You have just completed data collection for the first subject of your first fMRI experiment. After careful consideration, you decided to use a simple two-condition blocked design. During task blocks, a series of famous names (e.g., Bill Clinton, David Beckham) flashed across the screen, while during non-task blocks, the names were of people unknown to the subject. You hypothesized that the task condition would be associated with greater activation in the fusiform gyrus, which is critical for face processing, because subjects would imagine the faces that go with the famous names. So, to evaluate this hypothesis, you calculated that the mean activation in the fusiform gyrus was 500 units during the famous names blocks and 498 units during the unknown names blocks. As you stare at these numbers, you realize that your hypothesis remains unanswered. The averages are numerically different, to be sure, but is the difference meaningful?

This example illustrates the use of descriptive statistics, or summaries of a data set. Any set of numbers can be described using statistics such as the mean, median, or standard deviation. But regardless of the experiment, the recorded data from a single subject do not provide a complete and unerring description of the world. Instead, the data represent just a single sample, or one set of observations out of the many that might have occurred in the experiment. If our subject had performed the task differently, or if we had selected a different subject, our data could have been very different. It is possible that the numerical differences between blocks, that we measured in this subject, were due to random variation, and would disappear if we used the subject in the same experiment a second time. As experimenters, we want to do more than just describe our observations; we want to make inferences about the underlying processes involved. Stated another way, we do not want to know merely whether the fMRI activation in the task block was greater than that in the non-task block within the particular set of data we recorded. Instead, we want to know whether the difference between conditions would be replicated in repeated observations from the same subject, from the same group of subjects, or from the population as a whole. To make such judgments, we must use inferential statistics that provide estimates of our certainty in the experimental hypothesis.

Recall from Chapter 9 that experiments should be designed to test a research hypothesis, often symbolized as H₁, that states a possible relationship...
null hypothesis  The proposition that the experimental manipulation will have no effect on the experimental data. Most statistical analyses evaluate the probability that the null hypothesis is true, i.e., that the observed data reflect chance processes.
significance testing  The process of evaluating whether the null hypothesis is true. Also known as hypothesis testing.
hypothesis-driven analysis  The evaluation of data based on statistical tests of the validity of a null hypothesis.

between independent and dependent variables. Often only one hypothesis is stated, however, all experiments are designed to discriminate between two possible hypotheses, the research hypothesis and another, null hypothesis ($H_0$), which usually states that the manipulation has no effect. For the example described above, the null hypothesis ($H_0$) would be “Reading famous versus unknown names has no effect on fMRI activation in the fusiform gyrus.” All well-formed hypotheses must be falsifiable, such that either the research or the null hypothesis, but not both, must be true. This can be seen by examining the typical mathematical form of the hypotheses:

$$H_1 : \text{Condition}_1 \neq \text{Condition}_2$$
$$H_0 : \text{Condition}_1 = \text{Condition}_2$$

Since the null hypothesis assumes that the independent variable has no effect, it predicts that the observed values of the dependent variable will be similarly distributed between the conditions. In the blocked design in our example above, the null hypothesis could assert that data obtained in the two conditions were drawn from the same distribution (i.e., same baseline level of activation + random, normally distributed noise), such that any differences between the blocks were caused by random fluctuations in the MR signal. For example, if the mean fMRI response during different non-task blocks ranged from 450 to 550 units, a value of 500 in a task block would be consistent with the null hypothesis. But if typical values for non-task blocks were from 497.95 to 498.05 units, then a task value of 500 would be unlikely to be due to chance, and we could reject the null hypothesis. The process of evaluating whether the observed data reflect experimental manipulations or chance processes is known as significance testing.

In the following sections, we explore a number of different approaches to the statistical evaluation of fMRI data. All of these hypothesis-driven analysis approaches differ in their assumptions and goals, but they share some common features. First, they express significance as the probability that the results could occur under the null hypothesis. The color maps that form the basis of many figures in fMRI manuscripts almost always express this probability, often with brighter, more-intense colors indicating a very low probability that the experimental effects are due to chance (Figure 10.1). Second, voxels whose probabil-

Figure 10.1  Statistical maps of fMRI data. Functional MRI data are usually displayed using a background anatomical MRI image with an overlaid statistical map (A). In the statistical map, voxels whose activation levels pass some threshold value are shown in color, with the intensity of the color corresponding to the degree of significance (B). More extreme significance values (lower probabilities) are usually shown in brighter colors. It is important to recognize that the colors represent the output of some statistical test, not absolute data values.
Figure 10.2 Types of experimental errors. When testing research hypotheses, there are four possible outcomes. The experimenter may reject the null hypothesis when it should be rejected; this is sometimes called a "hit," and in fMRI it corresponds to successfully identifying an active voxel. Rejecting the null hypothesis when it is in fact true is known as a Type I error, and in fMRI, this corresponds with labeling a voxel as active when it is not. Accepting the null hypothesis when it is in fact false is a Type II error; in fMRI, this is labeling a voxel as inactive when it is active. Type II errors are common in fMRI. Finally, accepting the null hypothesis when it is indeed true is called a "correct rejection."

Basic Statistical Tests

The simplest and oldest of all statistical tests can be used on the data shown in Figure 10.3. The data are the BOLD changes in a motor cortex voxel, associated with brief hand movements made in response to the periodic presentation of a 3 s duration stimulus. Following every presentation of the stimulus, there was an increase in signal that lasted for about 10 s. How do we know whether this voxel exhibits task-related activation? In this case, we can use the venerable interocular trauma test, which can be stated succinctly: the comparison is significant if the data, when plotted, hit you between the eyes. In the

alpha value An a priori probability (e.g., 0.001) chosen as the threshold for statistical significance. If the probability that the data would be obtained under the null hypothesis is less than the alpha value, the data are considered to be statistically significant.

Type I error Rejecting the null hypothesis when it is in fact true. Also known as a false positive.

Type II error Accepting the null hypothesis when it is in fact false. Also known as an incorrect rejection or false negative.

Interocular trauma test An intuitive test of significance based on highly visible effects of the experimental manipulation. It states that data are significant if, when plotted, they hit you between the eyes.
Figure 10.3 Use of the interocular trauma test for statistical significance. For some data, the effect of the experimental manipulation can be readily ascertained just by looking at the data. In this voxel, each time a brief visual stimulus was presented (arrows), there were significant increases in activation. From looking at these data, it is obvious that stimulus presentation elicited an effect. Most fMRI data, however, are not so easily analyzed!

plotted data, the effect of the independent variable is obvious, because every time the stimulus was presented there was a very large change in the dependent variable. If only all fMRI data were so simple to analyze! As discussed in Chapter 8, the SNR in most fMRI experiments is quite low, so effects are rarely so easy to spot in the raw data. Researchers must instead use formal significance testing to evaluate whether the experimental manipulation has any effect.

In this section, we describe some simple forms of statistical analysis and how they can be applied to fMRI data. But recognize that, with minor exceptions, these types of analyses are no longer common in fMRI research. They have been replaced by the more powerful and flexible regression approaches described in the following section. Nevertheless, they provide a good introduction to the two core concepts of fMRI data analysis: assessing whether two or more independent variables have significantly different effects on the BOLD signal, and creating models for the expected effects of independent variables.

Contrasting experimental conditions: the t-test

In a standard two-condition blocked design, the null hypothesis is simple: the difference between the conditions has no effect on the fMRI data. As described in the introduction to this chapter, a simplenminded way to compare the conditions would be to calculate the difference between the means of the data produced under each condition (i.e., 500 units versus 498 units). This comparison follows the logic of subtraction advanced in the previous chapter. However, a difference between condition means, by itself, is uninformative. It is necessary to evaluate whether that difference is large compared with some measure of the variability in the data, such as the standard deviation.

So, under the null hypothesis, any difference between the mean of the fMRI data recorded in Condition 1 and the mean of the data recorded in Condition 2 is due to random chance. The t-distribution describes the expected difference between two random samples drawn from the same distribution (Figure 10.4). The mean of the t-distribution is zero, since the two samples should on average have the same mean value, and the standard deviation of the t-distr-
Figure 10.4 The normal distribution and the Student’s t-distribution. The normal distribution is a bell-shaped curve that describes the variability in many random processes, and is the distribution of data from a large number of independent events. Many variables are distributed normally. The Student’s t-distribution is a distribution of the means of samples drawn from a larger population, which may or may not be normally distributed. The t-distribution resembles the normal distribution at larger sample sizes (n > 30), but with fewer degrees of freedom, it is narrower near the center and has more of its values near the tails (shaded areas, percentages indicated below).

To conduct a t-test (Equation 10.1 and Figure 10.5), the researcher calculates the means for all data points in the two conditions (and ) and divides their difference by the shared standard error (σxy):

\[ t = \frac{\bar{x} - \bar{y}}{\sigma_{xy}} = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma^2_x + \sigma^2_y}} \]  

(10.1)

The resulting t-statistic can then be converted to a probability value based on the degrees of freedom (df), or the number of unconstrained data points. For many statistical tests, the number of degrees of freedom within a sample is equal to the number of data points minus one. As an example, for a sample of 20 data points with a known mean value, there are 19 degrees of freedom, since if you know 19 of the data points and the mean, you can calculate the twentieth data point. Once the probability of the t-test has been determined using a statistical table or calculator, the researcher compares that probability with the alpha value for the experiment. For example, imagine that 25 time points were collected for each condition, and that the difference between the means is seven units, and the shared standard error is two units. The resulting t-statistic has

t-test A test for statistical significance based on the Student’s t-distribution. The t-test typically evaluates whether the mean values of two sets of observations are sufficiently different to preclude their being drawn from the same distribution.

standard error A commonly used estimate of the likely discrepancy between a measured value and a true value, often calculated from a measure of variability in the data (i.e., the standard deviation) and the number of data points in the sample.

degrees of freedom (df) The number of independent observations within a data set. For many statistical tests, there are n - 1 degrees of freedom associated with n data points.

Figure 10.5 Conducting a t-test. The t-test compares the size of an effect (e.g., the difference between blocks) with the variability in the data (i.e., the shared standard error of that difference). Shown here are two simulated fMRI time courses. In plot A, the effect of the experimental manipulation is 2 units in amplitude, but the variability is relatively high, so the t-statistic is about 2.3. In plot B, the manipulation only has a 1-unit effect, but the variability is much smaller, and the resulting t-statistic (6.7) is much higher.
contrast (1) The intensity difference between different quantities being measured by an imaging system. It also can refer to the physical quantity being measured (e.g., $T_1$ contrast). (2) A statistical comparison of the activation evoked by two (or more) experimental conditions, in order to test a research hypothesis.

A value of 3.5. The researcher wants to evaluate this statistic against the experiment's alpha value, which has been set at 0.01. With 48 degrees of freedom (i.e., 24 from each group), we can calculate that there is less than a 0.001 chance that the data in these conditions were drawn from the same distribution. This probability is lower than the threshold alpha value, and thus the null hypothesis can be rejected.

Here we introduce the first of the core concepts for fMRI data analysis: experimental contrast. This very important term refers to a comparison between the activation levels evoked by two independent variables (or two levels of the same independent variable). The concept of experimental contrast has broad similarities to the idea of image contrast introduced in Chapter 1, where the term was used to describe differences between two forms of tissue in anatomical imaging. In a deep sense, the fundamental goal of fMRI analyses is to evaluate whether an experimental manipulation evoked a meaningful change in activation; in other words, whether a contrast between two conditions was statistically significant.

A primary challenge in adapting the logic of a $t$-test to fMRI lies in deciding which time points should be assigned to which experimental conditions. This problem is illustrated in Figure 10.6. Consider a standard alternating blocked-design fMRI study with two conditions, each of 20 s duration. If you repeated this design twice and collected fMRI data with a 1 s TR, then you would have a total of 80 time points in the dataset. Which time points should be assigned to condition A and which should be assigned to condition B? An obvious first option would be to assign the first 20 points to A, the next 20 to B, and so forth. But remember from Chapter 7 that the BOLD fMRI response lags behind neuronal activity. Therefore, a better approach would account for this lag by delaying the onset of all blocks (e.g., by about 6 s). But even this approach is imperfect. Because changes in the BOLD response are not instantaneous, there will be transition periods at the onset of each block where the measured fMRI signal will be changing from low to high or from high to low. By excluding these highly variable transition periods and sampling only the

![Figure 10.6](image_url)
latter parts of each block, the researcher can select time points where the BOLD response has reached a steady state, maximizing the power of the t-test.

The t-test is used with the assumption that the data are drawn from normal distributions with equal variability, however, one of the strengths of the t-test for fMRI is that it is relatively insensitive to violations of this assumption. Even in a short fMRI session, there will be at least a few tens of time points collected per voxel, and in many sessions, hundreds of data points are collected per voxel for each condition. At these sample sizes, deviations from normality have no meaningful effect on the outcome of the test. Likewise, with large samples, the conditions may have very different numbers of time points or have different variability without compromising the outcome of the t-test. However, although the t-test may be statistically valid for fMRI data, several potential concerns do exist. Any systematic difference between the experimental conditions, whether associated with meaningful BOLD activation or with uninteresting artifacts like scanner drift or head motion, could result in a significant t-test (Figure 10.7). This is particularly a problem in blocked studies that have only a few cycles of the task and non-task conditions. The t-test is also inappropriate for answering questions about the timing of activation, since it combines data from all time points within a condition. Note also that while the t-test evaluates differences between the means of two distributions, it is insensitive to differences in their variability or shape. To find such differences, a Kolmogorov-Smirnov (K-S) test can be used.

While this discussion has focused on blocked designs, t-tests are also useful for some analyses of event-related designs. Remember that the basic role of the t-test is to identify a significant difference between the means of two samples of data. If a research question focuses on specific data points within an event-related design, then a t-test may be appropriate. Event-related designs may compare the magnitude of fMRI activation at a given time point between two different experimental conditions. For example, in a study published in

Figure 10.7 Effects of scanner drift on t-tests. Shown here is an activation map of a phantom, with positively significant voxels shown in the green-to-yellow color range and negatively significant voxels shown in the blue-to-pink color range. The position of the phantom “moved” slightly along the frequency-encoding direction (top to bottom) due to slow changes in the center frequency of the scanner over time. Even though there was no true activation, this motion within the images was significant according to the t-test.
2006 Cantlon and colleagues presented subjects with a series of visual patterns, most of which had the same number of elements, all with the same shape. In deviant trials, the pattern changed to a different number of elements (e.g., 32 instead of 16) or to elements with a different shape (e.g., squares instead of circular dots). The authors calculated the mean signal change over a pre-stimulus baseline in each voxel, for each time point from 3 s before to 12 s after every trial. A set of t-tests were used to evaluate whether activation in each voxel at the expected peak of the hemodynamic response (4.5–7 s after each event) was significantly greater after deviants with a number change versus deviants with a shape change. (Note that another set of between-subjects t-tests were used to determine whether the distribution of significance values in each voxel differed from chance expectation. This will be discussed in the section on Inter-subject Analyses later in this chapter.)

In summary, regardless of the experimental design, if the research question can be answered by evaluating whether or not two samples have statistically different means, then a t-test may be appropriate. Yet, univariate statistical tests such as the t-test have significant weaknesses. Because they are used to compare samples that differ in only one independent variable they are not appropriate when the experiment involves multiple variables. These types of test are best used when data points can be classified into distinct categories (e.g., activated versus not activated); conversely, they are difficult to apply to complex event-related designs. And, because these tests compare the relative magnitude of activation between conditions, they cannot be used to determine whether a given event type evoked a pattern of activation that was consistent with a predicted shape for the fMRI hemodynamic response. In the next section we consider a type of analysis that overcomes some of these limitations (while introducing new limitations of its own).

Comparing experimental and predicted responses: correlation analyses

While the t-test can be applied to many fMRI studies, it cannot be used to analyze the shape of the hemodynamic response. Indeed, when excluding transitions between task blocks, information about the shape of the response is explicitly removed. Nevertheless, the fMRI signal does contain important timing information. As discussed in Chapter 7, the fMRI hemodynamic response takes about 5 s to rise to its maximum after the onset of brief neuronal activity. Following the cessation of neuronal activity, the hemodynamic response falls over a period of 5 to 10 s and then stabilizes at a below-baseline level for an extended interval. The consistency of this pattern in the hemodynamic response allows us to predict the change in fMRI activation that should be evoked in an active voxel. Using a correlation analysis, we can quantify how well the observed data match a canonical hemodynamic response, thus identifying voxels in which the fMRI time course reflects underlying neuronal activity. Correlation analyses were first reported in fMRI by Bandettini and colleagues in 1993 and have since been an important part of fMRI analyses. In early fMRI studies, correlation tests were used to identify voxels with significant levels of activation. Now, the idea of correlation forms the core of regression analyses.

Conducting a correlation analysis on fMRI data is very simple. First, identify an epoch from your experimental data that should contain some task-related signal change, and predict the hemodynamic response that should be observed during that epoch. Second, calculate the covariance between the
The experimental data (here, $x$) and the predicted hemodynamic response (here, $y$), which is indicated by the numerator in the right term of Equation 10.2:

$$r = \frac{1}{n-1} \sum \frac{(x-x)(y-y)}{\sigma_x \sigma_y}$$

(10.2)

In essence, that numerator takes each observation (e.g., each time point), compares its experimental and predicted values to their mean values, and then sums the product over every observation. A positive covariance indicates that when the experimental data were large, the predicted data tended to be large; likewise, when one was small, the other was small. But if the covariance is negative, the values of the experimental data tended to be small while the values of the predicted data were large, or vice versa.

The third step in conducting the correlation analysis is to normalize the covariance by dividing it by the product of the standard deviations of the two epochs ($\sigma_x, \sigma_y$). The resulting correlation coefficient, or $r$-value, can range from $-1.0$ to $+1.0$, or from perfect positive correlation to perfect negative correlation. A correlation of 0 indicates that the experimental data are not related to the predicted data. As with the $t$-test discussed in the previous section, the significance of the correlation coefficient can be evaluated using statistical tables based on the degrees of freedom; a correlation of 0.5 is more likely to be significant when based on 1000 data points than when based on 10. This basic correlation analysis is then repeated for every voxel in the brain to create the map of significant activation.

We now introduce the second of the core concepts for fMRI data analysis: building a model of the predicted fMRI response. In the language of statistics, a model is a collection of independent variables that together predict the likely outcome of the dependent variable. Each independent variable is called a regressor (also known as a predictor variable or a model factor). A basic fMRI correlation analysis is equivalent to a model with one regressor: a canonical hemodynamic response. If a voxel's data match the predicted values of the regressor, as determined by the significance of the correlation coefficient, then that voxel is considered to be activated by the experimental manipulation (Figure 10.8).

Given the differences between correlation and $t$-tests, it may surprise you that when they are applied to the same data set, they give identical results. Remember that the $t$-test evaluates whether data derived from one condition

correlation coefficient ($r$-value) A number between $-1$ and 1 that expresses the strength of the correlation between two variables.

model A collection of independent variables (and the relationships between those variables) that serve to predict a dependent variable.

regressor A hypothesized time course of BOLD activation caused by the manipulations of an independent variable or by another known source of variability.

![Figure 10.8](https://example.com/figure10.8.png)

**Figure 10.8** Correlation analyses match experimental data with a hypothesized hemodynamic response. Shown in blue is a hypothetical mean hemodynamic waveform, as would be obtained from averaging many trials of the same event (e.g., a flash of a visual checkerboard). The red curve indicates the hypothesized BOLD hemodynamic response. The correlation between these two time-courses can be calculated as $r = 0.81$, which would reflect a highly significant correlation. While few current fMRI studies use such literal correlations, the concept of comparing observed data to a predicted waveform underlies nearly all fMRI analyses.
differ from data obtained during another condition. Exactly the same test could be conducted by correlating the experimental data with a predicted data set that follows a boxcar waveform (i.e., in task blocks the predicted response would be 1, while in non-task blocks the response would be 0). This similarity is often seen in graphical representations of experimental designs, in which the different conditions are plotted as different values along the y-axis, and can be demonstrated using any sample data set and Equations 10.1 and 10.2. For any value of $r$, there is a corresponding value of $t$, given the degrees of freedom in the data. Note also that both tests measure signal change divided by non-signal variability, with the $t$-test using the difference between means and the correlation test using the more general measure of covariation. Thus, the power of the correlation coefficient (like the $t$-test) rests on having maximal variability in the signal of interest compared with noise. Furthermore, if the values of either the experimental or predicted data are distributed in a highly non-normal fashion, then the correlation statistic may not be meaningful. This may occur in fMRI studies if there are very long pre-stimulus or post-stimulus baseline periods, so that most data points in the prediction epoch are near zero.

Correlation analyses have often been used in conjunction with signal averaging. Averaging across stimulus epochs improves the estimate of the time course of activation in a voxel or region, increasing the power of a correlation test for detecting significant voxels. And, collecting epochs from different times within an experiment can minimize some potential sources of variability in the hemodynamic response (e.g., low-frequency drift, practice or fatigue effects). Signal averaging can also inform the choice of the predicted hemodynamic response. Using some canonical function may seem attractive, however, recall from Chapter 8 that the characteristics of the hemodynamic response may differ between subjects, brain regions, and stimuli. Such differences can reduce the significance of correlation tests. One way of overcoming this problem is to generate a unique hemodynamic response for each subject, based on averaged data collected during a screening run, using a task that targets some region of interest.

Correlation analyses have significant limitations. Identifying discrete epochs can be challenging if events of interest overlap or follow closely in time. Conversely, if events are widely spaced, some of the fMRI data will make minimal contributions to the analyses. Another problem is inherent in the correlation statistic, which depends only on the covariance between two variables. Suppose that you compare responses to two types of stimuli (e.g., upright and inverted faces) that evoke activation in the same brain region, but with one (e.g., upright faces) generating twice the activation amplitude as the other. The time course of activation evoked by each of these stimuli could be the same, making them each highly correlated with a predicted hemodynamic response. To test whether these two stimuli evoked different levels of activation, you would need to use a statistic that compares the amplitude of the evoked activation between two conditions, such as the $t$-test described in the previous section. And, for experimental designs that involve events or blocks that occur with a regular frequency, analysis methods that directly measure those frequency changes can be much more sensitive (see Box 10.1).

We emphasize that the techniques described so far, the $t$-test and correlation analyses, have complementary strengths. The former allows us to evaluate contrasts between experimental conditions, whereas the latter allows us to create models for expected fMRI activation. Both of these strengths are manifest in regression modeling, which now forms the basis for nearly all fMRI data analysis.
BOX 10.1 Identifying Task-Related Periodicity: Fourier Analyses

As discussed in Chapter 9, a blocked-design fMRI task presents stimulus conditions at regular intervals. As a consequence, the MR signal within an active voxel regularly rises during task blocks and falls during non-task blocks. The periodic nature of this signal change can be quantified using a Fourier transform. The Fourier transform expresses a temporally (or spatially) varying signal as the linear sum of a series of sine waves of different frequencies, amplitudes, and phases. A plot of the magnitude of each sine-wave component necessary to recreate the original signal is called a power spectrum. Thus, the power spectrum itself is another way of representing the original data. In the language of signal processing, the raw fMRI time series data are in the time domain, meaning that they show the relative intensity of the signal at each time point. The power spectrum represents the same data in the frequency domain, indicating the intensity of the signal at each component frequency. If a task-related signal rises and falls at a known frequency, then a peak will

Figure 1 Use of a Fourier analysis to calculate frequency and phase information from the BOLD signal. This patient, who had an arteriovenous malformation (AVM) in the left hemisphere (right side of image at center), participated in blocked-design motor squeeze tasks that began either with a left-hand squeeze or a right-hand squeeze. Data from primary motor cortex in each hemisphere are shown in the graphs, with the left hemisphere data shown on the right side of the figure. The raw data are shown in the upper graphs, the power spectra are shown in the middle graphs, and the phase at the task frequency (peak in spectra) is shown at bottom. Note that although both hemispheres had peaks at about the same frequencies, the phases were different, allowing dissociation of right-hand and left-hand activation.

(continued on next page)
occur at that frequency in the power spectrum. We discussed in Chapter 8 the use of the Fourier transform to remove unwanted variability from the data, and here we extend that discussion to consider its use for statistical analysis of task-related variability.

In practice, the frequencies that can be measured by a Fourier analysis depend on how often the BOLD time series is sampled. The basic rule of sampling, the Nyquist sampling theorem, states that to accurately measure a given frequency, you must sample at a minimum of twice that frequency. Thus, the Fourier transform of $n$ time points sampled at a given TR contains $n/2$ frequencies ranging from 0 Hz to $1/(2 \times \text{TR})$ Hz, which are represented along the $x$-axis of the power spectrum. The first frequency component, at 0 Hz, represents the mean intensity of the signal and is often called the DC component, after the electrical term for direct current (i.e., a constant-voltage power source). Since fMRI time courses are generally represented in arbitrary units with positive values proportional to the amount of current through the receiver coil, the DC component is generally positive and very large. Also present in almost all fMRI data, even from very stable scanners, is substantial low-frequency power associated with scanner drift, among other factors (see Chapter 8). There are also slow physiological changes due to vascular oscillations, although such effects are incompletely understood. Because of this power at low frequencies, very long block lengths are not ideal for fMRI.

Figure 2 The use of overlapping blocks with different frequencies. If several different classes of stimuli are presented at different frequencies, a Fourier transform can be used to separate activation associated with each. Here, faces are presented most frequently, animals are presented at an intermediate frequency, and objects are presented least frequently. Each is presented individually in an alternating blocked design, with scrambled objects as the control condition. Note that what subjects see changes at different points in the task, as shown in the sample visual displays at the top of the figure. Some of the time, all three categories will be present, while at other times, only one or two of the categories will be visible.
Regression Analyses

Nearly all fMRI data analyses share a single goal: model testing. Based on the timing and duration of events—or, more specifically, the timing and duration of the evoked neuronal activity—researchers generate a predicted hemodynamic response, guided by the concepts of scaling and superposition introduced in Chapter 7. These models contain predicted time courses for the entire session, rather than for individual epochs as described above for correlation analyses. And, they typically contain several distinct predictions (i.e., regressors) that correspond to different hypothesized processes (e.g., visual processing, retrieval from memory, and motor responses). The relative contribution of each of these regressors to the measured data, within each voxel, is then statistically evaluated using a technique known as multiple regression.

The core idea of regression is that the value of the observed data \( y \) can be attributed to two sources: a model comprising a linear combination of several regressors \( x_i \), each with a variable weighting \( \beta_i \); and residual noise in the data, or error in the measurements \( \epsilon \). The basic formula for a regression analysis is given in Equation 10.3:

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon
\]

By convention, the Greek letter beta represents the parameter weight \( \beta \), or how much each regression factor contributes to the overall data. The term \( \beta_0 \) reflects the total contribution of all factors that are held constant throughout the experiment. For fMRI data, this would include the baseline signal intensity (i.e., mean \( T_2^* \) signal given the pulse sequence parameters) in each voxel, as well as any evoked activation that is constant throughout the experiment. Researchers sometimes colloquially refer to the parameter weights, which are calculated during regression analyses, as “beta values” or “betas.” The symbol \( \beta \) should not be confused with the roman letter B, which is used to indicate magnetic field strength (or a magnetic field vector, \( B \)).

Regression models like that in Equation 10.3 have only one known quantity: the experimental data \( y \). The regressors \( x_i \) represent hypothesized factors that may or may not contribute to the data. Given the data and a specified set of regressors, the researcher can identify a combination of parameter weights that minimizes the error term. To evaluate the statistical significance of a regressor, the amount of variability it explains (when multiplied by its best-fitting parameter weight) is compared with the amount of variability explained by the error term. This statistical approach, when applied to data sets with many dependent variables, is known as the general linear model (GLM).

The general linear model: an overview

In fMRI experiments, the simple equation given above for the general linear model (Equation 10.3) is replaced by a set of matrices (Figure 10.9). The fMRI data \( y \) are represented as a two-dimensional data matrix consisting of \( n \) time points by \( V \) voxels. Note that the spatial structure of the fMRI data is not used in the general linear model, since the values of the parameter weights and error term are calculated independently for each voxel. Instead, all voxels in the imaging volume are arranged along one dimension, for ease of calculation. The design matrix, which specifies the linear model to be evaluated, consists of \( M \) regressors, each \( n \) time points in length. In some notation systems, the design...
matrix is denoted as \( G \). The \textbf{parameter matrix} contains \( M \) parameter weights and \( V \) voxels, such that each cell indicates the \( \beta \)-value for a given voxel. Finally, the \textbf{error matrix} expresses the residual error for each voxel, and thus is an \( n \)-by-\( V \) matrix. Of these four matrices, the data are obtained experimentally, the design matrix is constructed by the experimenter based on the hypothesized effects of the experimental manipulations, and the parameter weights and residual error are calculated during the analysis.

The general linear model is elegant in its simplicity. After the data are obtained and the design matrix has been established, only one question must be answered: what values for the parameter matrix lead to the smallest values in the error matrix? To understand this process, consider a simple experiment in which the subject squeezes her hand every 20 s, while fMRI data are recorded over 60 time points with a TR of 1 s. You hypothesize that voxels associated with motor processing should show three distinct hemodynamic responses, one following each of the three hand squeezes. The design matrix would thus contain a single column with 60 values. The highest values would occur at around 4-6 s after each hand squeeze (i.e., at the peak of the hemodynamic response) and the lowest values would occur immediately before each hand squeeze (i.e., following the cessation of the hemodynamic response).

Your next task is to evaluate how much this hypothetical time course contributed to the real data, compared with variability outside of the model. Since fMRI data consist of many time points, the residual errors for a given voxel must be combined from all time points into a single value. A formula for combining many error values into one summary statistic is known as a \textbf{cost function}. In the general linear model, the standard cost function is the \textbf{least-squares error}, or the sum of all squared residuals (thus introducing greater penalties for very large errors). Choosing the least-squares error as a cost function allows the general linear model to be solved using a small set of matrix operations (see the statistical references cited at the end of the chapter for details). In this example, the parameter matrix consists of a single parameter (\( \beta \)) for each of the \( V \) voxels. Note that the parameter only provides an estimate of the relative signal amplitude evoked by the experimental manipulation (i.e., the size of the response in that voxel). To obtain a statistic, the value of the parameter must be divided by the residual error. Under the null hypothesis, this quantity should follow a statistical distribution called the \( F \) distribution, and so its significance can be evaluated as a function of the available degrees of freedom (which depend on both the number of time points and the number of regressors).

You may have noticed that both the general linear model and a correlation analysis calculate significance based on how well the experimental data fit a prediction. If that prediction contains only one regressor based on the convolution of events with a hemodynamic response, then the correlation analysis and the general linear model will provide similar statistics. The \( t \)-test can
also be incorporated into the general linear model, by using a regressor with only two discrete levels, one for each of two conditions within a blocked design. Even the Fourier transform can be expressed using a general linear model, although the necessary design matrix would be very complex, because it would need to contain a large number of independent frequency components.

The general linear model provides the theoretical framework that underlies most fmri data analysis, regardless of the experimental design. All major fmri statistical packages include routines for data analysis using the general linear model, although the specific implementation routine depends on the package (Table 10.1). However, each shares the same basic set of simple algorithms and assumptions. In short, the general linear model assumes that the raw data can be modeled as the sum of separate factors, each of which may vary independently across voxels, along with additive Gaussian noise that is also independently distributed across voxels. In the following sections, we evaluate the implications of these assumptions for constructing and testing models for fmri data.

**Constructing a design matrix: regressors of interest**

Everything that fmri researchers do to ensure data quality (e.g., pulse sequence selection, training subjects, preprocessing) can be viewed as efforts to minimize values in the error matrix (\( \epsilon \)). Similarly, the primary purpose of fmri experimental design is to facilitate the creation of the best possible design matrix (\( G \)). Why is the design matrix so important? The maps generated by fmri experiments reflect the outcomes of statistical tests of a researcher's experimental hypotheses. If the hypotheses are poorly modeled within the design matrix (i.e., if the model mistakenly assigns a process of interest onto incorrect time points, or fails to include important contributors to the data) then those statistical tests will be underpowered, and the results will be incomplete or even misleading. It is no exaggeration to state that the greatest challenge in fmri data analysis—indeed, in all fmri experimentation—lies in creating the design matrix.

The regressors in the design matrix represent the hypothesized contributors to the fmri time course. In the general linear model, regressors associated with specific hypotheses are known as experimental regressors, of which there are two types. **Covariates** are factors that can take any of a continuous range of values, where the value of the factor represents the amount of some known quantity. **Indicators** are factors that have integral values that indicate a qualitative level. Within fmri design matrices, the most common experimental regressors are covariates that predict hemodynamic responses using the linear model.

---

**TABLE 10.1 Some of the Major Statistical Packages Available for the Analysis of fMRI Data**

<table>
<thead>
<tr>
<th>Package</th>
<th>Availability</th>
<th>Web site</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFNI</td>
<td>Freely available</td>
<td>afni.nimh.nih.gov/afni/</td>
</tr>
<tr>
<td>Brain Voyager</td>
<td>Commercial</td>
<td><a href="http://www.brainvoyager.com">www.brainvoyager.com</a></td>
</tr>
<tr>
<td>FSL</td>
<td>Freely available</td>
<td><a href="http://www.fmrib.ox.ac.uk/fsl/">www.fmrib.ox.ac.uk/fsl/</a></td>
</tr>
<tr>
<td>SPM</td>
<td>Freely available</td>
<td><a href="http://www.fil.ion.ucl.ac.uk/spm/">www.fil.ion.ucl.ac.uk/spm/</a></td>
</tr>
<tr>
<td>VoxBo</td>
<td>Freely available</td>
<td><a href="http://www.voxbo.org">www.voxbo.org</a></td>
</tr>
</tbody>
</table>
Figure 10.10 A design matrix for the general linear model, here shown for a mixed blocked/event-related design. The set of regressors that attempts to explain the experimental data using the general linear model is known as the design matrix. Here, three regressors represent different cognitive processes, each shown as a different column, within each of three runs. The first represents a blocked effect, while the second and third represent event-related effects associated with two different stimulus categories. The value of each regressor at each point in time is scaled using a dark–light color map. Black indicates that the regressor has its smallest values at that time point, while white indicates the greatest values. The three columns at far right reflect constant values included to remove the mean signal change for each run. Note that each run consists of roughly 130 brain images.

collinear regressors Model factors that are highly correlated with one another. The inclusion of collinear regressors reduces the validity of general linear model analyses.
should be independent of, or orthogonal to, all other regressors. Given that fMRI regressors express predicted hemodynamic amplitude over time, for two regressors to be orthogonal, the hemodynamic activation they model must occur with distinct temporal patterns or with unrelated amplitudes. This restriction has important consequences for choosing regressors. If using a blocked design (e.g., alternating 30 s of task with 30 s of baseline), one could choose to include separate task and baseline regressors. Unfortunately, because these two regressors are perfectly negatively correlated with each other (i.e., when the task has maximal value, the baseline value is minimal), they will explain exactly the same variance in the experimental data. Considered another way: one could not, in principle, distinguish between a voxel whose response to task blocks was increasingly negative and a voxel whose response to baseline blocks was increasingly positive (see Box 9.1). Only one such regressor is required, as shown in the left column of Figure 10.10, to account for both sorts of block effects. In short, the more orthogonal your regressors, the better your chances (i.e., the smaller the difference in BOLD signal required) of identifying effects.

The best way to ensure a well-formed design matrix is to run a well-designed experiment. Indeed, the goals of efficient fMRI design (see Box 9.2) are to maximize variability within a regressor and to minimize correlations between regressors. However, some research hypotheses may preclude perfect orthogonality. For example, experiments investigating attentional effects on target processing may require the presentation of two stimuli in rapid succession: a first cue indicating the likely upcoming location of a second, target stimulus. For ideal fMRI experimental design, two stimulus categories should not have any temporal contingency, but the very nature of this task requires that the cue precede the target (potentially, by a short interval). When the experimental design does not adequately separate two regressors, researchers can orthogonalize one regressor with respect to one or more other regressors in the design matrix (Figure 10.11). In doing so, the analysis program will change that regressor so that it no longer correlates with (one or more) other regressors in the model. A researcher interested in target effects might orthogonalize the target regressor with respect to the cue regressor, to identify aspects of fMRI activation unique to the cue stimuli. Orthogonalization should not be thought of as a panacea for poor experimental design, but as a method for clarifying the effects attributable to a specific regressor.

**Figure 10.11 Orthogonalization of a regressor.** (A) The two regressors represent hypothesized effects of an initial attentional cue (blue) and a subsequent target (red). Note that the two regressors are highly correlated over time, which prevents the regression analyses from identifying independent effects associated with each. (B) Orthogonalizing the target regressor with respect to the cue regressor changes the former's shape, leaving only that part of the signal that would reflect unique contributions from the target stimuli. Orthogonalization improves the interpretability of results when model regressors are not independent.
parametric effect A manipulation of some independent variable so that it takes a number of levels, to evoke regular changes (e.g., a linear increase) in the dependent variable.

An increasing proportion of fMRI experiments go beyond simple subtractive logic (i.e., comparing two conditions that differ in one factor) to include parametric effects. As the name implies, a parametric design incorporates multiple levels of some independent variable. Within the fMRI literature, common parameters include task difficulty, monetary reward, or the intensity of a perceptual stimulus. The parameter may be built into the task (e.g., using easy-, medium-, and hard-task blocks) or measured based on the subjects’ behavior (e.g., response time in each trial). There are two ways to model parametric effects in a design matrix. One is to introduce a separate regressor for each level of the parameter (Figure 10.12A). This can work well if there are only a few levels and if the design efficiently separates those levels in time. This approach has the major advantage of allowing an estimation of the activation associated with each level, making it robust to non-linear changes in activation amplitude. However, it may not be appropriate for non-categorical parameters (e.g., response time). A second approach is to use two regressors, one for the main process and one for how that process changes parametrically (Figure 10.12B). In a 2008 study, Knutson and colleagues investigated brain mechanisms that might underlie the “endowment effect,” which describes the tendency to over-value goods that you own. During a key part of each trial, they presented subjects with a potential prize (e.g., a digital camera) with an associated price (e.g., $25). They modeled the pricing phase of the trial, in part, with two regressors: a main-effect regressor (for overall decision-making processes) whose amplitude was constant throughout all trials, and a parametric regressor (for relative price) whose amplitude depended on the price of the goods. In designs like this one, the main effect and parametric regressors should be orthogonal with each other.

Note that in this latter approach, the regressor for the parametric effects is scaled so that the lowest values reflect negative signal changes and the highest values reflect positive signal changes. This may seem counterintuitive, given that we do not necessarily expect deactivation in voxels of interest (e.g., while evaluating prices). Remember, however, that the goal of multiple regression is

Figure 10.12 Creating models for parametric analyses. When an independent variable can take any of several levels along a continuum, researchers frequently set up their design matrix so that they can identify parametric changes on that variable. (A) One approach is to construct separate regressors for each level of the independent variable, here arranged from the lowest (1) to the highest (4) value. Subsequent contrasts between these regressors can reveal the effects of this variable. (B) Alternatively, two regressors can be used: one modeling a constant effect on every trial, and another modeling a variable effect across trials, depending on the parameter value.
to identify unique contributions of each of our regressors to the observed data, and thus regressors should be uncorrelated. If a voxel exhibited minimal activation, or even no activation, to the lowest levels of the parameter but maximal activation to the highest level of the parameter, then this two-regressor model could account for that activation, regardless of any main effect across trials.

Constructing a design matrix: nuisance regressors

In addition to the regressors of interest, the design matrix often includes additional regressors associated with known non-experimental sources of variability. These are known as **nuisance regressors**. Suppose that the MR scanner on which this study is conducted has a known linear drift (i.e., an increase in raw signal intensity over time) during experimental sessions. A savvy researcher could introduce an additional regressor into the design matrix to account for this drift. Or, if the subject’s respiration was measured during the session, the design matrix could include a regressor for artifacts associated with breathing. Of course, these factors have nothing to do with the experimental hypotheses, in that the experiment was not designed to test scanner drift or subject respiration. So why are they included in the design matrix?

Nuisance factors serve two related purposes in experimental analyses. First, they can reduce the amount of residual variation included in the error term. If the intensity of a voxel were to drift by a few percentage points through the course of a run, the overall variability in the voxel would be very large, compared with the BOLD effect of interest. But if a linear regressor were added to the design matrix, much of that intensity drift would be assigned to that regressor rather than to the error term, increasing the significance of the results. Second, assigning known variability to nuisance factors improves the validity of the general linear model. The model assumes that residuals are independent and identically distributed as Gaussian noise, which may not be the case if a regular source of variation is excluded from the design matrix. It is therefore critical to include all anticipated changes in the BOLD signal, whether of interest or not. However, the unnecessary inclusion of extra factors is not recommended. Each additional column in the design matrix reduces the number of degrees of freedom available. In the limiting case, one could reproduce perfectly any set of \( n \) time points with a combination of \( n - 1 \) different model factors. Since the significance of any individual factor is evaluated as a function of the number of available degrees of freedom, it is in the researcher’s interest for the number of factors to be as small as possible. In practice, the inclusion of a limited number of nuisance factors makes statistical testing more conservative, due to the reduced number of degrees of freedom, but can improve the validity of the general linear model.

No consensus exists about which nuisance regressors should be included in design matrices. The most commonly added nuisance regressors are head motion parameters, typically six regressors comprising three directions of translation and three axes of rotation. The value of each regressor at each point in time reflects the accumulated movement along that direction or around that axis, typically normalized to a range from \(-1\) to \(+1\). Within the design matrix, other regressors are then orthogonalized with respect to these parameters, so that variation in the data attributable to motion (e.g., an increase in activation along the edge of the brain each time a stimulus was presented) will be assigned to the nuisance, not task, regressors (Figure 10.13). Based on a 2006 comparison of analyses with and without motion parameters, Johnstone and colleagues...
Figure 10.13 A design matrix including nuisance regressors to account for head motion. By including motion parameters as nuisance regressors, design matrices can account for an increased proportion of variance in the experimental data. Usually, researchers include six regressors corresponding to three directions of translation and three axes of rotation; the value at each time point indicates the actual net translation or rotation. Shown here is a hypothetical event-related design with five regressors of interest and six nuisance (motion) regressors. For clarity here, the red lines indicate the regressor values at each point in time, with more positive values going to the right.

concluded that including motion parameters in the design matrix generally improved the detection of real activation, especially for analyses of event-related designs. But, because blocked conditions may evoke regular task-related movement, including motion parameters greatly reduced the sensitivity of the analyses. Other researchers argue against including motion regressors in the design matrix, advocating instead for the removal of motion effects during preprocessing.

Some researchers may include regressors for physiological parameters such as heart rate or respiration, if those are recorded during the experiment. Because of their oscillatory nature, these physiological changes can cause regular fluctuations in the BOLD signal, potentially obscuring task-related activation. Recent work was done by Lund and colleagues and by Birn and colleagues to investigate the effects of including physiological data as nuisance regressors. Note that because of the relatively coarse temporal resolution of most fMRI studies (e.g., with typical TRs of 1–2 s), rapid physiological changes like heart rate will be under-sampled. To account for this problem, both groups modeled each type of physiological noise with a set of sine and cosine regressors corresponding to the major frequencies in the physiological data, rather than with a single regressor. (For additional details, see also the work of Glover and colleagues cited in the Chapter References.) Both groups reported that the inclusion of the nuisance regressors improved the detection of true activation, in multiple independent simulations and experiments.

Analysis programs differ in how they combine data from different runs within an experimental session. Some programs treat each run separately, conducting basic contrasts within each run as a first-level analysis, and then combining the data from multiple runs (and subjects) in higher-level analyses. Other programs combine all runs for each subject into the same design matrix. If the runs are combined, then the design matrix should include, for each run, a nuisance regressor with a constant value, as shown in the right-most columns of Figure 10.10. These regressors capture variance associated with differences in mean signal intensity.
across runs. Alternatively, all runs can be normalized to the same mean signal intensity during preprocessing, obviating the need for these regressors.

**Modeling neuronal activity**

So far, we have emphasized that the regressors of interest in an fMRI design matrix represent predictions about hemodynamic changes in the brain, usually associated with the experimental manipulations. As a straightforward way of creating these regressors, a researcher can identify the onset of each stimulus in the experiment, and then convolve those onset times with a canonical hemodynamic response. Yet, this simple approach can lead to misleading results. Consider a simple experiment in which subjects see a picture of an object, and then must retrieve a specific, elaborated memory of a past event associated with that object. The pictures are presented for 2 s each with interstimulus intervals ranging between 20 and 30 s. After creating regressors using the simple stimulus-convolution approach, the researcher is surprised that visual cortical areas are active but memory-related areas are not. What could cause such a result? Imagine that you are the subject in this experiment, and you have been instructed to remember a detailed episode from your past, based on the picture on the screen. The first picture is a balloon. Now, recall a particular event in your life associated with balloons. If you are like most people, it took you between 5 and 15 s (or even longer, depending on the complexity of the memory) to recall and re-experience the particular event. Neurons associated with the retrieval of that memory would have a correspondingly extended period of activity, leading to hemodynamic changes that could span 20 s or more. To detect memory-related activation, the design matrix must accurately model the duration of neuronal processing, not just stimulus presentation (Figure 10.14).

This example illustrates the importance of thinking about the regressors in a design matrix as predictions of the BOLD time course, evoked by hypothesized neuronal activity. Before creating your design matrix for an experiment, you should think carefully about the separate types of brain processes that are evoked in your experiment, along with their timing and duration. In tasks
where several processes are likely to be evoked sequentially, researchers often identify different phases of the task. These phases may be explicit, such as when a visually presented word must be remembered over a delay interval, followed by a cue that requires a decision about what word was presented. The resulting design matrix could have a separate regressor for each process. Or the phases may be implicit, established by the experimenter based on an expectation of what the subjects will do in the task. If the task is to remember a complex display of shapes presented for 10 s, the researcher might distinguish between two phases: a beginning encoding phase (2 s), when the subjects study the display; and a rehearsal phase (8 s), during which time the subjects commit particular aspects of the display to memory. In summary, no fMRI statistical analysis can be valid when the statistical test, here determined by the predictions of BOLD activation in the design matrix, does not reflect the actual changes in the brain associated with the experimental manipulations.

Modeling hemodynamic convolution

Most researchers create design matrices by convolving predicted neuronal activity with a standard hemodynamic response, as provided by their fMRI analysis packages. The hemodynamic response takes the stereotyped form introduced in Chapter 7 (see Figure 7.11), with a rise to a peak around 5 s after stimulus onset, followed by a return to baseline and subsequent undershoot at around 12-15 s. (Note that while analysis packages differ slightly in their default hemodynamic response, all share the same essential features.) The use of a single standard hemodynamic response greatly simplifies analyses. However, it can also constrain statistical models, particularly if its timing or shape differs from that observed during an experiment. Researchers have therefore explored a number of approaches for improving the generalizability of their design matrices.

All analysis packages allow the researcher to select from a broad range of hemodynamic response functions. But, what function should you choose? Most common are mathematical distributions that can be described with a small set of parameters (Figure 10.15A). Gamma and Poisson functions can roughly match the rise and fall of the hemodynamic response if appropriate parameters are selected. Some analysis packages use a mixture of two Gamma functions: a faster, larger function to model the initial rise and fall, and a slower, smaller function to model the subsequent undershoot. As introduced in Chapter 8, subject-specific hemodynamic response functions can be derived from a simple experiment (e.g., a finger-tapping task to evoke activation in the motor cortex) and then applied throughout the brain in subsequent analyses.

Regardless of the functional form chosen, the regressors in the model often differ slightly from the observed data. Such differences could arise from imperfect estimation of the timing of neuronal activity or from variability in the evoked hemodynamic response. To model small differences in hemodynamic onset, or in the shape of the hemodynamic response, the design matrix can include additional regressors known as temporal derivatives and dispersion derivatives. When one adds a temporal derivative to a signal, the combination will be a time-shifted version of the original signal. Similarly, including temporal derivatives in a fMRI design matrix makes the analysis robust to small mismatches between the timing of the regressors and the observed BOLD signal. The dispersion derivative can correct for small mismatches in the width of the hemodynamic response. Note that when temporal and/or dispersion derivatives are included, they are generally calculated independently for each

**temporal derivative** A regressor that, when added to a model, improves the robustness of that model to small variations in the timing of the hemodynamic response.

**dispersion derivative** A regressor that, when added to a model, improves the robustness of that model to small variations in the width of the hemodynamic response.
regressor of interest. This reduces the degrees of freedom, and thus increases the threshold for significance, in the regression model. Researchers differ in how they use these derivatives when calculating the parameter amplitudes (i.e., calculating $\beta$ from just the regressor of interest, or from the combination of that regressor and its derivatives). Some laboratories regularly incorporate these derivatives into their analyses, while others do not.

An increasingly popular approach is to replace a single hemodynamic response function with a small number of basis functions, such as low-frequency sine and cosine waveforms or gamma functions. Since a wide range of hemodynamic responses can be modeled using a combination of multiple basis functions, this approach can be used to detect voxels whose time courses of activation are not standard, such as those with wider responses or with later peaks. Particularly promising are techniques that use finite impulse response (FIR) functions, which model each time unit with a separate basis function (Figure 10.15B). In principle, FIR modeling can be used to identify task-related changes in the BOLD signal regardless of the shape and timing of the evoked hemodynamic response. A researcher who lacks a strong hypothesis about the likely hemodynamic response might be well-served by using FIR or another set of basis functions, rather than a canonical response function. However, using basis functions increases the complexity of the design matrix, which in turn increases the challenge for detecting and interpreting significant results.

Finally, a frequent source of confusion when thinking about the use of the general linear model in fMRI is the idea of linearity itself. Remember from Chapter 8 that the BOLD response does not obey the assumptions of linearity

**Figure 10.15** Possible functions underlying the BOLD hemodynamic response. (A) The hemodynamic response evoked by a single, short-duration event can be approximated using any of several functions, including simple Gaussian and Gamma functions. The combined effects of the hemodynamic rise, fall, and undershoot can be modeled using two Gamma functions, one subtracted from the other (red line). (B) Another approach is to use a set of basis functions, or a series of separate regressors (shown here as overlapping waveforms) that can be added together to obtain the measured hemodynamic response. A major advantage of using basis functions is that they can more flexibly model almost any evoked response, even if it differs from the canonical hemodynamic shape.

**basis functions** A set of functions whose linear combination can take on a wide range of functional forms. In fMRI analyses, researchers often replace a single hemodynamic response function with basis functions, to improve the flexibility of their design matrices.

**finite impulse response (FIR)** A signal processing approach that treats each time unit with a separate function (i.e., an impulse); this has the major advantage of making no assumptions about the shape of the observed response function.
at short interstimulus intervals. So, how can it be analyzed using the general linear model? This confusion derives from a misunderstanding of the type of linearity under consideration. The hemodynamic response is nonlinear with respect to stimulus presentation, because the combined response to two stimuli in succession is less than the sum of the responses to the two stimuli presented independently. But, stimulus presentation is not itself important in the general linear model. What is important is that the overall BOLD time course, however it is modeled, adds linearly with other sources of variability in the data. So, refractory effects in the BOLD response can be incorporated directly into the experimental design matrix without compromising the validity of the model. One way to do this is to adjust the columns of the design matrix directly by including interaction effects. Another method, identified by Friston and colleagues in 1998, is to use Volterra kernels, which allow one to model the influence of a stimulus on subsequent stimuli, by specifying a second column in the design matrix. Without such corrections, the design matrix may not accurately capture the desired BOLD time course.

**Contrasts**

When analyzing fMRI data, researchers want to evaluate hypotheses about brain function. Because fMRI provides no information about absolute levels of activation, only about changes in activation over time, most research hypotheses involve comparison of activation between two conditions. For example, when identifying brain regions that support the perception of biological motion, the experiment might involve testing whether activation increases when subjects view a biological stimulus (e.g., a person walking) compared with a similar non-biological stimulus (e.g., a machine with moving levers and gears). As introduced in the discussion of $t$-tests, the statistical evaluation of whether the experimental manipulation evokes a significant change in activation is called a contrast.

Using the general linear model, an fMRI researcher constructs a design matrix consisting of a set of regressors, and then determines how strongly each of those regressors matches changes in the measured BOLD signal. Regressors that explain much of the BOLD signal have high-magnitude parameter weights (i.e., large $\beta$ values), whereas regressors that explain little of the BOLD signal have parameter weights near zero. To test an experimental hypothesis, the researcher evaluates whether the experimental manipulation caused a significant change in those parameter weights. The form of the hypothesis determines the form of the contrast, or which parameter weights contribute to the test statistic. We consider three types of hypotheses in the following examples.

The simplest type of contrast evaluates whether a single regressor causes significant changes in the BOLD signal. Suppose that your experiment uses an event-related design with the randomized presentation of three types of visual stimulus: biological motion (e.g., a person walking), non-biological motion (e.g., a machine moving), and perceptual control (e.g., a rotating shape). Each of these stimuli is presented as a short movie of 2 s duration with 8 s intervals between consecutive stimuli. You set up each of these as a separate regressor in your design matrix (Figure 10.16A), using the approaches described in the previous sections. If you want to identify voxels whose activation increased in response to the biological motion stimulus, you would use the set of **contrast weights** $(+1\,0\,0)$, as shown on the top row in Figure 10.16B. Conversely, to identify voxels whose activation decreased in response to biological motion, you would use the contrast weights $(−1\,0\,0)$. Effectively, these contrasts use the
Figure 10.16 Setting up experimental contrasts. (A) In this hypothetical example, the experimental task consisted of three types of events that occurred in a random order, with a relatively long interstimulus interval, over the course of a single run. The hypothesized BOLD timecourse associated with each event type has been modeled in this design matrix by convolving a standard hemodynamic response function with brief neuronal activity at the onset of each event. (B) The positive main effects associated with each event type (i.e., what voxels increased in activation to those events?) can be specified by a positive value for that event regressor, leaving the other regressors at zero. Negative main effects (i.e., decreases in activation) can be specified by a negative value for that regressor. An F-test can be used to combine across multiple main effects (i.e., those rows included as "Y" here), identifying voxels that showed significant changes in activation across any of the conditions. (C) More critical for most fMRI studies are contrasts between conditions. These involve comparing the relative amplitude of activation between two or more regressors; the example here shows both a direct contrast between two conditions (+1 -1 0) and a contrast between one condition and the other two (+2 -1 -1).

The parameter weight from the biological motion condition, but ignore the other conditions. The statistical software then multiplies the parameter weights by your chosen contrast weights, scales the resulting quantity by the residual error, and then evaluates (usually with an F-test) that scaled value against a null hypothesis of zero. Note that these single-condition contrasts, often referred to as main effects of a condition, can lack experimental control. Significant increases in one condition could arise from any of a host of factors, from visual stimulation to arousal. Thus, researchers often compare two or more regressors, following the subtractive logic described in the previous chapter.

If you want to test the (better controlled) hypothesis that biological motion evokes increased activation compared with non-biological motion, you would use the contrast weights (+1 -1 0), as shown in Figure 10.16C. This contrast provides a value for the difference between the biological motion and non-biological motion parameters, and then uses that value in the following statistical tests. Note that you could use other sets of contrast weights, depending on your hypothesis. To evaluate whether biological motion evokes greater activation than both other forms of motion, you would use the contrast weights (+2 -1 -1). Although statistical tests can incorporate any combination of weights (e.g., +1 -4 +3), researchers generally adopt relatively simple contrasts that reflect well-defined hypotheses. Note that contrasts between conditions generally use weights that sum to zero, reflecting the null hypothesis that the experimental manipulation had no effect. Also, contrasts are inherently directional, in that the contrast weights (+1 -1 0) are used to identify voxels with greater activation in response to the first stimulus type than to the second, whereas (-1 +1 0) is used to identify voxels with greater activation in response to the second stimulus type than the first. Researchers often include contrasts in both directions when testing hypotheses of particular interest.

Finally, research hypotheses sometimes involve combining over several different contrasts. Suppose you want to identify voxels that exhibit significant increases in activation in response to any of your three motion conditions. After setting up three of the single-condition contrasts described above, you could next enter all three contrasts into a single F-test (see Figure 10.16). The F-test evaluates whether any contrast or any combination of contrasts explains a significant amount of the variability in the measured data. Unlike t-tests, F-tests are non-directional; they do not indicate the direction of any of the contrasts. Nor do they provide information about which contrasts drive significance, only that

F-test A statistical test that evaluates differences among a set of distributions. For fMRI studies, F-tests can evaluate whether any of a set of contrasts exhibited a significant effect.
**homoscedastic** Having the property that the distributions of noise are similar for all experimental conditions.

**heteroscedastic** Having the property that the distributions of noise are different across experimental conditions.

A significant difference exists among the conditions. Researchers most commonly use *F*-tests to identify voxels that show some modulation in response to the experimental task, in advance of more targeted contrasts.

**Assumptions of the general linear model**

Provided that the design matrix is appropriate for testing the experimental hypotheses, several other conditions must be met for the general linear model to be appropriate. One assumption that has been the source of considerable debate is the use of the same design matrix throughout the brain. Although each voxel will have a different set of parameter weights, the model used to calculate those weights is identical. But we know that the properties of the hemodynamic response, especially its latency, may differ across brain regions. A model factor that is correct for one region may thus be incorrect for another, reducing the amount of variation explained by the model and increasing the residual error. The use of multiple basis functions provides some flexibility compared with using a single canonical hemodynamic response, but complicates the interpretation of the results. One way of overcoming the problem of regional variability is to combine a general linear model approach with region-of-interest analyses, as discussed later in this chapter.

Another assumption is that noise varies with a normal distribution that has similar properties at all time points. In other words, the contributions from noise do not vary over time, and thus are independent of the experimental task. That is, the data should be **homoscedastic**. If there is greater noise in one condition than another, the data are **heteroscedastic**. The general linear model assumes the former, however this assumption may not always be valid. Noise levels may be higher during BOLD activation than during rest, although whether such changes are due to hemodynamic variability or to variability in neuronal processing remains unknown. However, the contributions of noise may differ dramatically across voxels. A voxel that contains a major blood vessel or that is near the edge of the brain may have much higher noise levels than most others in the brain.

Another assumption is that all voxels represent independent statistical tests, even though adjacent voxels tend to have very similar properties. In fact, introducing correlation between adjacent voxels through spatial smoothing is a common step during preprocessing. While the general linear model framework cannot account for spatial correlation, the significance values it generates can be adjusted at later stages of the analyses, as discussed later in this chapter, in the section on cluster-thresholding. Likewise, the model assumes that each time point is independent of all others, in that the residuals should be similarly distributed throughout. Scanner drift, thermal variation, head motion, and many other factors can cause the overall MR signal to change dramatically over time, influencing the amplitude of the residual error. Therefore, it is critical to attempt to remove such unwanted variability before it reaches the error term, either during preprocessing or by including appropriate nuisance factors.

In summary, the general linear model has become the dominant statistical framework for fMRI analyses. It subsumes other tests, like the *t*-test and correlation analyses, that represent special cases of this framework with simplified assumptions. The power of the general linear model comes from its flexibility. By incorporating into the design matrix appropriate regressors of interest, and excluding unwanted variability through nuisance regressors, a researcher may test nearly any experimental hypothesis. However, researchers must carefully construct their design matrices and select their contrasts, or their analysis may result in incorrect conclusions.
Corrections for Multiple Comparisons

A typical fMRI data set may contain about 20,000 voxels within the brain and several times that number outside of the brain. Imagine that you first target a single voxel within the gray matter adjacent to the right calcarine sulcus. Using data from a simple event-related visual task, you calculate a t-statistic of 2.4 based on the voxel's correlation with a task-related regressor in the general linear model. The chance that such an extreme t-statistic could occur under the null hypothesis, based on the degrees of freedom in the test, is only about 1 in 50 (P = 0.02), less than your alpha value of 0.05. Given such a low probability, you confidently reject the null hypothesis for that voxel. Flush with the excitement of a significant result, you decide to analyze the remaining voxels.

You run all of the voxels through the correlation test, calculate a t-value for each, and compare those t-values with your alpha value. Now you find that about a thousand voxels, distributed in seemingly random fashion across the brain, pass your significance threshold (Figure 10.17). Even worse, a few thousand voxels outside the brain appear to be active! You stare at the computer screen in disbelief. Why have so many voxels produced significant results?

This example illustrates one of the central problems of fMRI data analysis, that of multiple comparisons. Stated succinctly, the greater the number of statistical tests conducted, the greater the chance of a false-positive result. To illustrate this point, we created a random data set of roughly similar size to an fMRI imaging volume (64 × 64 × 34 = 140,000 voxels), where the intensity of each voxel at each of 40 time points was distributed as Gaussian noise. We then conducted a t-test on a hypothetical blocked design, comparing an arbitrarily labeled task block to another arbitrarily labeled non-task block (20 time points each). Both conditions consisted of random data, so any difference between them was due solely to chance. At an alpha value of 0.05, there were about 6800 active voxels; at an alpha of 0.01, there were about 11,400 active voxels; and at an alpha of 0.001, there were still 155 active voxels. All of these voxels are false positives, since there was no signal present in the original data. In any data set with random noise, the number of false-positive results for n statistical tests is simply n × α. The probability of having no false-positive results is given by Equation 10.4:

\[ p(\text{no Type I error}) = (1 - \alpha)^n \]  

(10.4)

Note that for the alpha values that are typically used in social science experiments (e.g., 0.05, 0.01), this probability approaches zero for even small fMRI data sets. If you analyzed only a single slice of 4096 voxels at a < 0.01, the probability of finding no Type I errors is (1 - 0.01)^6800, which is virtually zero. If you cared only about the brain and not the thousands of other voxels represented in the slice, you could set your significance threshold high enough (e.g., 0.05) to avoid Type I errors. But is that the right approach? No! If you increase your threshold, you make it impossible to obtain a significant result, and you are left with no data.

**Figure 10.17** The problem of multiple comparisons. We simulated the effects of different alpha values (α = 0.05, 0.01, and 0.001) on the number of activated voxels, using random data. The imaging volume consisted of 34 slices, each with matrix size of 64 by 64, resulting in a total of about 140,000 voxels. Note that since the data were random, any activation was due merely to chance. (A) At an alpha threshold of 0.05, there were more than 6800 active voxels, representing 4.9% of the total brain. (B) When the threshold was reduced to 0.01, there were 1397 active voxels (1.0% of the brain). (C) At a 0.001 threshold, there were still 155 active voxels (0.1%) throughout the brain. Shown in each panel are single slices, with only those voxels passing the appropriate significance threshold highlighted in color.
familywise error rate (FWER) The probability of making at least one Type I error, given the total number of statistical tests.

Bonferroni correction A stringent form of familywise error rate correction for multiple comparisons in which the alpha value is decreased proportionally to the number of independent statistical tests.

The probability of having no false positives is $1.3 \times 10^{-18}$, or one quintillion to one. Stated differently, without correction for multiple comparisons, you are certain to make at least one Type I error of labeling a voxel as active when it is not.

Calculating the significance threshold

To overcome the problem of multiple comparisons, fMRI researchers always reduce the desired alpha value, so that voxels are less likely to pass the significance threshold by chance. The alpha value they select depends primarily on two factors: the types of error they want to avoid and the number of independent tests in the data. In this section, we focus on the first factor. Our examples make the simplifying assumption that the number of independent tests corresponds to the number of voxels. However, the spatial correlations within real fMRI data reduce the true number of independent statistical tests. We describe, in the following section, approaches for estimating the number of independent tests, replacing the idea of voxel-based thresholding with that of cluster-based thresholding.

The most common approach for the correction of multiple comparisons involves minimizing the number of false positive results (i.e., Type I errors), or controlling for familywise error rate (FWER). A stringent method for controlling the FWER is Bonferroni correction, which holds constant the overall probability of a false positive, given the number of statistical tests conducted. To implement Bonferroni correction, the alpha value is decreased proportionally to the number of independent statistical tests (here, $V$ voxels), as seen in Equation 10.5:

$$\alpha_{\text{bon}} = \frac{\alpha}{V}$$

Suppose that you decide on a target alpha value of 0.01. Even if your imaging volume included only 4096 voxels (i.e., a single slice), applying Bonferroni correction would reduce the alpha value from 0.01 to 0.000002. You would now have only a 1% chance of any Type I error. But, while Bonferroni correction effectively minimizes the chances of a Type I error, it also dramatically increases the probability of a Type II error, or failing to detect voxels with real activation (Figure 10.18). For many research questions, especially those that are exploratory or have clinical relevance, an increased rate of Type II errors may be unacceptable. Imagine that you have conducted an fMRI study to identify the cortex necessary for language processing in a patient about to undergo neuro-
surgery. A very conservative threshold might result in the identification of some active voxels, but there is a risk that other truly active voxels with lower significance values will be missed. In such a situation, Type II errors could lead to the removal of functional tissue and thus would have serious negative consequences for the patient.

An alternative and increasingly popular approach controls a different quantity, the false discovery rate (FDR). As its name implies, the false discovery rate describes the proportion of positive results (i.e., discoveries) that are actually false positives. For example, suppose that you choose a voxelwise false discovery rate of 0.05, calculate the necessary alpha value, and then observe 200 significant voxels using a given set of contrast weights. Based on your FDR criterion, you should expect that about 10 of those voxels are false positives. Now, suppose that you observed 400 active voxels using a different set of contrast weights within the same experiment. About 20 of those voxels would be expected to be false positives. In comparison, FWER correction would predict the same number of false positive results, regardless of the number of significant voxels. Because the alpha value used in FDR correction depends on the distribution of significance values, it is calculated using an iterated approach, as introduced to fMRI in 2002 by Genovese and colleagues.

First, an algorithm ranks the statistical tests (e.g., V voxels) according to their uncorrected probability of significance, from smallest probability ($p_i$; most likely to be significant) to greatest probability ($p_V$; least likely to be significant), as shown in Equation 10.6.

$$p_1 \leq p_2 \leq \cdots \leq p_V$$

(10.6)

Then, beginning with the voxel with the smallest probability ($p_1$), the algorithm identifies the voxel whose probability index ($p_i$) is greater than the voxel's ranking in the list ($i$) divided by the total number of tests ($V$), as corrected by the desired FDR ($q$). This is shown in Equation 10.7. All voxels ranked from $p_1$ to $p_i$ are labeled as significant.

$$p_i \leq \frac{iq}{V}$$

(10.7)

Controlling the FDR rather than the FWER provides two primary advantages for fMRI research. First, FDR uses less-stringent correction (i.e., a greater alpha value) than FWER, especially if there are many activated voxels. If there were only one activated voxel ($i = 1$), then that voxel would have to pass full Bonferroni correction. But, as more and more voxels become activated, then the ranking term ($i$) increases in Equation 10.7, driving down the threshold for significance. Many researchers find the gain in experimental power to be worth the additional chance of Type I errors. Second, whereas FWER correction controls the proportion of false tests, FDR controls the proportion of false claims. In fMRI studies, which typically have data sets with many tests, each with a low probability of significance, researchers may care more about the rate of errors among their claims (e.g., activation clusters in a table) than among their statistical tests (e.g., voxels).

If researchers are testing hypotheses about targeted anatomical regions, they may only examine data from a small portion of the brain (e.g., the hippocampus), thus restricting their analyses to perhaps a few hundred or a few thousand voxels. Such small-volume correction leads to a much less severe correction factor (i.e., a smaller denominator in Equations 10.5 or 10.7) than full-brain correction. Small-volume correction should only be done before analyses, based on strong a priori hypotheses. Furthermore, because it ignores most voxels in the brain, it
can readily to lead to the false impressions that activation is specific to a particular brain region, rather than distributed throughout the brain.

Finally, permutation approaches estimate the distribution of significance values that would be obtained by chance. These approaches permute (i.e., randomize the order of) the condition labels associated with the events within the design matrix. This approach falls into the class of "resampling" methods in the statistical literature. Suppose that your experiment consisted of 100 presentations of real words and 100 presentations of pseudo-words (e.g., "drelp"), in some random sequence. A permutation test takes those 200 events and randomly reassigned them to conditions, so half of the real words get included in the pseudo-word regressor, and vice versa. Because these regressors no longer systematically differ in their associated neural processes, any differences between them should be due to chance. The analyses are repeated many times, following each of a large number of permutations (e.g., > 100), recording the maximum significance value observed in any voxel (or any cluster). The distribution of maximum significance values then guides separation of the alpha value for the real analysis, such that, through the choice of a sufficiently high threshold, a false positive is unlikely to occur due to chance. Though computationally intensive, permutation approaches have the advantage of estimating false positives from the data set itself, and thus they can account for idiosyncratic features of particular data sets. However, some characteristics of fMRI data sets, notably their temporal autocorrelation, could violate assumptions of independent resampling in many designs (i.e., the exchangeability of labels). For such cases, researchers have proposed techniques for preprocessing the data to minimize effects of temporal autocorrelation. For details about permutation approaches, including their strengths and limitations, see the articles by Nichols and Hayasaka in the Chapter References.

**Thresholding based on clusters of activation**

Another approach to the correction of multiple comparisons is to use information about the sizes of any active clusters. If only a single isolated voxel passed a significance threshold, then that voxel's activation may result from mere chance. It is less likely, however, that a group of contiguous voxels will all be active by chance. This can be seen by careful examination of the data from Figure 10.17A. While many voxels in this figure are active due to chance, very few clusters of two or more adjacent voxels are present. Using cluster-size thresholding, which was first introduced in separate 1995 studies by Xiong and colleagues and Forman and colleagues, a researcher adopts a relatively liberal alpha value (e.g., $P < 0.01$) for voxelwise comparisons, and then increases the conservatism of the test by only counting clusters as significant if they are as large as some threshold. The cluster size to use as the threshold in a given experiment depends on several factors, including the desired alpha value and the number of independent tests in the imaging volume, and software analysis packages often suggest appropriate thresholds for a given data set.

Cluster-size thresholding works because as cluster size $C$ increases, the number of such clusters $(n_c)$ increases much more slowly than the probability that a given cluster is active. In a single slice of 4096 voxels ($64 \times 64$), there are approximately 16,000 distinct clusters of two contiguous voxels and approximately 55,000 clusters of three contiguous voxels. (In three-dimensional data, several times as many clusters are present for each cluster size.) If the alpha value for the cluster ($\alpha_c$) is set to 0.001, the expected number of false-positive voxels is $4096 \times 0.001$, or about 4. For the same alpha value, the joint probabilit-
ity of two given voxels being active is just $0.001 \times 0.001$, which comes out to one in one million. Thus, the expected number of false-positive clusters of two contiguous voxels is $16,000 \times 0.000001$, or 0.016. Finally, the joint probability of any three voxels being active is $0.001 \times 0.001 \times 0.001$, which comes out to one in one billion. So, the expected number of false-positive clusters of three contiguous voxels is about $55,000 \times 0.000000001$, resulting in an expectation of 0.000055 false-positive clusters.

In summary, the likelihood of a false-positive result decreases with increasing cluster size. The effects of cluster-size thresholding on the false-positive rate are indicated in Equation 10.8 (compare to Equation 10.4), where $n_c$ is the number of clusters in the data:

$$p = (\text{no clusterwise Type I error}) = (1 - \alpha_c)^{n_c} \quad (10.8)$$

By reducing the alpha value used in an experiment, cluster-size thresholding will often reduce the number of Type II errors, or misses of true activation. However, it makes several assumptions that, if violated, introduce potentially severe disadvantages. First, by definition, thresholding assumes that all areas of significant activation extend over a large number of voxels. This precludes small but meaningful activations. If your cluster-size threshold is six voxels, then detecting an active brain region of only four voxels in size becomes extremely unlikely. Some statistical programs use both the number of above-threshold voxels (as in the examples above) and the significance values of those voxels, to make better inferences about small but significant clusters of activation. Second, it assumes that activation foci are generally convex, or spherical, when calculating probabilities. If an active region has a very non-spherical shape, as when running linearly along the edge of a gyrus, then cluster size analyses may not be appropriate. Third, and most importantly, the above logic assumes that activation in adjacent voxels is uncorrelated, so that the probability of $n$ voxels all being active is given by $\alpha^n$. This assumption fails for fMRI data because of spatial correlation; significant voxels tend to cluster together, even if their significance results from noise processes. In the following section, we consider approaches that estimate this spatial correlation to improve calculation of correction factors.

**Estimating the number of independent tests**

When correcting for multiple comparisons, whether using FWER or FDR or some other approach, the adjusted alpha value derives from the number of independent statistical tests. For fMRI studies, this number could come from the number of voxels in the brain; if there are 20,000 voxels, there might be 20,000 tests. Yet, for many reasons, significance values tend to be highly correlated across adjacent voxels (see Figure 11.16 for an example). Inherent limitations in MRI data collection and image reconstruction allow signal to bleed across adjacent voxels (and even slices). Many sources of noise, notably head motion, systematically change the intensity of all voxels within a brain region. Activation itself often spans large regions, especially when generated by large vessels. Preprocessing steps, such as head motion correction and spatial normalization, introduce uncertainty regarding the contents of a voxel. And, in addition to all of these implicit factors, researchers usually apply explicit spatial smoothing before data analysis! Given these many sources of inter-voxel dependence in fMRI data, correction based on the number of voxels greatly overestimates the number of independent spatial units, resulting in a much too conservative alpha value. To determine a more accurate correction factor,
random field theory A branch of mathematics that deals with the properties of smooth, spatially extended data. Using random field theory, researchers can better calculate the number of independent tests within fMRI analyses.

smoothness The degree to which the time courses of nearby voxels are temporally correlated.

resolution elements (or resels) The independent statistical tests within an fMRI volume.

expected Euler characteristic The number of clusters of significant activation expected due to chance, as estimated from the number of independent statistical tests (i.e., resels).

voxelwise analyses Evaluations of hypotheses about the functional properties of individual voxels (or small clusters of voxels), often throughout the entire brain.

techniques have been proposed that modify the denominator in formulae like Equation 10.5, based on the degree of correlation between activated voxels. To determine a better correction factor, Worsley and colleagues applied (Gaussian) random field theory to fMRI data. Random field theory is used to estimate the number of independent statistical tests needed, based on the spatial correlation, or smoothness, of the experimental data (see Figure 8.28). Although smoothness depends heavily on the properties of the Gaussian filter used in preprocessing, intrinsic correlations also matter, and thus smoothness is typically estimated computationally by the statistical program used for the analysis. Based on the smoothness, which can be expressed in voxels or millimeters, the number of independent tests in a data set can be calculated. If a data set consisting of $x$ by $y$ by $z$ voxels had smoothness of $V$ voxels, the number of independent comparisons ($R$) would be given by:

$$R = \frac{x \times y \times z}{V^3}$$

(10.9)

The independent comparisons are sometimes known as resolution elements or resels. With even small to moderate amounts of smoothness in the data, the number of resels will be much less than the original number of voxels. At a smoothness of three voxels, there would be 1/27 as many resels as voxels. From the number of resels (with minor adjustments based on the shape of the brain volume), one can estimate how many clusters of activation should be found by chance at a given statistical threshold. The number of such clusters is known as the expected Euler characteristic. Note that for smooth random fields, like preprocessed fMRI data, the effect of threshold has a complex effect on the expected Euler characteristic. At very low statistical thresholds, only slightly above chance, there will be very few clusters. However, these clusters will be very large and interconnected since much of the brain will be labeled as active. At medium thresholds (i.e., at thresholds of about $P = 0.15$ for typical studies), there will be a very large number of smaller clusters labeled as active merely by chance. But as the threshold increases, the number of small clusters identified by chance should decrease. Here we have discussed smoothness and the resulting calculations of expected clusters of activation as something constant over the entire brain, however smoothness could differ from region to region. Recent work by Hayasaka and colleagues demonstrates that accounting for such regional differences can improve detection power.

Using random field theory, fMRI statistical software determines the statistical threshold whose expected Euler characteristic corresponds to the desired alpha value (e.g., 0.05). That is, given the number of resels in the data, at what threshold would there be only a 0.05 probability of an expected Euler characteristic of 1 (or greater), were there no true activation? For smoothed data, this threshold will always be much less conservative than that obtained through Bonferroni correction, leading to fewer Type II errors, and only a minimal risk of additional false positives. Nearly all fMRI research now uses some variant of random field theory to determine thresholds for statistical significance.

**Region-of-Interest Analyses**

Most fMRI studies involve statistical tests on individual voxels or clusters of voxels, often throughout an imaging volume that encompasses the entire brain. Such voxelwise analyses are appropriate for a wide range of research hypothe-
ses, especially those aimed at understanding particular cognitive processes. Yet some hypotheses require a more targeted analysis approach. If you are interested in a particular brain region, you may form your hypothesis about that region, rather than the entire brain. For example, you might ask, “Is the caudate nucleus active during recall of a word from memory?” In this example, the caudate nucleus becomes a region of interest, whose identity is based on anatomical criteria (Figure 10.19). In a deep sense, the basic question addressed by a voxelwise analysis, “What brain regions evince a particular pattern of fMRI activation?” is the inverse of the question posed by a region-of-interest (ROI) analysis, namely “What pattern of activation occurs in a particular brain region?”

For most studies, the establishment of a ROI is based on a priori expectations about the likely involvement of that brain region in a task. For example, a researcher studying motor function might draw an ROI that encompasses the precentral gyrus, which contains the primary motor cortex. The ROI is considered to be a homogenous and indivisible unit, at least for the purposes of the ROI analysis. Usually, ROIs are drawn on structural T_{1} or T_{2} images collected at the beginning of a scanner session. The structural images are used for two reasons: they typically have higher resolution, often four times that of the functional images in each spatial dimension; and they have much greater tissue contrast. The ROIs are then coregistered to the functional data. In most ROI approaches, the researcher draws the edges of a particular brain area, such as the sulci that demarcate a gyrus of interest, and then selects all circumscribed voxels. If anatomical ROIs are chosen beforehand and are drawn without consultation with functional activation maps, then they can provide an unbiased estimate of activation within a given brain area.

Anatomical ROI analyses have several advantages over voxelwise methods. First, because there are always many fewer ROIs than voxels, the total number of statistical comparisons is greatly reduced, ameliorating the need for correction for multiple comparisons. For example, in a study of the motor cortex, a researcher might draw just two ROIs, to encompass the primary motor cortex in each hemisphere. Second, each ROI combines data from many voxels, so there will be a corresponding increase in the SNR, to the extent that the ROI is functionally homogenous. This spatial signal averaging complements the temporal signal averaging common to voxelwise and ROI analyses. Another advantage is that ROI approaches allow the identification of brain topography, as reflected in changes in activation levels across spatial locations.

**region-of-interest (ROI) analysis**

Evaluations of hypotheses about the functional properties of brain regions (i.e., aggregated over a pre-determined set of voxels), often chosen to reflect a priori anatomical distinctions within the brain.

**anatomical ROI** Region of interest that is chosen based on anatomical criteria.

---

**Figure 10.19** Region-of-interest analysis. Region-of-interest (ROI) analyses divide the brain into sections based on anatomical or functional criteria. (A) Shown on a coronal slice of the brain are locations of the inferior (IFG), middle (MFG), and superior (SFG) frontal gyri, the anterior cingulate gyrus (ACG), and a white-matter control region (WHM). (B) Shown on a coronal slice of the brain are locations of the intraparietal sulcus (IPS) and the fusiform gyrus (FFG). These ROIs were drawn based on anatomical criteria. (From Jha and McCarthy, 2000.)
within the brain. Comparisons between ROIs, whether in different regions of the same slice or in the same region across slices, can be used to create simple and easily understood parametric activation maps. Finally, ROI approaches ameliorate many of the problems of comparing data between subjects. When brain regions are drawn on a subject-specific basis, they match spatial locations across the subject sample, thus eliminating inaccuracies introduced when normalizing an individual’s anatomy to a reference brain.

While anatomical ROI analyses provide important information about the functional properties of particular brain regions, they introduce a new problem: the potential mismatching of anatomical and functional regions of the brain. The idea that anatomically distinct regions of the brain are likely to have different functional properties is not new. An early and influential map of the brain was created by the German physician and neurobiologist Korbinian Brodmann, who divided the human cerebral cortex into nearly 50 anatomically distinct regions, known as Brodmann areas (see Box 6.2, Figure 4). Brodmann’s areas were defined by their cytoarchitecture—differences in the size, types, and distribution of neurons within a brain region—and do not necessarily correspond to specific gyri or sulci. Functional MRI, at least at the spatial resolution typical of human studies, provides no information about cytoarchitectural features of the brain; thus, Brodmann areas cannot be directly determined for a particular MRI subject. However, even with perfect anatomical mapping, there would remain the problem of linking anatomical regions to functional divisions within the brain. For example, a complex task like the visual recognition of objects relies on defined occipitotemporal and occipitoparietal pathways that cross many anatomical regions. This problem can be partially overcome by drawing several ROIs that encompass different anatomical components of a functional network. Conversely, a single anatomical region may contain multiple functional regions. If only a small subdivision of the anatomical ROI is activated by your task, then your functional SNR may be reduced by the inclusion of many inactive voxels. Even finely mapped anatomical ROIs may not reflect true functional divisions within the brain.

Another powerful approach is to create functional ROIs, which include only voxels that were activated by a particular stimulus. For example, in many studies of face processing in the fusiform gyrus, Kanwisher and colleagues used a screening task, also known as a localizer task, to identify voxels that are differentially activated by faces compared with other complex visual stimuli. These face-activated voxels then comprised a functional “face area” ROI that the researchers used to evaluate the effects of other experimental manipulations (e.g., identity changes, selective attention) on face processing. Creating functional ROIs is critical when the boundaries of a (functionally significant) brain region cannot be readily identified by anatomical landmarks (e.g., the frontal pole, the temporoparietal junction). For an extended discussion of the value and limitations of this approach, we refer interested readers to the 2006 articles by Friston and by Kanwisher and their colleagues.

Both practical and theoretical problems prevent the universal use of ROI approaches. Simply put, creating anatomical ROIs can be extremely challenging. There are some automated programs that partition the brain into anatomical regions based on segmentation algorithms and templates of typical brain structure. Though much progress has been made, variation among subjects in the size, shape, and local organization of the brain impedes the development of universally valid, but fully automated programs. Conversely, drawing anatomical ROIs manually on the structural MRI images requires substantial training, well-defined landmarks, and considerable labor. The subjective nature of ROI

**Brodmann areas** Divisions of the brain based on the influential cytoarchitectonic criteria of Korbinian Brodmann.

**cytoarchitecture** The organization of the brain on the basis of cell structure.

**functional ROI** A region of interest that is chosen based on functional criteria, such as the output of an independent voxelwise analysis.

**localizer task** An simple experimental paradigm designed to identify a set of voxels based on a known functional property, in preparation for subsequent analyses of that region in different paradigms.
drawing means that statistical evaluations are necessary to confirm that ROIs
drawn by different people correspond with each other accurately. Therefore,
ROI creation programs that combine the best features of automated and by-hand
drawing are of great interest. These programs require the user to identify
anatomical landmarks, such as major sulci, and the programs partition the
brain into ROIs based on those landmarks. This combination provides a good
compromise between accuracy and speed of creation.

Due to the variability in function within any anatomical region, ROI
approaches to fMRI analysis should be combined with voxelwise approaches
whenever possible. One potential combined approach is to use the anatomical
ROI as a grouping factor. An investigator can then count and compare, for
each experimental condition, the voxels identified in a voxelwise analysis
within that ROI. The investigator can also remove data for the active voxels
within the ROI and then examine the average time course of BOLD intensity
in the remaining voxels. This approach can be used to search for voxels that
may have a different pattern of activation than those identified using the gen-
eral linear model. Finally, information about the differential timecourses of
voxels within an ROI may lead to novel conclusions about local functions
within a brain region, as will be discussed in Chapter 11.

Intersubject Analyses

So far, we have focused on the issue of identifying areas of activation within a
single subject’s brain. Yet nearly all fMRI studies involve multiple subjects, usu-
ally at least 15 per experiment. Often many more subjects are included, in exper-
iments involving across-group comparisons. How can one use all of this data
to better test experimental hypotheses? Combining data from multiple subjects
presents several challenges. It is difficult to match anatomical locations between
subjects, given the wide variability in shape and size of the adult human brain.
Most experiments overcome this challenge by normalizing all subjects’ data
to a common stereotaxic space, as introduced in Chapter 8, either during prepro-
cessing or following each subject’s analysis. In conjunction with spatial smoothing,
normalization greatly reduces anatomical differences between subjects, at
a cost of functional resolution. Less common, but very useful, are ROI analyses
for identifying particular brain regions based on each subject’s anatomy, as
described in the preceding section. Even if one is confident that the same brain
region is identified in all subjects, whether by normalization or ROI, a theoreti-
cal problem remains: that of how to combine data from that region in multiple
subjects, into a single statistical test.

There are two common approaches for intersubject analyses. For simplic-
ity of exposition, we will consider how each approach affects the analysis of a
single experiment with eight subjects who participated in two blocks with ten
data points in each. However, the basic concepts discussed here apply to any
number of subjects, to any number of data points in any design, and to any
statistical test. The first, and most obvious, analysis approach involves combin-
ing all data points from all subjects into a single analysis. The 80 data points
in the task condition and the 80 data points in the control condition could be
compared using a t-test with 158 degrees of freedom. These approaches are
known as fixed-effects analyses (Figure 10.20A), because they assume that the
experimental effect is fixed, or constant across subjects, apart from the influ-
ence of random noise. Applied to fMRI data, fixed-effects models assume that
the experimental manipulation has the same effect on BOLD signal in every

fixed-effects analyses Intersubject analyses that assume that the effect of
the experimental manipulation is fixed across subjects, with differences
between subjects caused by random noise.
(A) Fixed effects

In a fixed-effects analysis, the experimental effect is assumed to be constant (i.e., fixed) across the subject population, so the experimental manipulation has the same effect on all subjects. The data from all subjects are combined and then undergo testing for significance.

(B) Random effects

In a random-effects analysis, the experimental effect is considered to vary between subjects. A statistical map is created for each subject, and then the output of those statistical tests is subjected to a further level of analysis. The advantage of random-effects analyses is that they allow researchers to make inferences about the populations from which the subjects were drawn.

Figure 10.20 Conceptual outline of fixed- and random-effects analyses. (A) In a fixed-effects analysis, the experimental effect is assumed to be constant (i.e., fixed) across the subject population, so the experimental manipulation has the same effect on all subjects. The data from all subjects are combined and then undergo testing for significance. (B) In a random-effects analysis, the experimental effect is considered to vary between subjects. A statistical map is created for each subject, and then the output of those statistical tests is subjected to a further level of analysis.

random-effects analysis Intersubject analysis that treats the effect of the experimental manipulation as variable across subjects, so that it could have a different effect on different subjects. (Note that, in the present section, the terms "fixed" and "random" refer solely to the subject factor. Other factors in the model are usually fixed, but might be random, depending on the experimental hypotheses.)

Though popular, fixed-effects analyses have an important disadvantage: they restrict statistical inferences to the particular sample of subjects used in the study. Suppose that in two of your subjects you measure a very large effect, while in the other six there is no effect at all. After averaging across the subjects, you compare your conditions by t-test and find that they differ significantly. You immediately recognize that this significant result seems inconsistent with the data, given that there was no effect in 75% of your subjects. This contradiction results from the sensitivity of fixed-effects models to extreme results from individual subjects. Under the assumption that the experimental manipulation affects subjects similarly, the best estimate for its true effect is the mean of the data from all subjects. But, if the manipulation does not affect all subjects similarly, then the mean value might be misleading.

In order to make inferences about the population from which subjects are drawn, analyses must include information about the distribution of the effect across subjects. Each subject can be considered as one of many possible subjects who could have participated in the experiment. The experimental manipulation could have a different effect on each of these potential subjects; that is, some could have a large BOLD response, while others could have a small BOLD response. Most fMRI studies use a multi-stage random-effects analysis (Figure 10.20B). Since the combination of data from multiple subjects almost always treats independent variables as having fixed effects at earlier stages of
analysis, random-effects analyses (at the intersubject level) can also be considered as mixed-effects analyses, if information about the variability at the subject level is carried up to the intersubject level.

For clarity of presentation, we will describe a three-level analysis that progresses from runs to subjects to the sample; some fMRI analysis software combines the first two of these levels into one step. In the first-level analysis, the researcher calculates summary statistics (e.g., parameter estimates for regressors of interest) for the data from each run from each subject, independently. Then, in the second-level analysis, these statistics are combined from all runs performed with each subject, using a fixed-effect analysis. If a voxelwise approach is used, this creates a statistical map of the contrasts of interest for each subject. In the third-level analysis, the distribution of data from all the subjects is itself tested for significance. As a rough approximation, this can be done using a t-test that evaluates whether the subjects' summary statistics are drawn from a distribution with a mean of zero. More powerful statistical approaches incorporate variances from earlier levels to better estimate the true significance of the effects. If this third-level statistical test is significant at the established alpha value, then the researchers can conclude that the experimental manipulation would have an effect on the population from which the subjects were drawn. Note that the subject population for many fMRI studies may itself be unrepresentative, in that subjects tend to be college-age, intelligent, physically healthy, and neurologically normal. The use of random-effects analyses does not allow extension of results to those individuals who are not within the subject population (e.g., the elderly, children, or patient groups).

**Thought Question**

Given the limitation of random-effects analyses discussed here, what sorts of experiments are necessary for fMRI results to be applicable to a wide range of subject groups? What problems with subject selection and experimental design would researchers encounter?

In summary, fixed-effects analyses allow inferences about the group of subjects who were used in a particular study; while random-effects analyses allow inferences about the population from which the subjects were drawn. Random-effects analyses can be conducted with minimal additional computation, and most statistical packages now include them within standard analysis procedures. While studies using fixed-effects models are still published, many journal reviewers and granting agencies have recognized the importance of random-effects models, and evaluate manuscripts and grants accordingly. Thus, random-effects analyses are strongly recommended for fMRI studies.

**Group and parametric effects**

Often, fMRI researchers want to go beyond the identification of significant activation to describe how some subject characteristics influence the amplitude of that activation. Between-group comparisons involve testing different groups, such as males and females, to look for differences in some key characteristic. This approach is particularly important for fMRI studies of clinical populations. To better understand the brain dysfunction that contributes to depression, researchers might compare a patient group consisting of individuals with major depression with a control group consisting of individuals without any
depressive symptoms. Similarly, to understand age-related changes in frontal lobe function, researchers might compare a group of 20–30-year-old individuals with a group of 65–75-year-olds. Between-session comparisons involve testing the same individuals in multiple sessions that differ only in the independent variable (e.g., before and after drug therapy). Both types of comparison require the use of different statistical tests than those used for single-group analyses. Remember that single-group analyses evaluate whether an independent variable caused systematic changes in the BOLD signal (i.e., the parameter effects or contrast values were unlikely to have occurred by chance). Between-group comparisons evaluate whether two groups of individuals were influenced differently by the independent variable (i.e., the group differences in their parameter effect were unlikely to have occurred by chance), usually through simple statistics like a two-sample t-test. Because between-session comparisons include two or more similar tests per subject, intersubject variability can be accounted for using a paired t-test.

Group comparisons can lead to valuable new results, but do require careful planning. As for any other between-subjects research, the primary challenge of fMRI group comparisons is matching the groups with regard to factors other than the independent variable of interest. For example, depressed and non-depressed individuals could differ in many factors (e.g., current job status, number of close friends, medication history) in addition to the depression itself. As introduced in Chapter 8, groups may differ in the amount of noise in the BOLD signal, through task-unrelated factors like increased head motion. Furthermore, the effects observed in between-group studies are generally small, given that they are usually modulations of some other effect. Session comparisons limit inter-individual variability, and thus provide greater experimental power. However, repeating an experiment over multiple sessions can introduce unwanted practice effects, confounding differences in brain function with changes in task difficulty or subject strategy. In general, sample sizes for group comparisons need to be much larger than for standard single-group studies, or Type II errors will be likely.

Another approach for understanding differences between groups is to use some parametric trait (e.g., extraversion, depression score, age) as a covariate in the across-subjects analyses. This approach is used to test hypotheses of the form: in what brain regions does the differential activation identified by contrast X vary with trait Y? In a study published in 2007, Spitzer and colleagues investigated the relationship between the neural response to social punishment and the personality trait of selfish opportunism, also known as "Machiavellianism." Each subject played an economic game in which they could divide up a pot of money between himself and an opponent outside of the scanner. In some trials, the opponent could sacrifice some of his own money to take money away from the subject (i.e., as punishment), if the division seemed unfair. The authors found that the lateral orbitofrontal cortex exhibited increased activation when there was the potential for punishment, compared with control trials when the opponent could not respond (Figure 10.21). Notably, the magnitude of this activation correlated well with the subjects’ Machiavellianism scores, as measured by an independent personality test, suggesting that the Machiavellian trait may reflect (or lead to) increased sensitivity to social punishments. Parametric across-subjects tests can illustrate novel and compelling relationships between brain and behavior. Researchers must be careful, when reporting their findings, not to overstate the meaning of their results by mapping complex traits onto single brain regions. Moreover, the direction of causality (i.e., from trait to process to brain
differences, or from brain differences to process to trait) can rarely be gleaned from fMRI data.

Displaying Statistical Results

The goal of most fMRI statistical tests, regardless of their complexity, is to evaluate the probability that the response in each voxel is consistent with the null hypothesis. When statistical tests from all voxels in the brain are combined, the result is a statistical map, or statistical parameter map, of brain activation (see also Figure 10.1). The statistical map is usually color-coded according to the probability value for each voxel. For example, if the (corrected) alpha value for an experiment was set at 0.01, a voxel with a near alpha probability value of 0.009 might be displayed in dark red, while a voxel with an extremely low probability value of 0.000001 might be shown in bright yellow. The association between probability values (or other statistic) and the colors that label them is known as a color map. In general, researchers use darker colors to indicate low significance levels and brighter colors to indicate high significance levels. The statistical map is usually displayed on top of a base image that illustrates brain anatomy. It is important not to confuse the properties of the base anatomical image with those of the overlaid statistical map. The former is usually of high resolution and has contrast dependent on a physical property of the brain (e.g., T₁), while the latter is a calculated statistical map reflecting the correspondence of the data to an experimental hypothesis. Note that although

statistical map (or statistical parameter map) In fMRI, the labeling of all voxels within the image according to the outcome of a statistical test.

color map The association between numerical values of a parameter and a set of colors.

base image The image on which a statistical map is displayed, often a high-resolution anatomical image.
**radiological convention** The practice of displaying images of the brain so that the left side of the image is the right side of the brain and vice versa, as if one were facing the subject.

**neurological convention** The practice of displaying images of the brain so that the left and right sides of the image correspond to the same sides of the brain, as if one were behind the subject.

The vast majority of color maps display statistical significance, other properties, such as percent signal change or latency, can also be displayed; see Figure 7.25 for an example.

There are many options for displaying fMRI data, each with advantages and disadvantages. Most commonly shown are single anatomical slices with overlaid color maps (Figure 10.22A). While these are relatively simple to create and interpret, readers may find it challenging to identify specific anatomical regions (i.e., which gyrus or sulcus is active). Depending on the area of interest, one slice orientation may be better than another. Gyri and sulci that run from left to right (e.g., the central sulcus) are difficult to identify in coronal slices, while those running from front to back (e.g., most frontal gyri) are harder to interpret in axial slices. Another limitation of single-slice displays is the choice of slices to include. Rarely will all the collected slices be displayed in a single poster, manuscript, or lecture slide, due to their sheer number. Instead, the researcher will display selected slices that illustrate the major activation locations found in the study. When showing single slices in research articles, it is critical to explicitly label the left and right hemispheres, due to the axial symmetry of the human brain. Historically, MRI data have been displayed in **radiological convention**, such that the left side of the image corresponds with the right side of the brain and vice versa. This convention results from the way in which radiologists typically interact with patients, who are generally facing them or lying in scanners with their feet toward them. Displaying fMRI data in normal or **neurological convention** (i.e., based on a surgeon looking from the head to the feet) has become increasingly common in recent years.

![Figure 10.22](image-url) Two- and three-dimensional representations of fMRI data. Statistical maps of fMRI data are typically shown either as two-dimensional slices (A) or as three-dimensional rendered brains (B). (B from Huetel et al., 2001, created using FreeSurfer, CorTechs/Marti nos Imaging Center, Boston, MA.)
While statistical maps are often calculated and displayed as two-dimensional slices, they can also be displayed in three-dimensional perspective (Figure 10.22B). Such displays are often called rendered images. The major advantage of three-dimensional rendering is that one can easily identify the locations of brain activation, especially with regard to prominent gyri and sulci. Such images are also more easily interpreted by naïve viewers who are more familiar with the general shape and external structure of the brain than with specific internal features. Hidden, however, are internal structures, such as the basal ganglia and thalamus, as well as internal cortical regions like the hippocampus, cingulate, and insula. In published figures, researchers often cut away a corner or side of the brain to reveal these and other deeper structures. It is also more difficult to show both hemispheres simultaneously in a single image, so authors often include more than one rendered image within a single figure. To overcome the problems introduced by the opacity of rendered images, some analysis programs generate transparent or semitransparent images, sometimes called glass-brain views (Figure 10.23A). These project all foci of activation onto an outline of the brain, allowing the researcher to see all of the activations at once. As with any projective display technique, a single glass-brain view is considerably underdetermined, in that many possible sets of activations could lead to the displayed image. For this reason, three orthogonal projections of the data are displayed. In addition, most programs also generate a set of three anatomical or reference slices that correspond with the glass-brain views, so that researchers can more easily localize activations (Figure 10.23B). Even experienced fMRI scientists can find complex glass-brain figures difficult to interpret, making them inappropriate for many purposes. They can, however, clearly show the specificity (or absence) of activation in a given contrast.

Rendering a brain image can be very computationally intensive, especially if the analysis program generates a smooth high-resolution surface based on an anatomical image. Many researchers, therefore, use simple two-dimensional rendered image A display of MRI data in three-dimensional perspective.
glass-brain view A two-dimensional projection of fMRI data, as if the brain were made transparent and only the activations were visible.

Figure 10.23 Glass-brain views of fMRI data. One common way to visualize fMRI data is to use a “glass-brain” view (A), which shows three orthogonal projections of the original data. The red arrow (>) in each view shows the same brain location. All activations are visible in each orientation, although it can be difficult to identify their anatomical locations. For this reason, although glass-brain views are useful when working with the data, other views may be preferable for a final display. Use of projection views in conjunction with anatomical or reference slices (B) makes identification of activation foci much easier. (Images created using SPM; Wellcome Department of Cognitive Neurology, London, UK.)
When working with data from an experiment, and only create the high-quality rendered images when making final figures for a manuscript. Besides standard surface views, several other types of rendered images are available. To illustrate activation that lies within deep cortical sulci, the outer cortical layer can be removed or de-emphasized. For applications where precise distinctions between adjacent brain regions must be made, researchers may display fMRI data on inflated brains or flat maps (Figure 10.24). This is most common for retinotopic mapping of the visual system. Remember that the cortex is basically a single folded sheet about 5 mm in depth. As its name implies, an inflated brain recovers this structure by expanding the cortical surface like a balloon while maintaining its basic shape. A flat map is obtained by cutting the inflated surface at different points and then laying out the cortical sheet in two dimensions. Since no changes in depth are visible in these techniques, the original gyral and sulcal patterns are marked using different colors or brightness levels. When a three-dimensional object like the brain is transformed into a two-dimensional map, there are by necessity some local distortions. Compare, for example, world globes and world maps; the latter either have cuts between adjacent areas (i.e., Goode’s interrupted map) or distortions in size (i.e., Mercator projections). Similarly, distortion is a significant problem in the generation of flat maps of the brain.

**inflated brain** A transformation of the cortical sheet into a balloon-like structure, removing gyral and sulcal folds so that activation can be more easily viewed.

**flat map** An unfolded and flattened representation of the cortical sheet to allow viewing of topographic changes over cortical space. Flat maps are most commonly used in fMRI to illustrate the organization of the visual cortex.
Summary

Since fMRI studies rely on the detection of a weak signal in the presence of substantial noise, careful statistical analysis is necessary. Most fMRI analyses are based on hypothesis testing. The researcher sets up two hypotheses, an experimental hypothesis with some prediction about the data and a null hypothesis based on random chance. The probability that the data could have occurred under the null hypothesis is compared with a threshold alpha value, if the alpha value is exceeded, then the result is declared significant. Two types of errors are possible in hypothesis testing. Type I errors occur when a non-significant result is declared significant (a false positive), while Type II errors occur when a significant result is missed (a false negative). Functional MRI analyses typically attempt to exclude Type I errors, but as a result they lead to many Type II errors. Significantly active voxels are displayed on statistical maps, usually with the degree of significance indicated using a color scale.

Nearly all statistical tests used for hypothesis testing in fMRI experiments are variants of the general linear model, which treats the data as a linear summation of independent regressors (i.e., condition effects). When using the general linear model to analyze fMRI data, an experimenter creates a design matrix that contains regressors of interest as well as other known, but uninteresting, sources of variability in the data (nuisance regressors). Based on this design matrix, statistical software is used to determine how a given regressor predicts changes in the BOLD signal (parameter estimates) and whether some regressors lead to larger changes than others (contrast effects). Regardless of the analysis approach used, a central problem for fMRI studies is the vast number of statistical tests they require, which leads to false positive results. Standard corrections like the Bonferroni method are too strict and may eliminate true activations. One approach is to use information about spatial properties of activation, either through Gaussian random field theory or cluster-size thresholds. Another approach is to reduce the number of tests by using region-of-interest analyses, which allow targeted studies of particular anatomical areas. Finally, for all statistical analyses, data should be combined from multiple subjects to increase experimental power. Random-effects analyses enable researchers to make inferences about the populations from which the subjects were drawn, and are therefore preferred over traditional fixed-effects analyses.

Suggested Readings


Chapter References


