations $x = (x_1, \ldots, x_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$. We begin with

[31] In Eq. (10.24) the statistic Q could follow a standard distribution, such as a t_ν-distribution, in which case the calculation would be based on the distribution of Q. However, Eq. (10.24) may be rewritten as

$$p = \int_R f_0(x)dx$$

where $R = \{x : Q \geq q_{obs}\}$.

[32] For the normal testing problem of Section 10.3.1, one may consider the class of all pdfs that are symmetric around $\mu = \mu_0$, and also have their mode at $\mu = \mu_0$. Selke et al. (2001) reported results based on this assumption. They also considered the distribution of the p-value. Under H_0 this distribution is uniform (see Section 10.4.1) and under H_A they assumed it to take the form $f(p) = \xi p^{\xi-1}$ for some ξ, which provided another way to formalize the family of alternatives and compute the minimum value of the Bayes factor.

Sellke

l

to data will produce latent factors, and the factor loadings become interpretable, this conception is very appealing. It suffers, however, from a serious difficulty: the unknown parameters are the components of the variance matrix $V(X) = \Sigma$ and for any orthogonal matrix P, if we define $B = AP$, using (12.59) and $PP^T = I_m$ we have

$$\begin{aligned} V(BS + \epsilon) = BV(S)B^T + I_m &= API_mP^TA^T + I_m \\ &= AA^T + I_m \\ &= \Sigma. \end{aligned}$$

In other words, we obtain the same variance matrix using both B and A, so an interpretation of factor loadings based on B would be neither more or less valid than an interpretation based on A. There are thus infinitely many equivalent interpretations. Various methods have been used to specify a unique factor loading matrix, but there often remains a degree of arbitrariness that leaves many practitioners wary of resulting interpretations.[6]

A related, but different approach is to begin by allowing the latent vector S to be non-normal, but with independent components, in the linear latent variable model

$$X = AS,$$

where S and X are both m-dimensional and A is taken to be orthogonal. The idea is that the independent components in S would drive the vector X through the linear combinations in A. If S is assumed to be normally distributed, then so is X, and the solution is given by PCA, i.e., S consists of the principal components. However, if S is allowed to be non-normal it can be quite different.

Let us assume the data vector $X = x$ has been standardized (or *pre-whitened*, see p. 557) so that its sample variance matrix is the m-dimensional identity. We wish to find A and s such that $x = As$. By orthogonality $A^TA = I_m$ so that $A^Tx = s$. The matrix A may be defined to minimize the mutual information among the components of $s = A^Tx$, where mutual information is the Kullback-Leibler divergence between the joint pdf and the independence pdf (estimated from the data), as in (4.28). That is, the components of s are taken to be as close to independent as possible, in the sense of mutual information. The resulting procedure is called *independent components analysis (ICA)*. It turns out that minimizing mutual information in A^Ts has the effect of making the distribution of s as far from normal as possible (measured in terms of entropy).

lower case

Example 17.4 Efficient Coding of Natural Sounds Lewicki (2002) used ICA to find components of auditory signals. Some of the components he found from human speech are shown in Fig. 17.4. For comparison, response properties of cochlear neurons are also displayed. There is a qualitative resemblance between the ICA components and the neural response functions. Lewicki argued that ICA may capture an efficient representation of auditory input. □

[6] The most famous example is Spearman's general intelligence index g, which is obtained from factor analysis. See, e.g., Gould (1996); Devlin et al. (1997).

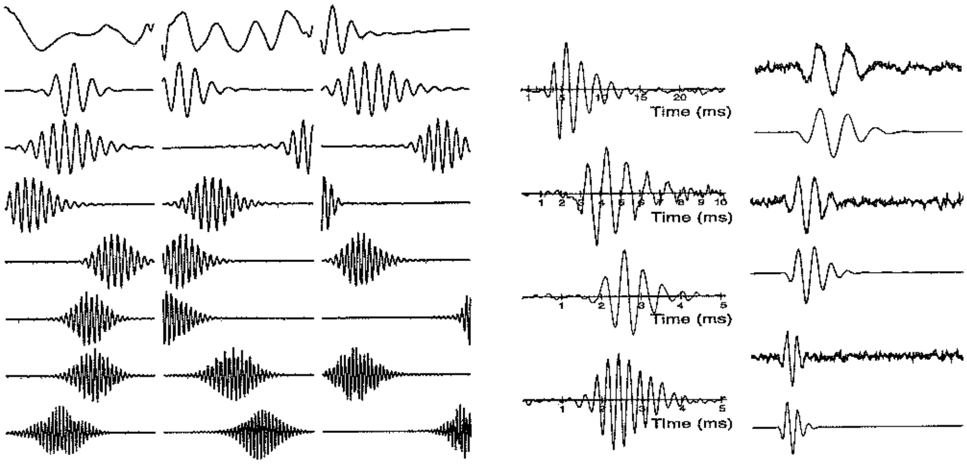


Fig. 17.4 *Left panel* components determined by ICA from human speech. *Right panel* response functions from cochlear neurons. The latter used linear regression of the binary spike train (see Chapter 19) on the input signal at multiple time lags (see p. 530). Adapted from Lewicki (2002).

17.4 Classification and Clustering

17.4.1 Bayes classifiers for multivariate normal distributions take a simple form.

Suppose each of many m-dimensional observation vectors $X = x$ comes from one of K classes $C_1, C_2, \ldots, C_K$, and when it comes from class k the random vector X has pdf $f_k(x)$, for $k = 1, \ldots, K$. The problem of classification (see Section 4.3.4) is to determine, for each observation $X = x$, the class to which x belongs. As we showed in Section 4.3.4, the expected number of classification errors is minimized by using a Bayes classifier. For each x the Bayes classifier finds the class C_k that maximizes the posterior probability given by Eq. 4.38, which we repeat here:

Eq. (4.38)

( )

$$P(C = C_k|X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{m} f_i(x)\pi_i}. \tag{17.22}$$

In the special case where, for each class k, we have $X \sim N_m(\mu_k, \Sigma)$ for some μ_k and Σ, the solution takes a simple form. If we write the ratio of posterior probabilities for two classes j and k by plugging the pdfs given by Eq. (5.17) into (17.22), and take logs, after some algebra we obtain

$$\begin{aligned}\log\frac{P(C = C_j|X = x)}{P(C = C_k|X = x)} &= \log\frac{f_j(x)}{f_k(x)} + \log\frac{\pi_j}{\pi_k} \\ &= \delta_j(x) - \delta_k(x)\end{aligned} \tag{17.23}$$

be related to one another, and thus no longer independent. In this case, specialized methods can produce powerful results. The term *time series*, refers both to data collected across time and to the large body of theory and methods for analyzing such data.

Let us switch over to the general notation for random variables and write a theoretical sequence of measurements as $X_1, X_2, \ldots$, and a generic random variable in the sequence as X_t. Another way to say the X_t variables are dependent is that knowing $X_1, X_2, \ldots, X_{t-1}$ should allow us to predict, at least up to some uncertainty, X_t. Predictability plays an important role in time series analysis.

Space ✓

Example 2.2 (continued from p. 27) On p. 27 we displayed severalEEG spectrograms taken under different stages of anesthesia. We noted earlier that both the roughly 10 Hz alpha rhythm and the 1–4 Hz delta rhythm are visible in the time series plot. In this scenario we can say a lot about the variation among the EEG values based on their sequence along time: in the time bin at time t the EEG voltage is likely to be close to that at time $t-1$ and from the voltage in multiple time bins preceding time t we could produce a good prediction of the value at time t. □

The spectrograms in Example 2.2 display the rhythmic, wave-like features of the EEG signals contrasting them across phases of anaesthesia. They do so by decomposing the signal into components of various frequencies, using one of the chief techniques of time series analysis. The decompositions are possible in this context because the EEGs may be described with relatively simple and standard time series models, but this is not true of all time series. The EEG series are, in a sense, very special because their variation occurs on a time scale that is substantially smaller than the observation interval. By contrast, if we go back to Fig. 1.5 of Example 1.6 we see another time series where the variation is on a longer time scale. The EPSC signal drops suddenly, and only once, shortly after the beginning of the series, then recovers slowly throughout the remainder of the series. In other words, the variation in the EPSC takes place on a time scale roughly equal to the length of the observation interval. Another way to put this is that the EEG at time x_t may be predicted reasonably well using only the preceding EEG values $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$, going back h time bins, where h is some fairly small integer, but a prediction of the EPSC at x_t based on earlier observations would require nearly the entire previous series and still might not be very good. The most common time series methods, those we describe here, assume predictability on relatively short time scales.

anesthesia

So far we have said that the EEG at time x_t may be predicted using the preceding EEG values $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$, but we did not specify which value of t we were referring to. Part of the point is that it doesn't much matter. In other words, it is possible to predict almost *any* x_t using the preceding h observations. (We say "almost" any x_t because we have to exclude the first few x_t observations, with $t \leq h$, where there do not exist h preceding observations from which to predict.) Furthermore, the formula we concoct to combine $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$ in order to predict x_t may be chosen independently of t. This is a very strong kind of predictability, one that is stable across time, or *time-invariant*. The notion of time invariance is at the heart of time series analysis.

We now begin to formalize these ideas. Let X_t be the measurement of a series at time t, with $t = 1, \ldots, n$. Let $\mu_t = E(X_t)$ and $\Sigma_{ij} = Cov(X_i, X_j)$. As soon as we contemplate estimation of this mean vector and covariance matrix we are faced with a serious difficulty. For simplicity consider time t and the problem of estimating μ_t and $\sigma_t^2 = \Sigma_{tt}$. If we have many replications of the measurements at time t (as is usually the case, for example, with evoked potentials) we can collect all the observations across replications at time t and compute their sample mean and sample variance. However, if we have only one time series, and therefore one observation at t, we do not have a sample from which to compute the sample mean and variance. The only way to apply any kind of averaging is by using observations at other values of time. Thus, we can only get meaningful estimates of mean and covariance by making assumptions about the way X_t varies across time. Let us introduce a theoretical time series, or *discrete-time stochastic process* $\{X_t;\, t \in \mathcal{Z}\}$, $\mathcal{Z}$ being the set of all integers. We are now in a position to define the kinds of time invariance we will need. We say that the series X_t is *strictly stationary* if it is time-invariant in the sense that the joint distribution of each set of variables $\{X_t, X_{t+1}, \ldots, X_{t+h}\}$ is the same as that of the variables $\{X_s, X_{s+1}, \ldots, X_{s+h}\}$ for all t, s, h. Because the time index takes all possible integer values it is an abstraction (no experiment runs indefinitely far into the past and future) but it is an extremely useful one. A standard notation in the time series context is $\gamma(s, t) = \Sigma_{st}$. The function $\gamma(s, t)$ is called the *autocovariance function* and the *autocorrelation function* (ACF) is defined by

time

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}}.$$

The prefix "auto," which signifies here that we are considering dependence of the time series on itself, is a hint that one might instead consider dependence across multiple time series, where we would instead have "cross-covariance" and "cross-correlation" functions (which we discuss in Section 18.5). A time series is said to be *weakly stationary* or *covariance stationary* if (i) μ_t is constant for all t and (ii) $\gamma(s, t)$ depends on s and t only through the magnitude of their difference $|s - t|$. This weaker sense of stationarity is all that is needed for many theoretical arguments. Under either form of stationarity we follow the convention of writing the autocovariance function in terms of a single argument, $h = t - s$, in the form $\gamma(h) = \gamma(t - h, t)$. Note that $\gamma(0) = V(X_t)$. It is not hard to show that $\gamma(0) \geq |\gamma(h)|$ for all h, and $\gamma(h) = \gamma(-h)$. In the stationary case the autocorrelation function becomes

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{18.2}$$

Illustration: The 3-point moving average process

$$X_t = \frac{1}{3}(U_t + U_{t-1} + U_{t-2})$$

which is an estimator of the autocorrelation function (18.2).

In this chapter we provide an overview of key concepts in time series analysis. Section 18.2 describes the two major approaches to time series analysis. Section 18.3 gives some details on methods used to decompose time series into frequencies, as in Example 2.2. There are several important subtleties, and we discuss these as well. Section 18.4 discusses assessing uncertainty about frequency components, and Section 18.5 reviews the way these methods are adapted to assess dependence between pairs of simultaneous time series.

18.2 Time Domain and Frequency Domain

In discussing Example 2.2, on p. 514, we alluded to the decomposition of the signal into frequency-based components. In general, time series analysis relies on two complementary classes of methods. As the name indicates, *time domain* methods view the signal as a function of time and use statistical models that describe temporal dependence. *Frequency domain* methods decompose the signal into frequency-based components, and describe the relative contribution of these in making up the signal. In this section we provide a brief introduction to these two approaches, starting with frequency-based analysis. Here are two examples.

Space

Example 18.1 Gamma oscillations inMEG during ~~learning~~ Cortical oscillatory activity in the gamma band (roughly 30–120 Hz) has been associated with many cognitive functions. Chaumon et al. (2009) used MEG imaging to investigate the role of gamma oscillations during unconscious learning. They used a paradigm in which subjects were to find the letter "T" within a set of distractors and determine its orientation. On some trials, which they called "predictive," the distractors were repeated and the location of the "T" remained the same. On other trials, which they called "nonpredictive," the distractors changed configurations and the location of the "T" changed. The subjects were shown many blocks containing 12 trials of each type. Although they remained unaware of the information provided by the configuration type, their reaction time decreased faster across blocks for the predictive trials than for the nonpredictive trials. The authors were interested in whether this unconscious learning was associated with changes in gamma band activity recorded with MEG. □

Space

fMRI BOLD

Example 18.2 fMRI BOLD signal and neural activity To investigate ~~the neural~~ basis of the fMRIBOLD signal, Logothetis et al. (2001) recordedlocal field potential (LFP) and multi-unit activity (MUA) together with fMRI from a region in primary visual cortex across 29 experimental sessions using 10 macaque monkeys. The stimulus involved rotating checkerboard patterns. In examining the relationship between LFP and BOLD, the authors focused on gamma band activity from 40 to 130 Hz. □

Space

recorded local

We now introduce another example, which we will use repeatedly in several parts of this chapter to demonstrate analytical techniques.

Based on (18.5) the statistical model for observations $y_1, \ldots, y_n$ at time points $t_1, \ldots, t_n$ is then

$$Y_i = \mu_{avg} + R_{amp} \cos(2\pi(\omega_1 t_i - \phi)) + \varepsilon_i$$

Corresponding to

to

where, for the core temperature data, $\omega_1 = 1/72$ cycles per 20 min is the frequency corresponding 1 cycle per day (a 24 h period). To simplify fitting, this model may be converted to a linear form, i.e., a form that is linear in the unknown parameters. Using

$$\cos(u - v) = \cos u \cos v + \sin u \sin v \tag{18.6}$$

with $u = 2\pi\omega_1 t_i$ and $v = 2\pi\phi$ we have

$$R_{amp} \cos(2\pi(\omega_1 t_i - \phi)) = A \cos(2\pi\omega_1 t_i) + B \sin(2\pi\omega_1 t_i) \tag{18.7}$$

where $A = R_{amp} \cos(2\pi\phi)$ and $B = R_{amp} \sin(2\pi\phi)$. We may therefore rewrite the statistical model as

$B \sin(2\pi\omega_1 t_i)$

t_i

$$Y_i = \mu_{avg} + A \cos(2\pi\omega_1 t_i) + B \sin(2\pi\omega_1) + \varepsilon_i, \tag{18.8}$$

should be same as in (18.7)

which has the form of a linear regression model, and may be fitted using ordinary linear regression. Specifically, we do the following:

1. Assume the data $(t_1, \ldots, t_n)$ and $(y_1, \ldots, y_n)$ are in respective variables `time` and `temp`.
2. Define

$$\texttt{cosine} = cos(2\pi\texttt{time}/72)$$
$$\texttt{sine} = sin(2\pi\texttt{time}/72).$$

3. Regress `temp` on `cosine` and `sine`.

For future reference we note that the squared amplitude of the cosine function in (18.7) is

$$R_{amp}^2 = A^2 + B^2 \tag{18.9}$$

and the phase is

$$\phi = \frac{1}{2\pi} \arctan(\frac{B}{A}). \tag{18.10}$$

In the core temperature data of Example 18.3 there is a clear, dominant periodicity, which is easily described by a cosine function using linear regression. We may do a bit better if we allow the fitted curve to flatten out a little, compared to the cosine function. This is accomplished by introducing a second frequency, $\omega_2 = 2\omega_1$ to produce the model

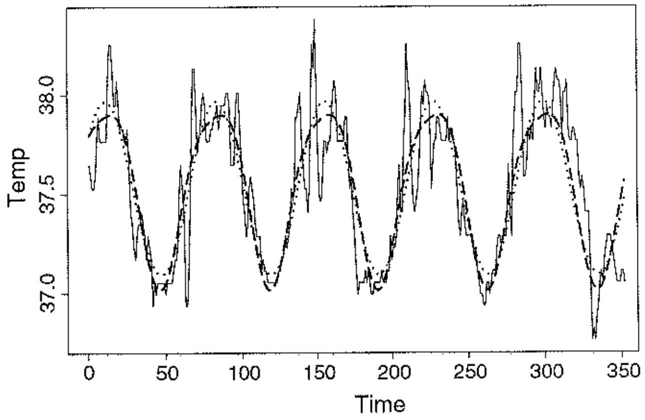


Fig. 18.3 Plot of core temperature, as in Fig. 18.2, together with fit of (18.8), shown in *dotted line*, using the fundamental frequency $\omega_1 = 1/72$ (one oscillation every 72 data points, i.e., every 24 h), and fit of (18.11), shown in *dashed line*. The latter improves the fit somewhat in the peaks and troughs.

$\sin(2\pi\omega_1 t_i)$ t_i

$\sin(2\pi\omega_2 t_i)$ t_i

$$Y_i = \mu_{avg} + A_1 \cos(2\pi\omega_1 t_i) + B_1 \sin(2\pi\omega_1 \underset{\wedge}{)} \\ + A_2 \cos(2\pi\omega_2 t_i) + B_2 \sin(2\pi\omega_2 \underset{\wedge}{)} + \varepsilon_i. \qquad (18.11)$$

Example 18.3 (continued from p. 519) Least-squares regression using model (18.11) yields a highly significant effect for the second cosine–sine pair ($p < 10^{-6}$) and Fig. 18.3 displays a modest improvement in fit. □

Model (18.8) was modified in (18.11) by introducing the additional cosine–sine pair corresponding to the frequency ω_2. In principle this process could be continued by introducing frequencies of the form $\omega_k = k\omega_1$ for $k = 3, 4, \ldots$. Here, ω_1 is called the *fundamental frequency*, the additional frequencies ω_k are *harmonic frequencies*, and the resulting regression model is often called *harmonic regression*. For the core temperature data it turns out that $k = 2$ is a satisfactory choice (see Greenhouse et al. (1987)) but, in general, one might use linear regression to fit many harmonics and ask how much variation in the data is explained by each cosine–sine pair. For this purpose one might use contributions to R^2, which is the germ of the idea behind one of the main topics in time series, *spectral analysis*. Spectral analysis can be a very effective way to describe wave-like behavior, as seen in the EEG signals of Example 2.2.

p. 351.) Harmonic trigonometric functions are orthogonal, so the interpretation is internally consistent.

These steps all involved major conceptual breakthroughs for mathematics.[4] Taken together they suggest that a signal represented by a smoothly varying function $f(t)$ may be decomposed into cosine and sine harmonic components. This is what Fourier analysis accomplishes.

To be a little more specific, suppose that $f(t)$ is a function on the interval $[0, 1]$ and let us consider time points $t_j = \frac{j}{n}$ for $j = 1, 2, \ldots, n$ where, for simplicity, we assume n is odd so that $(n-1)/2$ is an integer. If we evaluate $f(t)$ at the time points t_j we get an n-dimensional vector

$$y = (f(t_1), f(t_2), \ldots, f(t_n))^T. \tag{18.12}$$

Now define the harmonic trigonometric functions $f_k(t) = \cos(2\pi kt)$ and $g_k(t) = \sin(2\pi kt)$, for $k = 1, 2, \ldots, (n-1)/2$. By evaluating these functions at $t_1, t_2, \ldots, t_n$ we form vectors $f_k = (f_k(t_1), f_k(t_2), \ldots, f_k(t_n))^T$ and $g_k = (g_k(t_1), g_k(t_2), \ldots, g_k(t_n))^T$ and, it turns out, the collection of vectors

$$1_{vec}, f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$$

are orthogonal, where $1_{vec} = (1, 1, \ldots, 1)^T$. (This follows from straightforward algebraic manipulation, together with properties of sines and cosines, see Bloomfield (2000)). They therefore form an orthogonal basis for R^n (see Section ~~19.4~~), which means that any vector y, such as in (18.12), may be written in the form

Section A.9

A.9

$$\begin{aligned} y = {} & \mu_{avg} 1_{vec} + A_1 f_1 + \cdots + A_{(n-1)/2} f_{(n-1)/2} \\ & + B_1 g_1 + \cdots + B_{(n-1)/2} g_{(n-1)/2}. \end{aligned} \tag{18.13}$$

If we define

$$\begin{aligned} p_n(t) = {} & \mu_{avg} + A_1 f_1(t) + \cdots + A_{(n-1)/2} f_{(n-1)/2}(t) \\ & + B_1 g_1(t) + \cdots + B_{(n-1)/2} g_{(n-1)/2}(t) \end{aligned} \tag{18.14}$$

[4] The first requires the notion of function, which emerged roughly in the 1700s, especially in the work of Euler (the notation $f(x)$ apparently being introduced in 1735). The second may be considered intuitively obvious, but a detailed rigorous understanding of the situation did not come until the 1800s, particularly in the work of Cauchy (represented by a publication in 1821) and Weierstrass (in 1872). The notion of harmonics was one of the greatest discoveries of antiquity, and is associated with Pythagoras. The third and fourth steps emerged in work by D'Alembert in the mid-1700s, and by Fourier in 1807. Along the way, representations using complex numbers were used by Euler (his famous formula, given below, appeared in 1748), but they were considered quite mysterious until their geometric interpretation was given by Wessel, Argand, and Gauss, the latter in an influential 1832 exposition. A complete understanding of basic Fourier analysis was achieved by the early 1900s with the development of the Lebesgue integral. Recommended general discussions may be found in Courant and Robbins (1996), Lanczos (1966), and Hawkins (2001).

the lag-h correlation after adjusting for the effects of lags 1 through $h-1$, adjusting as in multiple linear regression. It may be computed as the normalized lag-h regression coefficient found from an $AR(h)$ model, normalized by dividing the series by the sample variance $\hat{\gamma}(0)$.

> *A detail:* Suppose X_t is a mean-zero stationary Gaussian series. Then the theoretical PACF is given by $\phi_{11} = Cor(X_t, X_{t+1})$ and for $h \geq 2$,
>
> $$\phi_{hh} = Cor(X_t, X_{t+h}|X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}).$$
>
> More generally, for any mean-zero stationary series let $X_t^{h-1} = \sum_{j=1}^{h-1} \beta_j X_{t-j}$ where the coefficients $\beta_1, \ldots, \beta_{h-1}$ minimize $E((X_t - \sum_{j=1}^{h-1} \alpha_j X_{t-j})^2)$ over the α_js. Then, for $h \geq 2$,
>
> $$\phi_{hh} = Cor(X_t - X_t^{h-1}, X_{t+h} - X_{t+h}^{h-1}).$$ □

Once again, using large-sample theory, horizontal lines may be drawn on the sample PACF to indicate where the coefficients stop being significant. The sample PACF is often used to choose the order of the autoregressive model.

Example 18.3 (continued from p. 525) Let us consider an $AR(p)$ model for the core temperature residuals following the cosine regression reported on p. 519, and then detrending (using BARS, see Section 15.2.6). We take $p = 22$. The fitted coefficients are plotted in Fig. 18.7. Here is an abbreviated table of coefficients:

space

Variable	Coefficient	Std. Err.	t-ratio	p-value
x_{B1}	.906	.057	15.9	$< 10^{-15}$
x_{B2}	−.205	.077	−2.7	.008
x_{B3}	−.147	.078	−1.9	.06
x_{B4}	.005	.078	.1	.95
x_{B5}	−0.154	.078	−1.9	.05
x_{B6}	.115	.078	.9	.35
...				
x_{B21}	−.031	.076	−.4	.69
x_{B22}	.011	.057	−.2	.84

Only the first two lagged variables have large t statistics, so it appears that only the first two lagged variables are likely to be helpful in predicting the response variable. Also shown in Fig. 18.7 is the sample ACF, together with horizontal lines drawn at $\pm 2/\sqrt{n}$. The PACF in Fig. 18.7 has nonzero lag-1 and lag-2 coefficients, but the remaining coefficients are not distinctly different from zero relative to statistical uncertainty. Using an $AR(2)$ fit to the residuals added to the fitted 24 h cycle produces the overall fit to the temperature data shown in Fig. 18.8. □

In general, autoregressive models may be fit by maximum likelihood. We now connect ML estimation with lagged least-squares regression (p. 531), by writing down the

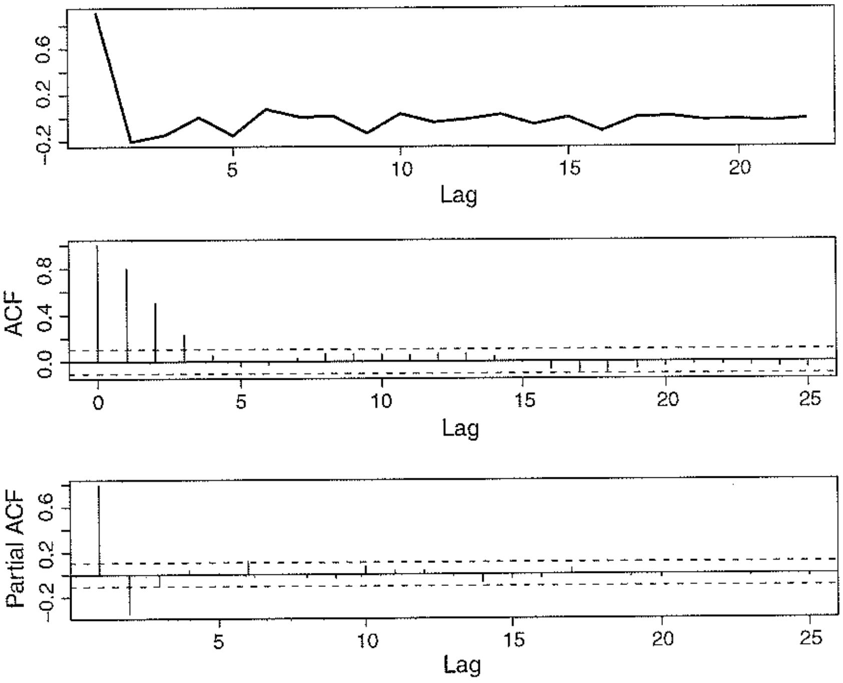


Fig. 18.7 Autoregressive model of order $p = 22$ for core temperature residuals. *TOP* Coefficients $\hat{\phi}_i$ as a function of lag i. *MIDDLE* The sample autocorrelation function. *BOTTOM* The sample partial autocorrelation function.

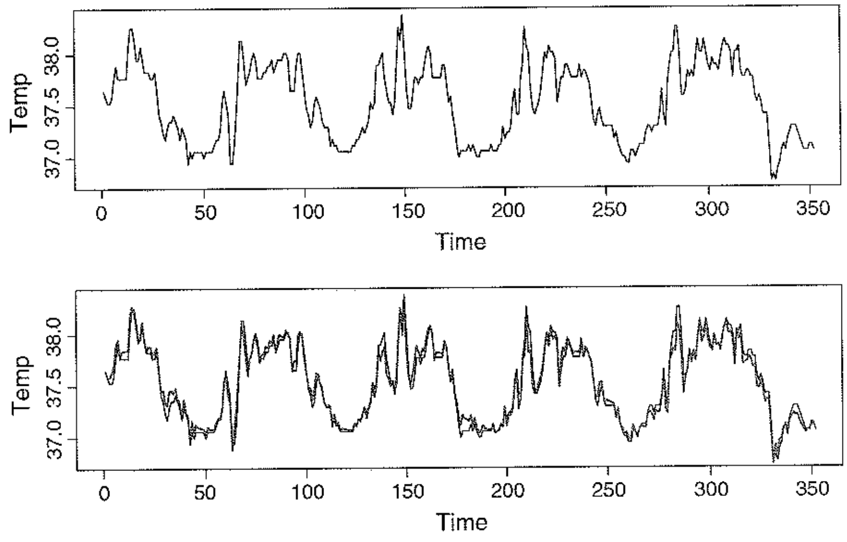


Fig. 18.8 Core temperature data together with fit. *TOP* plot of temperature data. *BOTTOM* Plot of temperature data together with fit based on the sum of an $AR(2)$ fit to residuals and the fitted 24 h cycle.

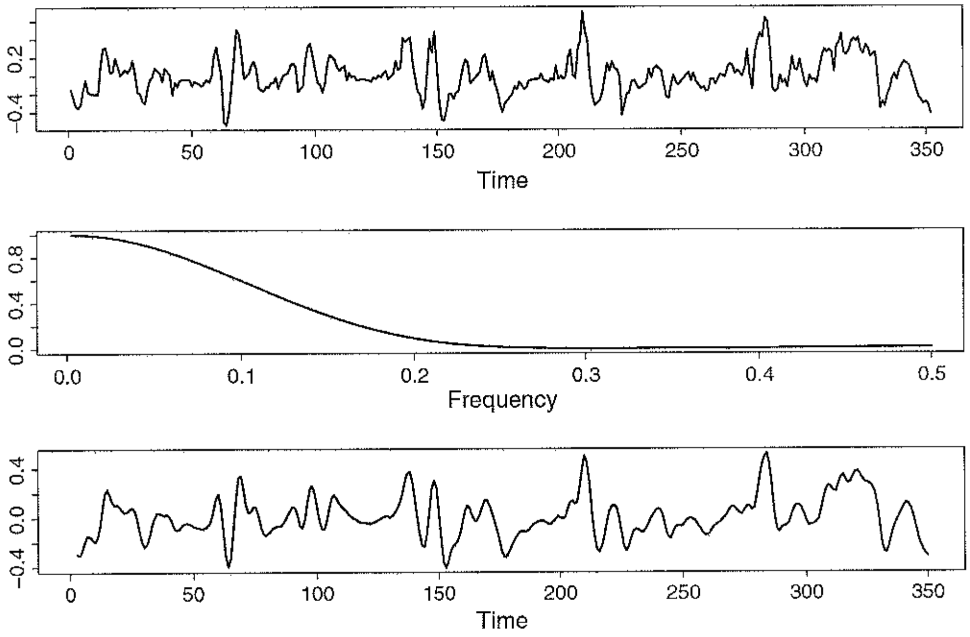


Fig. 18.10 *TOP* Core temperature data after removing dominant 24h effect, i.e., the residuals after simple harmonic regression. *MIDDLE* The power transfer function of the five-point linear filter with coefficients (1, 2, 3, 2, 1)/9, showing a strong diminution of the higher frequency components. *BOTTOM* Core temperature data after applying the five-point linear filter with coefficients (1, 2, 3, 2, 1)/9.

$$y_t = \frac{1}{9}(x_{t-2} + 2x_{t-1} + 3x_t + 2x_{t+1} + x_{t+2}) \tag{18.45}$$

$\sqrt{n}$ replaces n

for $t = 3, \ldots, n-2$. A Gaussian filter would be similar but would instead use a normal (Gaussian) pdf to define the coefficients.

It may be shown that the DFT of $\{y_t\}$ is related to the DFT of $\{x_t\}$ according to

$\sqrt{n}$.

$$d_y(\omega) = \sqrt{n} d_a(\omega) d_x(\omega) \tag{18.46}$$

where $d_a(\omega)$ is the Fourier transform of $\{a_r, a_{r+1}, \ldots, a_s, 0, 0, \ldots, 0\}$, with the zeroes being added to fill up the rest of the n data values. (This is called "padding" the sequence.) The quantity $\sqrt{n} d_a(\omega)$ is called the *transfer function* and its squared magnitude is the *power transfer function*. Expression (18.46) makes it possible to analyze easily the effects of linear filters. This, coupled with their simplicity and the high speed with which they may be computed makes them a very common method of choice for smoothing a time series and the resulting periodogram.

Example 18.3 (continued) We applied the 5-point linear filter described above to the residuals from the core temperature data following simple harmonic regression, yielding a series of the form (18.45). The top panel of Fig. 18.10 shows the residual series and the middle panel shows the power transfer function. The power transfer

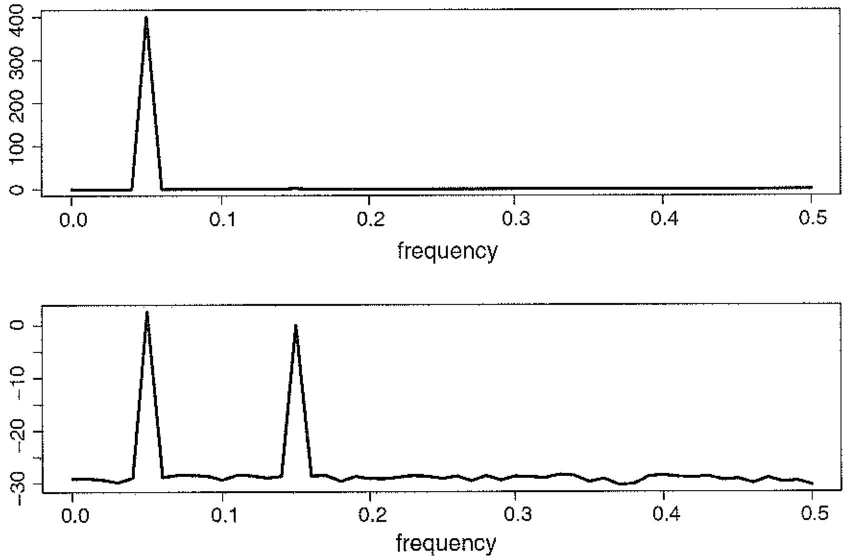


Fig. 18.13 *TOP* Periodogram of $x_t = 20\cos(2\pi\omega_1 t) + \cos(2\pi\omega_2 t)$, where $n = 100$, $\omega_1 =$.~~0~~05 ← .05 and $\omega_2 = .15$. *BOTTOM* Log periodogram of x_t. In the log scale the second peak becomes visible.

periodogram indicates misleadingly that those other frequencies are present in the data.

The problem of leakage is very dramatic when the *dynamic range* of the data is large. Dynamic range refers to the ratio of the largest to smallest positive periodogram values (usually measured on the $\log_{10}$, or decibel, scale).

Illustration: As an illustration, consider

$$x_t = 20\cos(2\pi\omega_1 t) + \cos(2\pi\omega_2 t) \tag{18.50}$$

where $n = 100$, $\omega_1 = .05$ and $\omega_2 = .15$. Its periodogram is shown in the top panel of Fig. 18.13. To see the second frequency it is necessary to use a log scale to plot the periodogram, as shown in the bottom panel of Fig. 18.13. Log periodogram plots are used as defaults in many contexts. Now consider the leakage-prone variant where we take $\omega_1 = 1/22$ rather than $1/20$. Its periodogram is shown in Fig. 18.14. In this case leakage obscures the second peak almost entirely, and if the periodogram were noisy (as it is with real data) it would be extremely difficult to see the second peak at all. □

Leakage is also a problem when there are trends, which cause large low-frequency coefficients in the periodogram.

Example 15.2 (continued from p. 528) We previously showed the log periodogram for the LFP data in Fig. 18.5. The very low frequency trends cause leakage, which obscures the higher frequencies of interest. □

Example 2.2 (continued from p. 514) The spectrograms in Fig. 2.2 on p. 27 displayed nicely some changes in the frequency content of EEGs across the course of the experiment. Specifically, the alpha rythm appeared during an epoch in which the subject's eyes closed, and during induction of anesthesia. □

rhythm

h

Spectrograms, such as that in Example 2.2, may be created by segmenting the observation time interval $[0, T]$ into a set of subintervals $[0, T_1], [T_1, T_2], \ldots, [T_k, T]$, and then computing spectral density estimates within each interval. The estimated spectrum is then plotted on the y-axis for every time interval, with time labeled along the x-axis. The intervals must be chosen to be long enough so that there are substantial series from which to estimate the spectrum, yet short enough that the series may be considered stationary within each interval. Some spectrogram software takes as a default 512 observations per interval (with corrections to this to allow for T not being divisible by 512). Some smoothing (and tapering) of the spectral density estimates across time is often incorporated. One way to smooth across time, which is available as an option in most spectrogram software, is to choose the analysis intervals to be overlapping. In some experiments there are repeated trials, in which case the spectrograms may be averaged across trials.

Example 18.2 (continued from p. 518) To display the LFP response to the stimulus Logothetis et al. (2001) used a spectrogram that incorporated tapering and was averaged across trials and across subjects. It showed strong power in the gamma range after onset of the stimulus. □

Time-frequency analysis is often performed using wavelets (Section 15.2.8). Because of the scaling property (the narrowing range) in the definition (15.9), wavelet regression provides a representation that is localized in both time and frequency, with frequency here defined by the scale of the wavelets. See Percival and Walden (2000).

Example 18.1 (continued from p. 518) In their study of MEG oscillatory activity during learning Chaumon et al. (2009) used Morlet wavelets (see p. 429) to decompose MEG sensor signals across time and frequency. They analyzed the log-transformed power within a 30–48 Hz, band at time 100–400 ms after target onset, from one group of sensors over the occipital lobe and another group of sensors over the frontal lobe. They found that during the learning phase (the first few blocks) of the experiment this gamma band power in the sensors over the occipital lobe was higher for the predictive trials than for the nonpredictive trials ($p < .005$ based on an across-subject paired t-test, using 16 subjects) with the power for the predictive trials being elevated above baseline. On the other hand, during the same learning period, the gamma band power in the sensors over the frontal lobe was depressed for the nonpredictive trials ($p < .0001$), but not for the predictive trials (with the predictive and nonpredictive gamma band power being different, $p < .01$). □

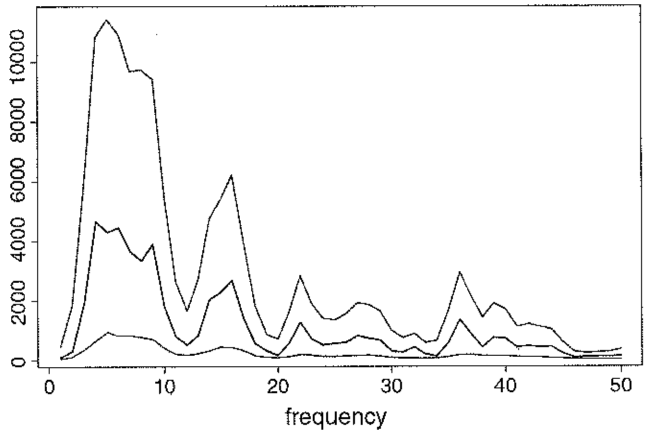


Fig. 18.15 Smoothed periodogram and approximate, pointwise 95 % confidence bands, from the beginning-period LFP detrended series.

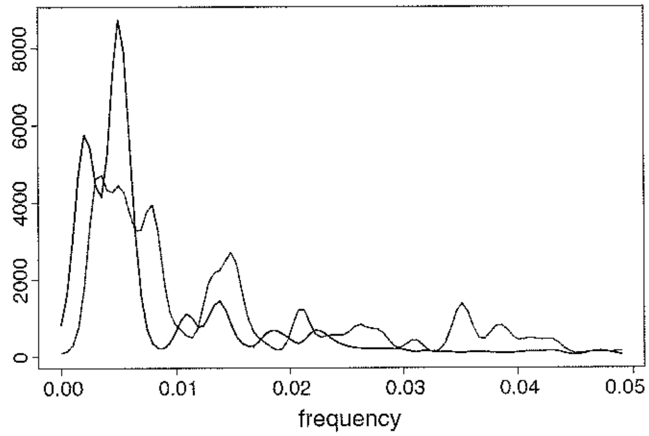


Fig. 18.16 Smoothed periodograms from beginning and end periods, overlaid.

computed. This is usually called a bootstrap, analogously to the bootstrap procedures in Chapter 9.

Example 15.2 (continued from p. 528) Returning to the pair of 1 s average LFP recordings, we noted previously, in Figs. 18.1 and 18.5, the need to detrend the time series before looking for periodicities under the assumption of stationarity. Figure 18.6 displayed the smoothed periodograms of the detrended series. Pointwise 95 % confidence bands together with the smoothed periogram the for the first period, obtained by propagation of uncertainty, are shown in Fig. 18.15.

periodogram

do

We next consider whether the first and last periods have the same spectral density (an indication of stationarity). Figure 18.16 shows the two smoothed periodograms overlaid. A significance test may be based on the integrated squared difference between the two smooth curves. Specifically, if $\hat{f}_1(\omega)$ and $\hat{f}_2(\omega)$ are the two spectral density estimates, then we use

$$t_{obs} = \sum_k (\hat{f}_1(\omega_k) - \hat{f}_2(\omega_k))^2$$

as the test statistic. To compute a p-value under $H_0 = f_1(\omega) = f_2(\omega)$ for all ω, we take as a "pooled" estimate

$$\hat{f}(\omega_k) = \frac{1}{2}(\hat{f}_1(\omega_k) + \hat{f}_2(\omega_k))$$

for $k = 1, \ldots, m$. We then generate a pseudo-sample of pairs of periodograms using $\hat{f}(\omega)$ as the spectral density, and for each generated pair of periograms, apply smoothing and compute t. We then see what fraction of the generated t values is greater than t_{obs}. This is our approximate p-value. In this case, we obtained $p = 0.53$, indicating no evidence that the spectra from the two recording intervals are different. □

periodograms

do

18.4.2 *Uncertainty about functions of time series may be obtained from time series pseudo-data.*

The method above propagates the uncertainty from the asymptotic distribution of the periodogram to anything computed from it. If, however, an analytical technique bypasses the periodogram a different method must be used to propagate uncertainty. A more general idea is to use the approximate normal distributions on the coefficients, in order to propagate the uncertainty from the DFT itself. In other words, one may begin with the uncertainty in the DFT obtained from the data, and then apply an inverse DFT to generate time series that behave the same as the original series in the sense of having (approximately) the same spectrum. The resulting time series pseudo-data are sometimes called *surrogate data.*

An efficient method of carrying out such simulations (based on "circulant embedding") is described in Percival and Constantine (2006). Code by these authors is available in the CRAN library of R packages, within the package `fractal`. See below. As described in the Percival and Constantine paper, the method is closely related to *surrogate time series*, e.g., Schreiber and Schmitz (2000). Additional "bootstrap" resampling methods for spectral analysis, with an emphasis on theoretical results, are discussed in Chapter 9 of Lahiri (2003b). We omit detailed discussion of this topic and note only that the pseudo data generated by this approach are normal (Gaussian), and so do not reflect any sources of uncertainty arising from substantial non-normal variation in the data.

above, are consistent. For example, if we consider intervals (t_1, t_2) and (t_2, t_3) we must be sure that the Poisson distributions for the counts in each of these are consistent with the Poisson distribution for the count in the interval (t_1, t_3). Specifically, in this case, we must know that the sum of two independent Poisson random variables with means $\mu = \lambda(t_2 - t_1)$ and $\mu = \lambda(t_3 - t_2)$ is a Poisson random variable with mean $\mu = \lambda(t_3 - t_1)$. But this follows from the fact that if $W_1 \sim P(\mu_1)$ and $W_2 \sim P(\mu_2)$ independently, and we let $W = W_1 + W_2$, then $W \sim P(\mu_1 + \mu_2)$. We omit the details. □

We now come to an important characterization of homogeneous Poisson processes.

Theorem: A point process is a homogeneous Poisson process with intensity λ if and only if its inter-event waiting times are i.i.d. *Exp*(λ).

Proof: We derive the waiting-time distribution for a homogeneous Poisson process. Recalling that X_i is the length of the inter-event interval between the $(i-1)^{\text{st}}$ and ith event times, we have $X_i > t$ precisely when $\Delta N_{(S_{i-1}, S_{i-1}+t]} = 0$. From the definition of a homogeneous Poisson process, $P\left(\Delta N_{(S_{i-1}, S_{i-1}+t]} = 0\right) = e^{-\lambda t}$. Therefore, the CDF of X_i is $F_{X_i}(t) = P(X_i \le t) = 1 - e^{-\lambda t}$, which is the CDF of an *Exp*(λ) random variable.

The converse of this theorem involves additional calculations and is omitted. □

Recall from Section 5.4.2 that the exponential distribution is memoryless. According to this theorem, for a homogeneous Poisson process, at any given moment the time at which the next event will occur does not depend on past events. Thus, the homogeneous Poisson process "has no memory" of past events.

Another way to think about homogeneous Poisson processes is that the event times are scattered "as irregularly as possible." One characterization of the "irregularity" notion is that, as noted on p. 120, the exponential distribution *Exp*(λ) maximizes the entropy among all distributions on $(0, \infty)$ having mean $\mu = 1/\lambda$. Here is another.

Result: Suppose we observe $N(T) = n$ events from a homogeneous Poisson process on an interval $(0, T]$. Then the distribution of the event times is the same as that of a sample of size n from a uniform distribution on $(0, T]$.

Proof: This appears as a corollary to the theorem on p. 577, where it is also stated more precisely. □

patch

Example 19.4 Miniature excitatory post-synaptic currents Figure 19.2 displays event times of miniature excitatory postsynaptic currents (MEPSCs) recorded from neurons in neonatal mice at multiple days of development. To record these events, the neurons are patch~~ed~~ clamped at the cell body and treated so that they cannot propagate action potentials. These MEPSCs are thought to represent random activations of the dendritic arbors of the neuron at distinct spatial locations, so that the two assumptions of a Poisson process are plausible. The sequence of events in Fig. 19.2 looks highly

$$\hat{\gamma}_{XY}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_t - \bar{x})(y_{t+h} - \bar{y})$$

with $\hat{\gamma}_{XY}(-h) = \hat{\gamma}_{YX}(h)$, and

$$\hat{\rho}(h) = \frac{\hat{\gamma}_{XY}(h)}{\hat{\sigma}_X\hat{\sigma}_Y}.$$

The univariate Eqs. (18.29)–(18.31) have immediate extensions to the bivariate case: if

$$\sum_{h=-\infty}^{\infty} |\gamma_{XY}(h)| < \infty$$

then there is a *cross-spectral density function* $f_{XY}(\omega)$ for which

$$\gamma_{XY}(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i\omega h} f_{XY}(\omega)d\omega \tag{18.51}$$

and

$$f_{XY}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{XY}(h)e^{-2\pi i\omega h}.$$

The cross-spectral density is, in general, complex valued. Because $\gamma_{YX}(h) = \gamma_{XY}(-h)$ we have

$$f_{YX}(\omega) = \overline{f_{XY}(\omega)} \tag{18.52}$$

i.e., $f_{YX}(\omega)$ is the complex conjugate of $f_{XY}(\omega)$. In Section 18.3.1 we said that a smoothed periodogram could be considered an estimator of the theoretical spectral density, and we based that interpretation on a finite-sample expression (18.33), which gave the periodogram as a scaled DFT of the sample covariance function. Similarly, an estimate $\hat{f}_{XY}(\omega)$ of $f_{XY}(\omega)$ may be obtained by smoothing a scaled DFT of the sample cross-covariance function $\hat{\gamma}_{XY}(h)$. In Section 18.5.1 we discuss the important concept of *coherence*, which is defined in terms of the cross-spectral density.

18.5.1 The coherence $\rho_{XY}(\omega)$ between two series X and Y may be considered the correlation of their ω-frequency components.

There is a very nice way to decompose into frequencies the linear dependence between a pair of stationary time series. This frequency-based measure of linear dependence forms an analogy with ordinary correlationwhich, as we noted in

space

Section 4.2.1, may be interpreted as a measure of linear association. To substantiate this interpretation for the ordinary correlation ρ between two random variables Y and X we provided on p. 81 a theorem concerning the linear prediction of Y from $\alpha + \beta X$, giving the formula for α and β that minimized the mean squared error of prediction, $E\left((Y - \alpha - \beta X)^2\right)$ and showing that when these optimal values of α and β are plugg~~g~~ed in, the minimum mean squared error became

plugged

$$E\left((Y - \alpha - \beta X)^2\right) = \sigma_Y^2(1 - \rho^2), \tag{18.53}$$

which was Eq. (4.11).

In Eq. (18.53) we considered the linear prediction of Y based on X, meaning the prediction of Y based on a linear function of X. The analogous problem for $\{(X_t, Y_t), t \in \mathcal{Z}\}$ is to assume

$$Y_t = \sum_{h=-\infty}^{\infty} \beta_h X_{t-h} + W_t, \tag{18.54}$$

where W_t is a stationary process independent of $\{X_t\}$, with $E(W_t) = 0$ and $V(W_t) = \sigma_W^2$, and to minimize the mean squared error

$$MSE = E\left(Y_t - \sum_{h=-\infty}^{\infty} \beta_h X_{t-h}\right)^2. \tag{18.55}$$

Some manipulations show that the solution satisfies

$$\min MSE = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_Y(\omega)(1 - \rho_{XY}(\omega)^2)d\omega \tag{18.56}$$

where

$$\rho_{XY}(\omega)^2 = \frac{|f_{XY}(\omega)|^2}{f_X(\omega)f_Y(\omega)} \tag{18.57}$$

is the *squared coherence*. Thus, in analogy with (18.53), $f_Y(\omega)(1 - \rho_{XY}(\omega)^2)$ is the ω-component of the minimum-MSE fit of (18.54). In (18.56) we have $MSE \geq 0$ and $f_Y(\omega) \geq 0$, which together imply that $0 \leq \rho_{XY}(\omega)^2 \leq 1$ for all ω, and when

$$Y_t = \sum_{h=-\infty}^{\infty} \beta_h X_{t-h}$$

we have $\rho_{XY}(\omega)^2 = 1$ for all ω. These facts, together with (18.56), give the interpretation that the squared coherence is a frequency-based analogue to squared correlation between two theoretical time series.

can lie in a higher-dimensional physical or abstract space. In PET imaging, for example, a radioisotope that has been incorporated into a metabolically active molecule is introduced into the subject's bloodstream and after these molecules become concentrated in specific tissues the radioisotopes decay, emitting positrons which may be detected. These emissions represent a four-dimensional *spatiotemporal* point process because they are localized occurrences both spatially, throughout the tissue, and in time. Here, however, we focus on point processes in time and their application to modeling spike trains.

The simplest point processes are *Poisson processes*, which are *memoryless* in the sense that the probability of an event occuring at a particular time does not depend on the occurrence or timing of past events. In Section 19.2.1 we discuss *homogeneous* Poisson processes, which can describe highly irregular sequences of event times that have no discernable temporal structure. When an experimental stimulus or behavior is introduced, however, time-varying characteristics of the process become important. In Section 19.2.2 we discuss Poisson processes that are *inhomogeneous* across time. In Section 19.3 we describe ways that more general processes can retain some of the elegance of Poisson processes while gaining the ability to describe a rich variety of phenomena.

occurring

r

discernible

i

Spike trains are fundamental to information processing in the brain, and point processes form the statistical foundation for distinguishing signal from noise in spike trains. We have already seen in Chapters 14 and 15 examples of spike train analysis using Poisson regression with spike counts. For this purpose, the Poisson regression model may be conceptualized as involving counts observed over time bins of width Δt based on a neural firing rate $FR(t)$. In Poisson regression, each Poisson distribution has mean equal to $FR(t) \cdot \Delta t$ and then $FR(t)$ is related to the stimulus (or the behavior) by a formula we may write in short-hand as

$$\log FR(t) = \text{stimulus effects}, \tag{19.4}$$

meaning that $\log FR(t)$ is some function that is determined by the stimulus or behavior. In Example 14.5, for instance, the right-hand side of (19.4) involved a quadratic function that represented the effective distance of a rat from the preferred location of a particular hippocampal place cell, and the result was a Poisson regression model of the place cell's activity. This sort of model may be considered a kind of simplified prototype. When we pass to the limit as in (19.2) and use instantaneous firing rate, the Poisson regression model becomes a Poisson process regression model.

Poisson processes are important, and they are especially useful for analyzing the trial-averaged firing rate. When, in Example 15.1, we displayed the smoothed PSTH under two experimental conditions, we were comparing two trial-averaged firing-rate functions. We spell this out in Section 19.3.3. On the other hand, many phenomena can only be studied *within trials*. For instance, oscillatory behavior, bursting, and some kinds of influences of one neuron on another show substantial variation across trials and may be difficult or impossible to detect from across-trial summaries like the PSTH. Careful examination of spike trains within trials usually reveals non-Poisson behavior: neurons tend not to be memoryless, but instead exhibit effects in a variety of settings, probability models for spike trains make dependence on spiking history explicit.

Example 19.1 Retinal ganglion cell under constant conditions Neurons in the retina typically respond to patterns of light displayed over small sections of the visual field. When retinal neurons are grown in culture and held under constant light and environmental conditions, however, they will still spontaneously fire action potentials. In a fully functioning retina, this spontaneous activity is sometimes described as background firing activity, which is modulated as a function of visual stimuli. A short segment of the spiking activity from one neuron appeared in Fig. 16.1. A histogram of the ISIs appears in the left panel of Fig. 19.10. Even though this neuron is not responding to any explicit stimuli, we can still see structure in its firing activity. Although most of the ISIs are shorter than 20 ms, some are much longer: there is a small second mode in the histogram around 60–120 ms. This suggests that the neuron may experience two distinct states, one in which there are bursts of spikes (with short ISIs) and another, more quiescent state (with longer ISIs). From Fig. 16.1 we may also get an impression that there may be bursts of activity, with multiple spikes arriving in quick succession of one another. □

Example 19.2 Beta oscillations in Parkinson's disease Parkinson's disease, a chronic progressive neurological disorder, causes motor deficits leading to difficulty in movement. Clinical studies have shown that providing explicit visual cues, as guides, can improve movement in many patients, a possible explanation being that cortical drive associated with cues may lead to dampening of pathological beta oscillations (10–30 Hz) in the basal ganglia. To investigate this phenomenon, Sarma et al. (2012) recorded from neurons in the basal ganglia (specifically, the substantia nigra) while patients carried out a hand movement task. Because the period associated with a 20 Hz oscillation is 50 ms, if a neuron's activity is related to a beta oscillation it will tend to fire roughly every 50 ms. Therefore, its probability of firing at time *t* will be elevated if it fired previously 50 ms prior to time *t*. This is a form of history effect, which the authors built into their neural models in order to examine whether it was dampened due to visual cues. □

Example 19.3 Spatiotemporal correlations in visual signalling To better understand the role of correlation among retinal ganglion cells, Pillow et al. (2008) examined 27 simultaneously-recorded neurons from an isolated monkey retina during stimulation by binary white noise. The authors used a model having the form of (19.5). They concluded, first, that spike times appear more precise when the spiking behavior of coupled neighboring neurons is taken into account and, second, that in predicting (decoding) the stimulus from the spike trains, inclusion of the coupling term improved prediction by 20 % compared with a method that ignored coupling and instead assumed independence among the neurons. □

signaling

To accommodate event probabilities that change across time, we generalized from homogeneous to inhomogeneous Poisson processes. This eliminated the assumption of stationary increments but it preserved the independence assumption, which entailed history independence. Systems that produce point process data, however, typically have physical mechanisms that lead to history-dependent variation among the events, which cannot be explained with Poisson models. Therefore, it is necessary to further generalize by removing the independence assumption.

The simplest kind of history-dependent behavior occurs when the probability of the ith event depends on the occurence time of the previous event s_{i-1}, but not on any events prior to that. If the ith waiting time X_i is no longer memoryless, then $P(X_i > t + h|X_i > t)$ may not be equal to $P(X_i > u + h|X_i > u)$ when $u \neq t$, but X_i is independent of event times prior to S_{i-1}, and is therefore independent of all waiting times X_j for $j < i$. Thus, the waiting time random variables are all mutually independent. In the time-homogeneous case, they also all have the same distribution. A point process with i.i.d waiting times is called a *renewal process*. We already saw that homogeneous Poisson processes have i.i.d. exponential waiting times. Therefore, renewal processes may be considered generalizations of homogeneous Poisson processes.

occurrence

r

A renewal model is specified by the distribution of the inter-event waiting times. Typically, this takes the form of a probability density function, $f_{X_i}(x_i)$, where x_i can take values in $[0, \infty)$. In principle we can define a renewal process using any probability distribution that takes on positive values, but there are some classes of probability models that are more commonly used either because of their distributional properties, or because of some physical or physiological features of the underlying process.

For example, the gamma distribution, which generalizes the exponential, may be used when one wants to describe interspike interval distributions using two parameters: the gamma shape parameter gives it flexibility to capture a number of characteristics that are often observed in point process data. If this shape parameter is equal to one, then the gamma distribution simplifies to an exponential, which as we have shown, is the ISI distribution of a simple Poisson process. Therefore, renewal models based on the gamma distribution generalize simple Poisson processes, and can be used to address questions about whether data are actually Poisson. If the shape parameter is less than one, then the density drops off faster than an exponential. This can provide a rough description of ISIs when a neuron fires in rapid bursts. If the shape parameter is greater than one, then the gamma density function takes on the value zero at $x_i = 0$, rises to a maximum value at some positive value of x_i, and then falls back to zero. This can describe the ISIs for a relatively regular spike train, such as those from a neuron having oscillatory input. Thus, this very simple class of distributions with only two parameters is capable of capturing, at least roughly, some interesting types of history dependent structure.

While the gamma distribution is simple and flexible, it doesn't have any direct connection with the physiology of neurons. For neural spiking data, a renewal model with a stronger theoretical foundation is the inverse Gaussian. As described in Section 5.4.6, the inverse Gaussian also has two parameters and is motivated by as for a Poisson process, so that homogeneity holds, approximately. As far as independence is concerned, the key point is that the renewal processes are independent of one another, so that the only dependence in the superposition is due to events from the same process, which are very rare among the large numbers of events in the superposition process. That is, if we assume n is so large that, for all k, $P(\Delta N_{(t,t+h]} = 1) >> P(\Delta N^k_{(t,t+h]} = 1)$, then when we consider two non-overlapping intervals $(t_1, t_1 + h]$ and $(t_2, t_2 + h]$, relative to the superposition process, the probability that the kth process has events in both intervals is negligible. This is another way of saying that the identity of events in the superposition gets washed out as the number of processes increases. □

By combining this superposition result and the renewal theorem we obtain a practical implication: the superposition of multiple renewal processes will be approximately a Poisson process, but we can expect the approximation to be better for large t, after initial conditions die out. If, for example, we take multiple spike trains, and if time $t = 0$ has a physiological meaning related to the conditions of the experiment, then we may expect the initial conditions to affect the spike trains in a reproducible way from trial to trial so that even after pooling we might see non-Poisson behavior near the beginning of the trial; as such effects dissipate across time we would expect the pooled spike trains to exhibit Poisson-process-like variation.

19.3.2 The conditional intensity function specifies the joint probability density of spike times for a general point process.

In Section 19.2.2 we described the structure of an inhomogeneous Poisson process in terms of an intensity function that characterized the instantaneous probability of firing a spike at each instant in time, as in (19.6). In an analagous way, a general point process may be characterized by its *conditional intensity function*. Poisson processes are memoryless but, in general, if we want to find the probability of an event in a time interval $(t, t + \Delta t]$ we must consider the timing of the events preceding time t. Let us denote the number of events prior to t by $N(t-)$, analogous

$$N(t-) = max_{u<t}N(u).$$

We call the sequence of event times prior to time t the *history* up to time t and write it as $H_t = (S_1, S_2, \ldots, S_{N(t-)})$. For a set of observed data we would write $H_t = (s_1, s_2, \ldots, s_n)$ with the understanding that $N(t-) = n$. The conditional intensity function is then given by

$$\lambda(t|H_t) = \lim_{\Delta t \to 0} \frac{P(\Delta N_{(t,t+\Delta t]} = 1|H_t)}{\Delta t}, \tag{19.16}$$

The distinction between conditional and marginal intensities is so important for spike train analysis that we emphasize it, as follows.

> If we consider spike trains to be point proces~~s~~ses, within trials the instantaneous firing rate is $\lambda(t|H_t)$ and we have
>
> $$P(\text{spike in } (t, t+dt]|H_t) = \lambda(t|H_t)dt,$$
>
> while the across-trial average firing rate is $\lambda(t)$ and we have
>
> $$P(\text{spike in } (t, t+dt]) = \lambda(t)dt.$$

processes

19.3.4 Conditional intensity functions may be fitted using Poisson regression.

On p. 576 we discussed the way Poisson regresion may be used to fit inhomogeneous Poisson process models. The key theoretical result that made this possible was Eq. (19.11) in conjunction with (19.10). As we said on p. 584, that theorem holds again for conditional intensity functions using Eq. (19.21). This means that Poisson regression can again be used for non-Poisson point processes.

regression

5

We now give some examples in which conditional intensity functions have been fitted to spike train data.

Example 19.1 (continued from p. 569) Let us take time bins to have width $\Delta t = 1$ ms and write $\lambda_k = \lambda(t_k|H_{t_k})$, where t_k is the midpoint of the kth time bin. Defining

$$\log \lambda_k = \alpha_0 + \sum_{j=1}^{120} \alpha_j \Delta N_{(k-j-1,k-j]}, \tag{19.28}$$

we get a model with 120 history-related explanatory variables, each indicating whether or not a spike was fired in a 1 ms interval at a different time lag. The parameter α_0 provides the log background firing rate in the absence of prior spiking activity within the past 121 ms. Using Poisson regression with ML estimation (as in Section 14.1) we obtained $\hat{\alpha}_0 = 3.8$ so that, if there were no spikes in the previous 121 ms, the conditional intensity would become $\lambda_k = \exp(\hat{\alpha}_0) = 45$ spikes per second, corresponding to an average ISI of 22 ms. The MLEs $\hat{\alpha}_i$ obtained from the data are plotted in Fig. 19.4, in the form $\exp\{\hat{\alpha}_i\}$. The $\hat{\alpha}_i$ values related to 0–2 ms after a spike are large negative numbers, so that $\exp\{\hat{\alpha}_i\}$ is close to zero, leading to a refractory period when the neuron is much less likely to fire immediately after

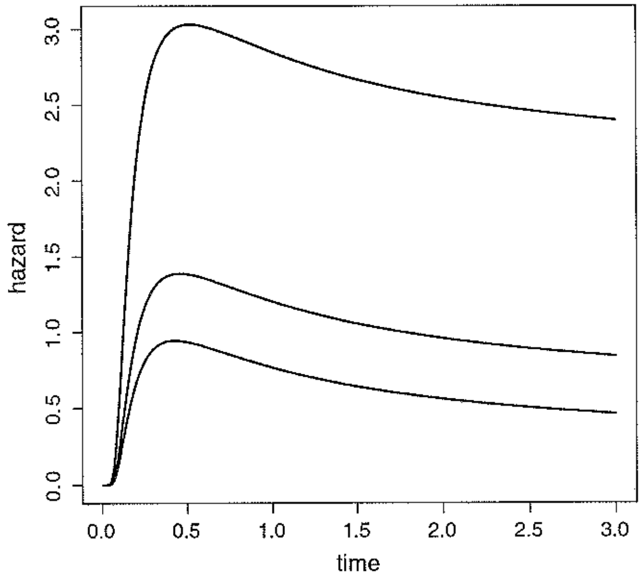


Fig. 19.9 Plots of inverse Gaussian hazard function for three different values of the coefficient of variation, .7 (*top curve*), 1 (*middle curve*), and 1.3 (*bottom curve*). These values correspond to the rough range of those commonly observed in cortical interspike interval data. The theoretical coefficient of variation is given by Eq. (5.16).

The non-monotonic behavior of the recovery function $g_1(t - s_*(t))$ in the foregoing analysis of Example 1.1 may seem somewhat suprising, but anecdotal evidence suggests it may be common. Interestingly, Adrian and Lucas (1912) found a qualitatively similar result by a very different method. They stimulated a frog's sciatic nerve through a second electrode and examined the time course of "excitability," which they defined as the reciprocal of the voltage threshold required to induce an action potential. Figure 19.8 plots this excitability as a function of time since the previous stimulus. There is again a relative refractory period of approximately 10 ms followed by an overshoot and a gradual return to the baseline value. Furthermore, the theoretical inter-spike interval distribution for an integrate-and-fire neuron (following a random walk generated by excitatory and inhibitory post-synaptic potentials) is inverse Gaussian (see Section 5.4.6), and the hazard function for an inverse Gaussian has a non-monotonic shape, shown in Fig. 19.9, that closely resembles the typical recovery function. The qualitative shape of the recovery function shown in Fig. 19.7 is thus consistent with what we would expect from the point of view of theoretical neurobiology.

surprising

r

In many experimental settings spike trains are collected to see how they differ under varying experimental conditions. The conditions may be summarized by a variable or vector, often called a *covariate* (as in regression, see p. 332). Furthermore, there may be other variables that may be related to spiking activity, which could be time-varying, such as a local field potential. Let us collect any such covariates into a vector denoted by u_t if we regard them as fixed by the experimenter, and V_t if they should be considered stochastic. We then write $X_t = (H_t, u_t, V_t)$ and let the conditional intensity become a function not only of time and history, but also of the covariate vector X_t. Thus, for an observation $X_t = x_t$ we write the conditional intensity in the form $\lambda(t|x_t)$. With this in hand we may generalize the statement on p. 586, allowing it to cover the interesting cases implied by our discussion surrounding Eq. (19.5), as follows:

If we consider spike trains to be point proces~~s~~ses, within trials the instantaneous firing rate is $\lambda(t|x_t)$ and we have

processes

$$P(\text{spike in } (t, t+dt]|H_t) = \lambda(t|x_t)dt. \tag{19.32}$$

We may also generalize formula (19.20).

Theorem If the conditional intensity of an orderly point process on an interval $(0, T]$ depends on the random process X_t, so that when $X_t = x_t$ it may be written in the form $\lambda(t|x_t)$, then, conditionally on $X_t = x_t$, the event time sequence $S_1, S_2, \ldots, S_{N(T)}$ has joint pdf

$$f_{S_1,\ldots,S_{N(T)}|X_t}(s_1, \ldots, s_n|X_t = x_t) = \exp\left\{-\int_0^T \lambda(t|x_t)dt\right\} \prod_{i=1}^{n} \lambda(s_i|x_t). \tag{19.33}$$

Proof: The proof is the same as that given for (19.20) in Section 19.4 with x_t replacing H_t. □

A detail: If we are interested in the variation of the conditional intensity with the random vector X_t we can emphasize this by writing it in the form $\lambda(t|X_t)$. For example, in a multi-trial experiment, the firing rate may vary across trials, and the conditional intensity could include a component that changes across trials (see Ventura et al. (2005b)). In such situations, the model includes two distinct sources of variability: one due to variability described by the point process pdf in (19.33) and the second due to the way the conditional intensity varies with X_t. The resulting point process is often called *doubly stochastic.* □

Example 16.6 (continued from p. 472) We now give some additional details about the model used by Frank et al (2002). They applied a multiplicative IMI model to characterize spatial receptive fields of neurons from both the CA1 region of the hippocampus and the deep layers of the entorhinal cortex (EC) in awake, behaving rats. In their model, each neuronal spike train was described in terms of a conditional intensity function of the form (19.31), where the temporal factor $g_0(t)$ became

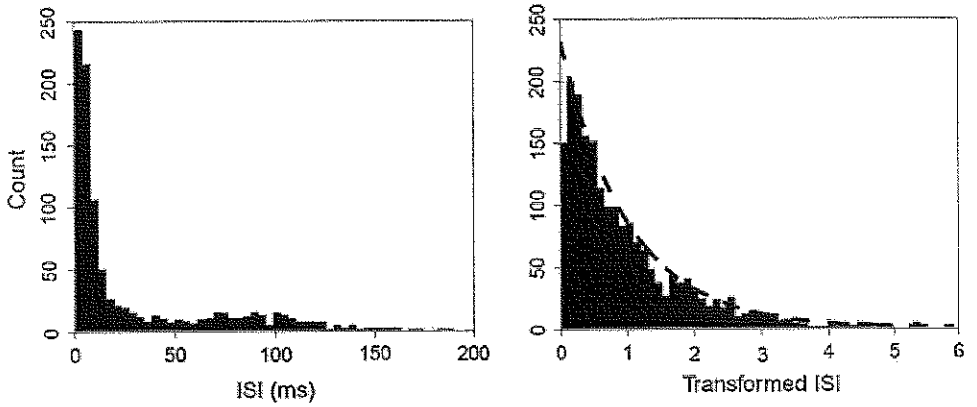


Fig. 19.10 *Left* Histogram of ISIs for the retinal ganglion cell spike train. *Right* Histogram of time-rescaled ISIs. *Dashed line* is the *Exp*(1) pdf.

the null hypothesis that the transformed waiting times follow an *Exp*(1) distribution, which becomes an assessment of fit of the conditional intensity function. If the P–P plot consists of pairs (x_r, y_r), for $r = 1, \ldots, n$, the usual approach is to use the points $(x_r, y_r + 1.36/\sqrt{n})$ and $(x_r, y_r - 1.36/\sqrt{n})$ to define upper and lower bands for visual indication of fit, as illustrated in Fig. 19.11. Specifically, to make a P–P plot for a conditional intensity function $\lambda(t|x_t)$ used to model spike times $s_1, s_2, \ldots, s_n$ we do the following:

1. From (19.36) and (19.37) find transformed spike times $z_1, \ldots, z_n$;
2. for $j = 1, \ldots, n$ define $u_j = 1 - \exp(-z_j)$;
3. put the values $u_1, \ldots, u_n$ in ascending order to get $u_{(1)}, \ldots, u_{(n)}$;
4. for $r = 1, \ldots, n$ (see p. 67) plot the (x, y) pair $\left(\frac{r-.5}{n}, u_{(r)}\right)$;
5. produce upper and lower bands: for $r = 1, \ldots, n$ plot the (x, y) pair $\left(\frac{r-.5}{n}, u_{(r)} + 1.36/\sqrt{n}\right)$ and $\left(\frac{r-.5}{n}, u_{(r)} - 1.36/\sqrt{n}\right)$.

586 6 **Example 19.1 (continued from p. 58~~4~~9)** Using the conditional intensity of Eq. (19.28) we may apply time rescaling. Figure 19.10 displays a histogram of the original ISIs for this data. The smallest bin (0–2 ms) is empty due to the refractory period of the neuron. We can also observe two distinct peaks at around 10 and 100 ms respectively. It is clear that this pattern of ISIs is not described well by an exponential distribution, and therefore the original process cannot be accurately modeled as a simple Poisson process. However the histogram in the right panel of the figure shows the result of transforming the observed ISIs according to the conditional intensity model. Figure 19.11 displays a P–P plot for the intervals in the right panel of Fig. 19.10. Together, these figures show that the model in Eq. (19.28) does a good job of describing the variability in the retinal neuron spike train. □

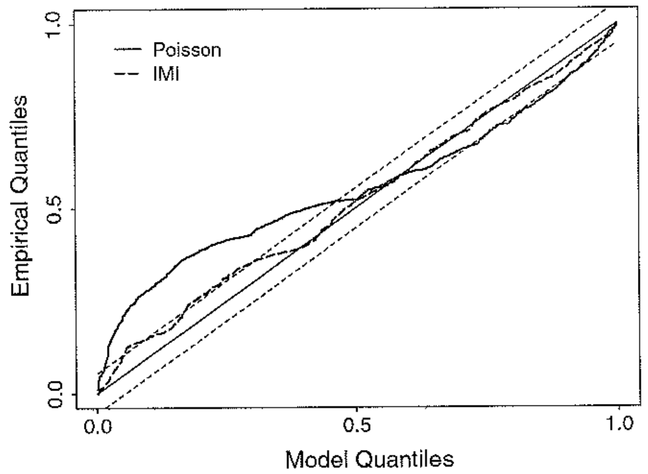


Fig. 19.12 P–P plots of inhomogeneous Poisson and multiplicative IMI models for spike train data from a locust olfactory bulb. For a perfect fit the curve would fall on the diagonal line $y = x$. The data-based (empirical) probabilities deviate substantially from the Poisson model but much less so from the IMI model. When the curve ranges outside the diagonal bands above and below the $y = x$ line, some lack of fit is indicated according to the Kolmogorov-Smirnov test (discussed in Section 10.3.7).

Poisson process with rate λ, we can draw a random sample from an $Exp(\lambda)$ distribution and take the ith event time to be $s_i = \sum_{j=1}^{i} x_j$.

Generating event times from a general point process is more complicated. One simple approach, based on the Bernoulli approximation, involves partitioning the total time interval into small bins of size Δt: in the kth interval, centered at t_k, we generate an event with probability $p_k = \lambda(t_k|x_{t_k})\Delta t$, where x_{t_k} depends on the history of previously generated events. This works well for small simulation intervals. However, as the total time interval becomes large and as Δt becomes small, the number of Bernoulli samples that needs to be generated becomes very large, and most of those samples will be zero, since $\lambda(t|x_t)\Delta t$ is small. In such cases the method becomes very inefficient and thus may take excessive computing time. Alternative approaches generate a relatively small number of i.i.d. observations, and then manipulate them so that the resulting distributions match those of the desired point process.

Thinning To apply this algorithm, the conditional intensity function $\lambda(t|x_t)$ must be bounded by some constant, $\lambda_{\max}$. The algorithm follows a two-stage process. In the first stage, a set of candidate event times is generated as a simple Poisson process with a rate $\lambda_{\max}$. Because $\lambda_{\max} \geq \lambda(t|x_t)$, these candidate event times occur more frequently than they would for the point process we want to simulate. In the second stage they are "thinned" by removing some of them according to a stochastic scheme. We omit the details. In practice, thinning is typically only used when simulating inhomogenous Poisson processes with bounded intensity functions.

inhomogeneous

e

this spectrum might presume that this spiking process has no very low frequency firing, tends to fire around 120 Hz, but also has considerable high frequency activity, suggesting no refractoriness. However, this interpretation is incorrect. The point process generating this spike train actually has an average firing rate around 28 Hz and reflects realistic spiking features including a 5 ms refractory period and an increased probability of firing 8 ms after a previous spike. The error here does not come from the computation of the estimated spectrum, but rather from the näive interpretation.

We do not pursue further the estimation of point process spectra. Our discussion of Fig. 19.13 is intended to show that point process spectra must be interpreted carefully.

discussion

19.4 Additional Derivations

Derivation of Equation (19.9) We start with a lemma.

Lemma The pdf of the ith waiting-time distribution is

$$f_{S_i}(s_i|S_{i-1}=s_{i-1}) = \lambda(s_i)\exp\left\{-\int_{s_{i-1}}^{s_i}\lambda(t)dt\right\}. \tag{19.40}$$

Proof of the lemma: Note that $\{S_i > s_i|S_{i-1}=s_{i-1}\}$, is equivalent to there being no events in the interval $(s_{i-1}, s_i]$. Therefore, from the definition of a Poisson process on p. 574 together with the Poisson pdf in Eq. (5.3), we have $P(S_i > s_i|S_{i-1}=s_{i-1})$ $= P\left(\Delta N_{(s_{i-1},s_i]}=0\right) = \exp\left\{-\int_{s_{i-1}}^{s_i}\lambda(t)dt\right\}$, and the ith waiting time CDF is therefore $P(S_i \leq s_i|S_{i-1}=s_{i-1}) = 1-\exp\left\{-\int_{s_{i-1}}^{s_i}\lambda(t)dt\right\}$. The derivative of the CDF

$$f_{S_i}(s_i|S_{i-1}=s_{i-1}) = \frac{d}{ds_i}\left(1-\exp\left\{-\int_{s_{i-1}}^{s_i}\lambda(t)dt\right\}\right)$$

gives the desired pdf. □

Proof of the theorem: We have

$$\begin{aligned} &f_{S_1,\ldots,S_{N(T)}}(s_1,\ldots,s_n) \\ &= f_{S_1}(s_1)f_{S_2}(s_2|S_1=s_2)\cdots f_{S_{N(T)}}(s_n|S_{n-1}=s_{n-1})\cdot P(\Delta N_{(s_n,T]}=0). \end{aligned}$$

The factors involving waiting-time densities are given by the lemma. The last factor is

$$P(\Delta N_{(s_n,T]}=0) = \exp\left(-\int_{s_n}^{T}\lambda(t)dt\right).$$

Combining these gives the result. □

Derivation of Equation (19.20) We need a lemma, which is analogous to the lemma used in deriving (19.9).

Lemma For an orderly point process with conditional intensity $\lambda(t|H_t)$ on $[0, T]$, the pdf of the ith waiting-time distribution, conditionally on $S_1 = s_1, \ldots, S_{i-1} = s_{i-1}$, for $t \in (s_{i-1}, T]$ is

$$f_{S_i|S_1,\ldots,S_{i-1}}(s_i|S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = \lambda(s_i|H_t)\exp\left\{-\int_{s_{i-1}}^{s_i} \lambda(t|H_t)dt\right\}. \tag{19.41}$$

Proof of the lemma: Let X_i be the waiting time for the ith event, conditionally on $S_1 = s_1, \ldots, S_{i-1} = s_{i-1}$. For $t > s_{i-1}$ we have $X_i \in (t, t + \Delta t)$ if and only if $\Delta N_{(t,t+\Delta t)} > 0$. Furthermore, if the ith event has not yet occurred at time t we have $H_t = (s_1, \ldots, s_{i-1})$. We then have

$$\lim_{\Delta t \to 0} \frac{P(X_i \in (t, t + \Delta t)|X_i > t, S_1 = s_1, \ldots, S_{i-1} = s_{i-1})}{\Delta t}$$
$$= \lim_{\Delta t \to 0} \frac{P(\Delta N_{(t,t+\Delta t)} > 0|H_t))}{\Delta t}$$

and, because the point process is regular, the right-hand side is $\lambda(t|H_t)$. Just as we argued in the case of hazard functions, in Section 3.2.4, the numerator of the left-hand side may be written

$$P(X_i \in (t, t + \Delta t)|X_i > t, H_t) = \frac{F(t + \Delta t|H_t) - F(t|H_t)}{1 - F(t|H_t)}$$

subscript t

where F is the CDF of the waiting time distribution, conditionally on H_t. Passing to the limit again gives

$$\lim_{\Delta t \to 0} \frac{P(X_i \in (t, t + \Delta t)|X_i > t, H_t)}{\Delta t} = \frac{f(t|H_t)}{1 - F(t|Ht)}.$$

H_t

In other words, just as in the case of a hazard function, the conditional intensity function satisfies

$$\lambda(t|H_t) = \frac{f(t|H_t)}{1 - F(t|Ht)}.$$

H_t

Proceeding as in the case of the hazard function we then get the conditional pdf

$$f(t|H_t) = \lambda(t|H_t)e^{-\int_{s_{i-1}}^{t} \lambda(u|x_u)du}$$

as required. □

Proof of the theorem: The argument follows from the lemma by the same steps as the theorem for inhomogeneous Poisson processes. □

if and only if it is of full rank). Thus, a positive semi-definite matrix is non-singular if and only if all its eigenvalues are positive.

The spectral decomposition has a very nice geometrical interpretation. First, the set of two-dimensional points (u_1, u_2) satisfying

E_{11} E_{22}

$$\frac{u_1^2}{\cancel{D}_{11}} + \frac{u_2^2}{\cancel{D}_{22}} = c^2 \tag{A.21}$$

$2c\sqrt{E_{11}'}$

E E

E E

where $\cancel{D}_{11}$ and $\cancel{D}_{22}$ are positive numbers, forms an ellipse centered at the origin. Furthermore, the ellipse is oriented so that its two axes ~~fall~~ along the u_1 and u_2 coordinate axes, and the lengths of its two axes are $2c\sqrt{\cancel{D}_{11}}$ and $2c\sqrt{\cancel{D}_{22}}$. If we let $u = (u_1, u_2)$ then Eq. (A.21) may be written

E E

$D_{22} = E_{22}^{-1}$

$$u^T Du = c^2 \tag{A.22}$$

$D_{11} = E_{11}^{-1}$

where D is the diagonal matrix with diagonal elements D_{11} and D_{22}. Now let R_θ be the 2×2 orthogonal matrix that rotates each vector counter-clockwise through an angle θ. As pointed out above, R_θ^T is the 2×2 orthogonal matrix that rotates each vector clockwise through an angle θ. If we define $x = R_\theta u$ then $u = R_\theta^T x$, and from (A.22) we have

$$x^T R_\theta D R_\theta^T x = c^2 \tag{A.23}$$

E E

so that (A.23) must be the equation of an ellipse whose axes fall ~~along~~ the axes defined by the vectors $\text{col}_1(R_\theta)$ and $\text{col}_2(R_\theta)$ and have lengths $2c\sqrt{\cancel{D}_{11}}$ and $2c\sqrt{\cancel{D}_{22}}$. Because every orthogonal matrix is a rotation followed by a possible re-orientation of the axes, and such a re-orientation of axes defining x would not change the location of the ellipse defined by (A.23), for any 2×2 orthogonal matrix P, the equation

$$x^T PDP^T x = c^2, \tag{A.24}$$

E E

is the equation of an ellipse whose axes fall along the axes defined by the vectors $\text{col}_1(P)$ and $\text{col}_2(P)$ and have lengths $2c\sqrt{\cancel{D}_{11}}$ and $2c\sqrt{\cancel{D}_{22}}$. An analogous interpretation of Eq. (A.24) holds when x is k-dimensional and P and D are $k \times k$ matrices. Thus, for a positive definite matrix A, the equation $x^T Ax = 1$ defines an ellipse, and the spectral decomposition of A shows that the axes of this ellipse are oriented along the eigenvectors of A and have lengths equal to twice the square-root of the reciprocal of the corresponding eigenvalues.

A.9 Vector Spaces

The n-dimensional vectors $e_1 = (1, 0, 0, \ldots, 0)$, $e_2 = (0, 1, 0, 0, \ldots, 0)$, ..., $e_n = (0, \ldots, 0, 1)$ play a special role because they specify the axes or coordinate directions corresponding to each component of an n-dimensional vector $x = (x_1, x_2, \ldots, x_n)$

References

Abeles, M. (2009). “Synfire chains.” *Scholarpedia*, 4, 1441.

Adolph, K. (2002). “Babies steps make giant strides toward a science of development.” *Infant Behavior and Development*, 25, 86–90.

Adrian, E. and Lucas, K. (1912). “On the summation of propagated disturbances in nerve and muscle.” *The Journal of Physiology*, 44, 68–124.

Adrian, E. and Zotterman, Y. (1926). “The impulses produced by sensory nerve endings: Part II: The response of a single end organ.” *Journal of Physiology*, 61, 151–171.

Agresti, A. (1990). *Categorical data analysis*. Wiley.

Agresti, A. and Caffo, B. (2000). “Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures.” *The American Statistician*, 54, 280–288.

Akaike, H. (1974). “A new look at the statistical model identification.” *Automatic Control, IEEE Transactions on*, 19, 716–723.

Amirikian, B. and Georgopoulos, A. (2000). “Directional tuning profiles of motor cortical cells.” *Neuroscience Research*, 36, 73–79.

Anderson, C. and Stevens, C. (1973). “Voltage clamp analysis of acetylcholine produced end-plate current fluctuations at frog neuromuscular junction.” *Journal of Physiology*, 235, 655–691.

Anderson, J. (1990). *Cognitive Psychology and its Implications*. MacMillan.

Anderson, J. and Schooler, L. (1991). “Reflections of the environment in memory.” *Psychological Science*, 2, 396–408.

Anscombe, F. (1973). “Graphs in statistical analysis.” *The American Statistician*, 27, 1, 17–21.

Arlot, S. and Celisse, A. (2010). “A survey of cross-validation procedures for model selection.” *Statistics Surveys*, 4, 40–79.

Bar-Gad, I., Ritov, Y., Vaadia, E., and Bergman, H. (2001). “Failure in identification of overlapping spikes from multiple neuron activity causes artificial correlations.” *Journal of Neuroscience Methods*, 107, 1–13.

Bates, D. and Watts, D. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Behrmann, M., Ghiselli-Crippa, T., Sweeney, J., DiMatteo, I., and Kass, R. (2002). “Mechanisms underlying spatial representation revealed through studies of hemispatial neglect.” *Journal of Cognitive Neuroscience*, 14, 272–290.

Olson, C.R.

R E

Kass, R.E.

Behseta, S., Berdyyeva, T., Olson, C., and Kass, R. (2009). “Bayesian correction for attenuation of correlation in multi-trial spike count data.” *Journal of Neurophysiology*, 101, 2186–2193.

Behseta, S. and Kass, R. E. (2005). “Testing equality of two functions using BARS.” *Statistics in Medicine*, 24, 3523–3534.

functions

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

R, Behseta, S., Kass, R. E., Moorman, D., and Olson, C. (2007). "Testing equality of several functions: Analysis of single-unit firing rate curves across multiple experimental conditions." *Statistics in Medicine*, 26, 21, 3958–3975.

Behseta, S., Kass, R., and Wallstrom, G. (2005). "Hierarchical models for assessing variability among functions." *Biometrika*, 92, 419–434.

Bengio, Y. and Granvalet, Y. (2004). "No unbiased estimator of the variance of K-fold cross-validation." *Journal of Machine Learning Research*, 5, 1089–1105.

Benjamini, Y. and Yekutieli, D. (2001). "The control of the false discovery rate in multiple testing under dependency." *Annals of Statistics*, 29, 1165–1188.

Bernoullli, J. (1713). *Ars conjectandi*. Basel; Thurnisiorum.

Bickel, P. and Doksum, K. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. 1. Prentice Hall.

Billingsley, P. (1995). *Probability and Measure*. 3rd ed. New York: Wiley.

Bliss, C. (1936). "The size factor in the action of arsenic upon silkworm larvae." *Journal of Experimental Biology*, 13, 95–110.

Bloomfield, P. (2000). *Fourier Analysis of Time Series*. Wiley.

Bollen, K. (2002). "Latent variables in psychology and the social sciences." *Annual Review of Psychology*, 53, 605–634.

Box, G., Jenkins, G., and Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control*. 4th ed. Wiley.

Box, G. E. P. (1979). "Robustness in the strategy of scientific model building." In *Robustness in Statistics*, eds. R. Launer and G. Wilkinson. New York: Academic Press.

Brillinger, D. (1972). "The spectral analysis of stationary interval functions." *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 483–513.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. Expanded ed. Holden Day.

Brillinger, D. (2002). "John W. Tukey: His life and professional contributions." *The Annals of Statistics*, 30, 1535–1575.

Brion, M., Lawlor, D., Matijasevich, A., Horta, B., Anselmi, L., Araújo, C., Menezes, A., Victora, C., and Smith, G. (2011). "What are the causal effects of breastfeeding on IQ, obesity and blood pressure? Evidence from comparing high-income with middle-income cohorts." *International Journal of Epidemiology*, 40, 670–680.

Brockwell, A., Kass, R., and Schwartz, A. (2007). "Statistical signal processing and the motor cortex." *Proceedings of the IEEE*, 95, 881–898.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S. (2004). "Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality." *Proceedings of the National Academy of Sciences*, 101, 9849–9854.

Brown, E. N., Barbieri, R., Eden, U., and Frank, L. (2003). "Likelihood methods for neural data analysis." In *Computational Neuroscience: A comprehensive approach*, ed. J. Feng, chap. 9, 253–286. London: CRC.

Quotes missing

" Brown, E. N., Frank, L. M., Tang, D., Quirk, M. C., and Wilson, M. A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18: 7411–7425. "

Brown, E. N. and Kass, R. E. (2009). "What is statistics? (with discussion)." *American Statistician*, 63, 105–123.

Brown, E. N., Kass, R. E., and Mitra, P. (2004). "Multiple neural spike trains analysis: state-of-the-art and future challenges." *Nature Neuroscience*, 7, 456–461.

Brownlee, K. (1965). *Statistical Theory and Methodology in Science and Engineering*. Wiley.

Bundesen, C. (1998). "A computational theory of visual attention." *Philosophical Transactions of the Royal Society B, London*, 353, 1271–1281.

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E., and Munafó, M. (2013). "Power failure: Why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience*, 14, 365–376.

Casey, B., Somerville, L., Gotlib, I., Ayduk, O., Franklin, N., Askren, M., Jonides, J., Berman, M., Wilson, N., Teslovich, T., Glover, G., Zayas, V., Mischel, W., and Shoda, Y. (2011). "Behavioral and neural correlates of delay of gratification 40 years later." *Proceedings of the National Academy of Sciences*, 108, 14998–15003.

Chaumon, M., Schwartz, D., and Tallon-Baudry, C. (2009). "Unconscious learning versus visual perception: Dissociable roles for gamma oscillations revealed in MEG." *Journal of Cognitive Neuroscience*, 21, 2287–2299.

Chen, C. (1985). "On asymptotic normality of limiting density functions with Bayesian implications." *Journal of the Royal Statistical Society. Series B*, 47, 540–546.

Churchland, A.K. Kiani, R., Chaudhuri, R., Wang, X., Pouget, A., and Shadlen, M. (2011). "Variance as a signature of neural computations during decision-making." *Neuron*, 69, 818–831.

Cleveland, W., Diaconis, P., and McGill, R. (1982). "Variables on scatterplots look more highly correlated when the scales are increased." *Science*, 216, 1138–1141.

Colquhoun, D. (2007). "Classical perspective: What have we learned from single ion channels?" *The Journal of Physiology*, 581, 425–427.

Colquhoun, D. and Sakmann, B. (1985). "Fast events in single-channel currents activated by acetylcholine and its analogues at the frog muscle end-plate." *Journal of Physiology*, 369, 501–557.

Courant, R. and Robbins, H. (1996). *What is Mathematics?*. 2nd ed. revised by Ian Stewart: Oxford.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. New York: Wiley.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Dantzig, T. (1954). *Number: The Language of Science*. 4th ed. Doubleday.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. 2nd ed. Cambridge University Press.

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press.

del Castillo, J. and Katz, B. (1954). "Quantal components of the end-plate potential." *The Journal of Physiology*, 124, 560–573.

italics Devlin, B., Fienberg, S., Resnick, D., and Roeder, K., eds. (1997). Intelligence, Genes, & Success: Scientists Respond to The Bell Curve. New York: Copernicus (Springer-Verlag).

DiCiccio, T. and Efron, B. (1996). "Bootstrap confidence intervals." *Statistical Science*, 11, 189–228.

E. DiMatteo, I., Genovese, C., and Kass, R. (2001). "Bayesian curve-fitting with free-knot splines." *Biometrika*, 88, 1051–1077.

Dinstein, I. (2008). "Human cortex: Reflections of mirror neurons." *Current Biology*, 18, R956–R959.

Dinstein, I., Thomas, C., Humphreys, K., Minshew, N., Behrmann, M., and Heeger, D. (2010). "Normal movement selectivity in autism." *Neuron*, 66, 461–469.

Edwards, W., Lindman, H., and Savage, L. (1963). "Bayesian statistical inference for psychological research." *Psychological Review*, 70, 193–242.

Efron, B. (1979a). "Bootstrap methods: Another look at the jackknife." *The Annals of Statistics*, 7, 1–26.

Efron, B. (1979b). "Computers and the theory of statistics: Thinking the unthinkable." *SIAM Review*, 21, 460–480.

Efron, B. (2004). "The estimation of prediction error: Covariance penalties and cross-validation (with discussion)." *Journal of the American Statistical Association*, 99, 619–642.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Ernst, M. and Banks, M. (2002). "Humans integrate visual and haptic information in a statsitically statistically optimal fashion." *Nature*, 415, 429–433.

Faber, D. and Korn, H. (1991). "Applicability of the coefficient of variation method for analyzing synaptic plasticity." *Biophysical Journal*, 60, 1288–1294.

Fan, J. and Kreutzberger, E. (1998). "Automatic local smoothing for spectral density estimation." *Scandanavian Journal of Statistics*, 25, 359–369.

Scandinavian

i

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. 3rd ed. New York: Wiley.

Feynman, R. P., Leighton, R. B., and Sands, M. (1963). "Lectures on Physics, Vol. I." *Addison-Wesley*, 49–1.

Fienberg, S. E. (2006). "When did Bayesian inference become "Bayesian?"." *Bayesian Analysis*, 1, 1–40.

lower case

Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society, A*, 222, 309–368.

Fisher, R. A. (1924). "On a distribution yielding the error functions of several well known statistics." In *Proceedings of the international congress of mathematics*, vol. 2, 805–813. Toronto.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Hafner Press.

Fisher, R. (1935). "Statistical tests'.' *Nature*, 136, 474.

Formisano, E., Di Salle, F., and Goebel, R. (2006). "Fundamentals of data analysis methods in fMRI." In *Advanced imaging processing in magnetic resonance imaging*, eds. L. Landini, V. Positano, and M. Santarelli, 481–503. Boca Raton (FL): CRC Taylor & Francis.

Fox, M., Snyder, A., Vincent, J., Corbetta, M., Van Essen, D., and Raichle, M. (2005). "The human brain is intrinsically organized into dynamic, anticorrelated functional networks." *Proc. National Acad. Sciences*, 102, 9673–9678.

Francq, C. and Zakoiän, J.-M. (2005). "A central limit theorem for mixing triangular arrays of variables whose dependence is allowed to grow with the sample size." *Econometric Theory*, 21, 1165–1171.

Frank, L. M., Eden, U. T., Solo, V., Wilson, M. A., and Brown, E. N. (2002). "Contrasting patterns of receptive field plasticity in the hippocampus and the entorhinal cortex: An adaptive filtering approach." *The Journal of Neuroscience*, 22, 3817–3830.

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics*. 4th ed. New York: W.W. Norton.

Frezza, M., di Padova, C., Pozzato, G., Terpin, M., Baraona, E., and Lieber, C. S. (1990). "High blood alcohol levels in women." *New England Journal of Medicine*, 322, 95–99.

Gagliardo, A., Ioaleé, P., Odetti, F., Bingman, V., Siegel, J., and Vallortigara, G. (2001). "Hippocampus and homing in pigeons: left and right hemispheric differences in navigational map learning." *Eur J Neruosci*, 13, 1617–1624.

Gasser, T. and Muller, H. (1984). "Estimating regressive functions and their derivatives by the kernel method." *Scandinavian Journal of Statistics*, 11, 171–185.

Gaunt, P. and Lambert, B. (1987). "Single dose ciprofloxacin for the eradication of pharyngeal carriage of Neisseria meningitidis." *Journal of Antimicrobial Chemotherapy*, 21, 489–496.

Geisler, W. S. (2011). "Contributions of ideal observer theory to vision research." *Vision Research*, 51, 771–781.

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 6, 721–741.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). "Thresholding of statistical maps in functional neuroimaging using the false discovery rate." *NeuroImage*, 15, 870–878.

and

Georgopoulos, A. P., Ashe, J., ~~et al.~~ (2000). "One motor cortex, two different views." *Nature Neuroscience*, 3, 963–965.

Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., and Massey, J. T. (1982). "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex." *The Journal of Neuroscience*, 2, 1527–1537.

Gerstein, G. and Mandelbrot, B. (1964). "Random walk models for the spike activity of a single neuron." *Biophysical Journal*, 4, 41–68.

Geurts, H. M., Verté, S., Oosterlaan, J., Roeyers, H., and Sergeant, J. A. (2004). "How specific are executive functioning deficits in attention deficit hyperactivity disorder and autism?" *Journal of Child Psychology and Psychiatry*, 45, 836–854.

Georgopoulos, A.P. and Ashe, J.

Geweke, J. (1982). "Measurement of linear dependence and feedback between multiple time series." *Journal of the American Statistical Association*, 77, 304–313.

Gilovich, T., Vallone, R., and Tversky, A. (1985). "The hot hand in basketball: On the misperception of random sequences." *Cognitive Psychology*, 17, 295–314.

Glover, G. H. (1999). "Deconvolution of impulse response in event-related BOLD fMRI." *NeuroImage*, 9, 416–429.

Goebel, R., Esposito, R., and Formisano, E. (2006). "Analysis of FIAC data with BrainVoyager QX: From single-subject to cortically aligned group GLM analysis and self-organizing group ICA." *Human Brain Mapping*, 27, 392–401.

Gold, G., Giannakopoulos, P., Montes-Paixao, C., Herrman, F., Mulligan, R., Michel, J., and Bouras, C. (1997). "Sensitivity and specificity of newly proposed clinical criteria for possible vascular dementia." *Neurology*, 49, 690–694.

Goodman, S. N. et al. (1999a). "Toward evidence-based medical statistics. 1: The *P* value fallacy." *Annals of Internal Medicine*, 130, 995–1004.

italics

Goodman, S. N. et al. (1999b). "Toward evidence-based medical statistics. 2: The Bayes factor". Annals of Internal Medicine, 130, 1005–1013.

Gordon, A., Glazko, G., Qiu, X., and Yakovlev, A. (2007). "Control of the mean number of false discoveries, Bonferroni and stability of multiple testing." *Annals of Applied Statistics*, 1, 179–190.

Gould, S. (1996). *The Mismeasure of Man*. Norton.

Grace, A. A., Floresco, S. B., Goto, Y., and Lodge, D. J. (2007). "Regulation of firing of dopaminergic neurons and control of goal-directed behaviors." *Trends in Neurosciences*, 220–227.

Granger, C. (1969). "Investigating causal relations by econometric models and cross-spectral methods". *Econometrica*, 37, 424–438.

Greenhouse, J. B., Kass, R. E., and Tsay, R. S. (1987). "Fitting nonlinear models with ARMA errors to biological rhythm data." *Statistics in Medicine*, 6, 167–183.

Griffiths, T. L., Chater, N., Norris, D., and Pouget, A. (2012). "How the Bayesians got their beliefs (and what those beliefs actually are)." *Psychological Bulletin*, 138, 415–422.

Harrison, M., Amarasingham, A., and Kass, R. (2013). "Statistical identification of synchronous spiking." In *Spike Timing: Mechanisms and Function*, eds. P. DiLorenzo and J. Victor. Taylor and Francis, pp. 77–120.

Hartline, H. and Graham, C. (1932). "Nerve impulses from single receptors in the eye." *Journel of Cellular Comparative Physiology*, 1, 227–295.

Journal

a

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57, 97–109.

Hawkins, D. (1989). "Using U statistics to derive the asymptotic distribution of Fisher's Z statistic." *The American Statistician*, 43, 235–237.

Hawkins, T. (2001). *Lebesgue's Theory of Integration: Its Origins and Development*. American Mathematical Society.

Hébert, R. and Brayne, C. (1995). "Epidemiology of vascular dementia." *Neuroepidemiology*, 14, 240–257.

Hecht, S., Shlaer, S., and Pirenne, M. H. (1942). "Energy, quanta, and vision." *The Journal of General Physiology*, 25, 819–840.

Hill, A. (1971). *Principles of medical statistics*. 9th ed. Oxford University Press.

Hoeft, F., McCandliss, B., Black, J., Gantman, A., Zakerani, N., Hulme, C., Lyytinen, H., Whitfield-Gabrieli, S., Glover, G., Reiss, A., and Gabrieli, J. (2011). "Neural systems predicting long-term outcome in dyslexia." *Proceedings of the National Academy of Sciences*, 108, 361–366.

Hursh, J. B. (1939). "Conduction velocity and diameter of nerve fibers." *American Journal of Physiology*.

Ikegaya, Y., Aaron, G., Cossart, R., Aronov, D., Lampl, I., Ferster, D., and Yuste, R. (2004). "Synfire chains and cortical songs: Temporal modules of cortical activity." *Science*, 304, 559–564.

italics Theory of Probability, Oxford University Press.

Ikegaya, Y., Matsumoto, W., Chiou, H.-Y., Yuste, R., and Aaron, G. (2008). “Statistical significance of precisely repeated intracellular synaptic patterns.” *PloS ONE*, 3, 12, e3983.

Iyengar, S. and Liao, Q. (1997). “Modeling neural activity using the generalized inverse Gaussian distribution.” *Biological Cybernetics*, 77, 289–295.

Jacobs, R. A. and Kruschke, J. K. (2010). “Bayesian learning theory applied to human cognition.” *Wiley Interdisciplinary Reviews: Cognitive Science*, 2, 8–21.

Jarosiewicz, B., Chase, S. M., Fraser, G. W., Velliste, M., Kass, R. E., and Schwartz, A. B. (2008). “Functional network reorganization during learning in a brain-computer interface paradigm.” *Proceedings of the National Academy of Sciences*, 105, 19486–19491.

Jeffreys, H. (1931). *Scientific Inference*. Cambridge University Press.

Jeffreys, H. (1961). “~~Cartesian Tensors.~~” ~~Probability~~

Jeffreys, H. and Wrinch, D. (1921). “On certain fundamental principles of scientific inquiry.” *Philosoophical Magazine*, 42, 369–390.

Philosophical

Kalman, R. E. et al. (1960). “A new approach to linear filtering and prediction problems.” *Journal of Basic Engineering*, 82, 35–45.

Karpicke, J. D. and Roediger, H. L. (2008). “The critical importance of retrieval for learning.” *Science*, 319, 966–968.

Kass, R. E. (2011). “Statistical inference: the big picture (with discussion).” *Statistical Science*, 26, 1–20.

Kass, R. E., Kelly, R., and Loh, W.-L. (2011). “Assessment of synchrony in multiple neural spike trains using loglinear point process models.” *Annals of Applied Statistics*, 5, 1262–1292.

Kass, R. E. and Natarajan, R. (2006). “A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper).” *Bayesian Analysis*, 1, 535–542.

Kass, R. E. and Raftery, A. E. (1995). “Bayes factors.” *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E. and Steffey, D. (1989). “Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models).” *Journal of the American Statistical Association*, 84, 717–726.

Kass, R. E. and Ventura, V. (2001). “A spike-train probability model.” *Neural Computation*, 13, 8, 1713–1720.

Kass, R. E., Ventura, V., and Brown, E. (2005). “Statistical issues in the analysis of neuronal data.” *Journal of Neurophysiology*, 94, 8–25.

Kass, R. E., Ventura, V., and Cai, C. (2003). “Statistical smoothing of neuronal data.” *Network-Computation in Neural Systems*, 14, 5–16.

Kass, R. E. and Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley.

Kass, R. E. and Wasserman, L. (1995). “A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion.” *Journal of the American Statistical Association*, 90, 928–934.

Kass, R. E. and Wasserman, L. (1996). “The selection of prior distributions by formal rules.” *Journal of the American Statistical Association*, 91, 1343–1370.

Kaufman, C. G., Ventura, V., and Kass, R. E. (2005). “Spline-based non-parametric regression for periodic functions and its application to directional tuning of neurons.” *Statistics in Medicine*, 24, 2255–2265.

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). “Identifying natural images from human brain activity.” *Nature*, 452, 352–355.

Kelly, R. and Kass, R. E. (2012). “A framework for evaluating pairwise and multiway synchrony among stimulus-driven neurons.” *Neural Computation*, 24, 2007–2032.

Kelly, R. C., Smith, M. A., Kass, R. E., and Lee, T. S. (2010). “Local field potentials indicate network state and account for neuronal response variability.” *Journal of Computational Neuroscience*, 29, 567–579.

Kelly, R. C., Smith, M. A., Samonds, J. M., Kohn, A., Bonds, A., Movshon, J. A., and Lee, T. S. (2007). "Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex." *The Journal of Neuroscience*, 27, 261–264.

Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. Wiley.

Kent, L., Middle, F., Hawi, Z., Fitzgerald, M., Gill, M., Feehan, C., and Craddock, N. (2001). "Nicotinic acetylcholine receptor [alpha] 4 subunit gene polymorphism and attention deficit hyperactivity disorder." *Psychiatric Genetics*, 11, 37–40.

Klingsberg, T., Fernell, E., Olesen, P., Johnson, M., Gustafsson, P., Dahlstrom, K., Gillberg, C., Forssberg, H., and Westerberg, H. (2005). "Computerized training of working memory in children with ADHD - A randomized controlled trial." *Journal of the American Academy of Child and Adolescent Psychiatry*, 44, 177–186.

Knill, D. C. and Pouget, A. (2004). "The Bayesian brain: The role of uncertainty in neural coding and computation." *TRENDS in Neurosciences*, 27, 712–719.

Kolers, P. A. (1976). "Reading a year later." *Journal of Experimental Psychology: Human Learning and Memory*, 2, 554–565.

Kolmogorov, A. N. (1933). *Grundbegriffe der wahrscheinlichkeitsrechnung*. Springer-Verlag.

Konishi, S. and Kitagawa, G. (2007). *Information Criteria and Statistical Modeling*. Springer.

Körding, K. (2007). "Decision theory: What "should" the nervous system do?" *Science*, 318, 606–610.

Körding, K. P. and Wolpert, D. M. (2004). "Bayesian integration in sensorimotor learning." *Nature*, 427, 244–247.

Koyama, S., Chase, S. M., Whitford, A. S., Velliste, M., Schwartz, A. B., and Kass, R. E. (2010). "Comparison of brain-computer interface decoding algorithms in open-loop and closed-loop control." *Journal of Computational Neuroscience*, 29, 73–87.

Koyama, S. and Kass, R. E. (2008). "Spike train probability models for stimulus-driven leaky integrate-and-fire neurons." *Neural Computation*, 20, 1776–1795.

Kriegeskorte, N., Lindquist, M. A., Nichols, T. E., Poldrack, R. A., and Vul, E. (2010). "Everything you never wanted to know about circular analysis, but were afraid to ask." *J. Cereb. Blood Flow Metab.*, 30, 1551–1557.

Kullingsbaek, S. (2006). "Modeling visual attention." *Behavioral Research Methods*, 38, 123–133.

Kwon, H., Reiss, A. L., and Menon, V. (2002). "Neural basis of protracted developmental changes in visuo-spatial working memory." *Proceedings of the National Academy of Sciences*, 99, 13336–13341.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis*, 5, 369–411.

Lahiri, S. (2003a). "A necessary and sufficient condition for asymptotic independence of discrete Fourier transforms under short-and long-range dependence." *The Annals of Statistics*, 31, 613–641.

Lahiri, S. N. (2003b). *Resampling Methods for Dependent Data*. Springer.

Lanczos, C. (1966). *Discourse on Fourier series*, vol. 3. Edinburgh: Oliver & Boyd.

distribution
t

Levine, M. (1991). "The disribution of invervals between neural impulses in the maintained discharges of retinal ganglion cells." *Biological Cybernetics*, 65, 459–467.

intervals
t

Lewicki, M. (1998). "A review of methods for spike sorting: The detection and classification of neural action potentials." *Network: Computation in Neural Systems*, 9, R53–R78.

Lewicki, M. (2002). "Efficient coding of natural sounds." *Nature Neuroscience*, 5, 356–363.

Lewis, S. M., Jerde, T. A., Tzagarakis, C., Gourtzelidis, P., Georgopoulos, M.-A., Tsekos, N., Amirikian, B., Kim, S.-G., Uğurbil, K., and Georgopoulos, A. P. (2005). "Logarithmic transformation for high-field BOLD fMRI data data." *Experimental Brain Research*, 165, 447–453.

Li, D., Held, U., Petkau, J., Daumer, M., Barkhof, F., Fazekas, F., Frank, J., Kappos, L., Miller, D., Simon, J., et al. (2006). "MRI T2 lesion burden in multiple sclerosis: A plateauing relationship with clinical disability." *Neurology*, 66, 1384–1389.

Loader, C. (1999). *Local regression and likelihood*. New York: Springer.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). "Neurophysiological investigation of the basis of the fMRI signal." *Nature*, 412, 150–157.

Lucas, A., Morley, R., Cole, T., Lister, G., and Leeson-Payne, C. (1992). "Breast milk and subsequent intelligence quotient in children born preterm." *The Lancet*, 339, 261–264.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). "The BUGS Book: A Practical Introduction to Bayesian Analysis."

italics

Chapman and Hall/CRC Press.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Madiman, M. and Barron, A. (2007). "Generalized entropy power inequalities and monotonicity properties of information." *Information Theory, IEEE Transactions on*, 53, 2317–2329.

Makris, S. L., Raffaele, K., Allen, S., Bowers, W. J., Hass, U., Alleva, E., Calamandrei, G., Sheets, L., Amcoff, P., Delrue, N., et al. (2009). "A retrospective performance assessment of the developmental neurotoxicity study in support of OECD Test Guideline 426." *Environmental Health Perspectives*, 117, 17–25.

Manly, B. F. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall/CRC.

Marshall, J. C. and Halligan, P. W. (1988). "Blindsight and insight in visuo-spatial neglect." *Nature*, 336, 766–767.

Matsuzaka, Y., Picard, N., and Strick, P. L. (2007). "Skill representation in the primary motor cortex after long-term practice." *Journal of Neurophysiology*, 97, 1819–1832.

Mayo, D. G. and Spanos, A. (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge University Press.

McCullagh, P. and Nelder, J. (1989). *General linear models. Chapman and Hall.*

McGrayne, S. (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.

Metropolis, N. (1987). "The beginning of the Monte Carlo method." *Los Alamos Science (Special Issue dedicated to Stanislaw Ulam)*, 15, 125–130.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics*, 21, 1087–1091.

Mitra, P. P. and Pesaran, B. (1999). "Analysis of dynamic brain imaging data." *Biophysical Journal*, 76, 691–708.

Mosteller, F. and Tukey, J. (1968). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley.

Mudholkar, G. S. and Tian, L. (2002). "An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test." *Journal of Statistical Planning and Inference*, 102, 211–221.

Mullins, P. G., Rowland, L. M., Jung, R. E., and Sibbitt, W. L. (2005). "A novel technique to study the brain's response to pain: Proton magnetic resonance spectroscopy." *NeuroImage*, 26, 642–646.

Nagelkerke, N. (1991). "A note on a general definitino of the coefficient of determination." *Biometrika*, 78, 691–692. definition

Nelder, J. A. and Wedderburn, R. W. (1972). "Generalized linear models." *Journal of the Royal Statistical Society*. Series A, 135, 370–384.

Neyman, J. (1937). "Outline of a theory of statistical estimation based on the classical theory of probability." *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236, 333–380.

Nielsen, T. A., DiGregorio, D. A., and Silver, R. A. (2004). "Modulation of glutamate mobility reveals the mechanism underlying slow-rising AMPAR EPSCs and the diffusion coefficient in the synaptic cleft." *Neuron*, 42, 757–771.

Marshel, J., Garrett, M., Nauhaus, I. and Callaway, E. (2011) "Functional specialization of seven mouse visual cortical areas," Neuron, 72, 1040-1054.

Sellke
l

Selke, T., Bayarri, J., and Berger, J. (2001). "Calibration of p-values for testing precise hypotheses." *American Statistican*, 55, 62–71.

Statistician
i

Shadlen, M. N. and Newsome, W. T. (1998). "The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding." *The Journal of Neuroscience*, 15, 3870–3896.

Shruti, S., Clem, R. L., and Barth, A. L. (2008). "A seizure-induced gain-of-function in BK channels is associated with elevated firing activity in neocortical pyramidal neurons." *Neurobiology of disease*, 30, 323–330.

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications, with R Examples.* Springer New York.

Simmons, J., Nelson, L., and Simonsohn, U. (2011). "False-positive psychology: Undisclosed flexibility in data collection and analyis allow presenting anything as significant." *Psychological Science*, 22, 1359–1366.

analysis
5

Sklar, R. and Strauss, B. (1980). "Role of the *uvrE* gene product and of inducible *O6* -methylguanine removal in the induction of mutations by N-methyl- N-nitro- N-nitrosoguanidine in *Escherichia coli*." *Journal of molecular biology*, 143, 343–362.

Smith, A. C., Stefani, M. R., Moghaddam, B., and Brown, E. N. (2005). "Analysis and design of behavioral experiments to characterize population learning." *Journal of Neurophysiology*, 93, 1776–1792.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. (2011). "Network modelling methods for fMRI." *Neuroimage*, 54, 875–891.

Solt, K., Cotten, J., Cimenser, A., Wong, K., Chemali, J., and Brown, E. (2011). "Methylphenidate actively induces emergence from general anesthesia." *Anesthesiology*, 115, 791–803.

Sperling, G. (1967). "Successive approximations to a model for short term memory." *Acta psychologica*, 27, 285–292.

Stefani, M. R., Groth, K., and Moghaddam, B. (2003). "Glutamate receptors in the rat medial prefrontal cortex regulate set-shifting ability." *Behavioral Neuroscience*, 117, 728–737.

Stevens, S. (1961). "To Honor Fechner and Repeal His Law." *Science*, 133, 80–86.

Stevens, S. (1970). "Neural events and the psychophysical law." *Science*, 170, 1043–1050.

Stigler, S. M. (1986). *The History of Statistics. The Measurement of Uncertainty before 1900.* Cambridge, Mass.: Harvard.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions (with discussion)." *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 111–147.

Stopfer, M., Jayaraman, V., and Laurent, G. (2003). "Intensity versus identity coding." *Neuron*, 39, 991–1004.

Teich, M. C., Prucnal, P. R., Vannucci, G., Breton, M. E., and McGill, W. J. (1982). "Multiplication noise in the human visual system at threshold: 3. The role of non-Poisson quantum fluctuations." *Biological Cybernetics*, 44, 157–165.

Thompson, J. A., Wu, W., Bertram, R., and Johnson, F. (2007). "Auditory-dependent vocal recovery in adult male zebra finches is facilitated by lesion of a forebrain pathway that includes the basal ganglia." *The Journal of Neuroscience*, 27, 12308–12320.

Thomson, D. J. (1982). "Spectrum estimation and harmonic analysis." *Proceedings of the IEEE*, 70, 1055–1096.

Tibshirani, R. (2011). "Regression shrinkage and selection via the lasso: A retrospective (with discussion)." *J. Royal Statist. Soc. B*, 73, 273–282.

Tokdar, S., Xi, P., Kelly, R. C., and Kass, R. E. (2010). "Detection of bursts in extracellular spike trains using hidden semi-Markov point process models." *Journal of Computational Neuroscience*, 29, 203–212.

Tuckwell, H. (1988). *Introduction to Theoretical Neurobiology*, vol. 2: Nonlinear and Stochastic Theories. Cambridge.

Tukey, J. (1987). *The Collected Works of John W. Tukey*, vol. 4. Wadsworth.

Turner, R. and DeLong, M. (2000). "Corticostriatal activity in primary motor cortex of the macaque." *Journal of Neuroscience*, 20, 7096–7198.

Uhlhaas, P. J., Pipa, G., Lima, B., Melloni, L., Neuenschwander, S., Nikolić, D., and Singer, W. (2009). "Neural synchrony in cortical networks: History, concept and current status." *Frontiers in Integrative Neuroscience*, 3.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press ~~(Cambridge, UK and New York, NY, USA)~~.

Vandenbergh, R., Nelissen, N., Salmon, E., Ivanoiu, A., Hasselbalch, S., Andersen, A., Korner, A., Minthon, L., Brooks, D., Van Laere, K., and Dupont, P. (2013). "Binary classification of ^{18}F-flutemetamol PET using machine learning: Comparison with visual reads and structural MRI". Neuroimage, 64, 57–25.

Ventura, V., Cai, C., and Kass, R. (2005a). "Statistical assessment of time-varying dependence between two neurons." *J. Neurophys.*, 94, 2940–2947.

Ventura, V., Cai, C., and Kass, R. E. (2005b). "Trial-to-trial variability and its effect on time-varying dependency between two neurons." *Journal of neurophysiology*, 94, 2928–2939.

Ventura, V., Carta, R., Kass, R., Gettner, S., and Olson, C. (2002). "Statistical analysis of temporal evolution in single-neuron firing rates." *Biostatistics*, 3, 1–20.

Vu, V. Q., Ravikumar, P., Naselaris, T., Kay, K. N., Gallant, J. L., and Yu, B. (2011). "Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models." *The Annals of Applied Statistics*, 5, 1159–1182.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). "Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method." *Cognitive psychology*, 60, 158–189.

Wallstrom, G. L., Kass, R. E., Miller, A., Cohn, J. F., and Fox, N. A. (2002). "Correction of ocular artifacts in the EEG using Bayesian adaptive regression splines." In *Case Studies in Bayesian, Statistics*, 351–366. Springer-Verlag.

Wallstrom, G. L., Kass, R. E., Miller, A., Cohn, J. F., and Fox, N. A. (2004). "Automatic correction of ocular artifacts in the EEG: A comparison of regression-based and component-based methods." *International Journal of Psychophysiology*, 53, 105–119.

Wang, W., Sudre, G. P., Xu, Y., Kass, R. E., Collinger, J. L., Degenhart, A. D., Bagic, A. I., and Weber, D. J. (2010). "Decoding and cortical source localization for intended movement direction with MEG." *Journal of Neurophysiology*, 104, 2451–2461.

Wasserman, L. (2004). *All of Statistics*. Springer.

Watson, R. and Tang, D. (1980). "The predictive value of prostatic acid phosphates as a screening test for prostatic cancer." *New England Journal of Medicine*, 303, 497–499.

Weinberg, S. (2002). *It Must Be Beautiful: Great Equations of Modern Science*, chap. Afterword: How great equations survive. Granta Press.

Welch, P. (1967). "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms." *Audio and Electroacoustics, IEEE Transactions on*, 15, 70–73.

Whitmore, G. and Seshadri, V. (1987). "A heuristic derivation of the inverse Gaussian distribution." *The American Statistician*, 41, 280–281.

Wolpert, D. M., Diedrichsen, J., and Flanagan, J. R. (2011). "Principles of sensorimotor learning." *Nature Reviews Neuroscience*, 12, 739–751.

Wood, F., Black, M. J., Vargas-Irwin, C., Fellows, M., and Donoghue, J. P. (2004). "On the variability of manual spike sorting." *Biomedical Engineering, IEEE Transactions on*, 51, 912–918.

Wu, C.-F. (1981). "Asymptotic theory of nonlinear least squares estimation." *The Annals of Statistics*, 9, 501–513.

Wu, W., Gao, Y., Bienenstock, E., Donoghue, J. P., and Black, M. J. (2006). "Bayesian population decoding of motor cortical activity using a Kalman filter." *Neural computation*, 18, 80–118.

Xu, Y., Sudre, G. P., Wang, W., Weber, D. J., and Kass, R. E. (2011). "Characterizing global statistical significance of spatio-temporal hot spots in MEG/EEG source space via excursion algorithms." *Statistics in Medicine*, 30, 2854–2866.

Example Index

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

Index

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

Analysis of variance, see ANOVA

ALL INDEX ENTRIES WITH ABBREVIATIONS SHOULD BE LISTED TWICE
WITH PAGE REFERENCES APPEARING WITH THE ABBREVIATED VERSION
I.i.d. (independent and identically distributed), 137

Springer Series in Statistics

Robert E. Kass · Uri T. Eden · Emery N. Brown

Analysis of Neural Data

Continual improvements in data collection and processing have had a huge impact on brain research, producing data sets that are often large and complicated. By emphasizing a few fundamental principles, and a handful of ubiquitous techniques, *Analysis of Neural Data* provides a unified treatment of analytical methods that have become essential for contemporary researchers. Throughout the book ideas are illustrated with more than 100 examples drawn from the literature, ranging from electrophysiology, to neuroimaging, to behavior. By demonstrating the commonality among various statistical approaches the authors provide the crucial tools for gaining knowledge from diverse types of data. Aimed at experimentalists with only high-school level mathematics, as well as computationally-oriented neuroscientists who have limited familiarity with statistics, *Analysis of Neural Data* serves as both a self-contained introduction and a reference work.

Robert E. (Rob) Kass is Professor in the Department of Statistics, the Machine Learning Department, and the Center for the Neural Basis of Cognition at Carnegie Mellon University. Since 2001 his research has been devoted to statistical methods in neuroscience. Together with Emery Brown he has organized the highly successful series of international meetings, Statistical Analysis of Neural Data (SAND).

Uri T. Eden is Associate Professor in the Department of Mathematics and Statistics at Boston University. He received his Ph.D. in the Harvard/MIT Medical Engineering and Medical Physics program in the Health Sciences and Technology Department. His research focuses on developing mathematical and statistical methods to analyze neural spiking activity, using methods related to model identification, statistical inference, signal processing, and stochastic estimation and control.

Emery N. Brown is Edward Hood Taplin Professor of Medical Engineering, Professor of Computational Neuroscience, and Associate Director of the Institute of Medical Engineering and Science at MIT; he is also the Warren M. Zapol Professor of Anaesthesia at Harvard Medical School and Massachusetts General Hospital. He is both a statistician and an anesthesiologist. Since 1998 his research has focused on neural information processing, and his experimental work characterizes the way anesthetic drugs act in the brain to create the state of general anesthesia.

Statistics/Life Sciences,
Medicine, Health Sciences

ISBN 978-1-4614-9601-4

9781461496014

▶ springer.com