7

Statistical inference on images

The goal of statistical inference is to make decisions based on our data, while accounting for uncertainty due to noise in the data. From a broad perspective, statistical inference on fMRI data is no different from traditional data analysis on, say, a response time dataset. Inference for fMRI is challenging, however, because of the massive nature of the datasets and their spatial form. Thus, we need to define precisely what are the features of the images that we want to make inference on, and we have to account for the multiplicity in searching over the brain for an effect.

We begin with a brief review of traditional univariate statistical inference and then discuss the different features in images we can make inference on and finally cover the very important issue of multiple testing.

7.1 Basics of statistical inference

We will first briefly review the concepts of classical hypothesis testing, which is the main approach used for statistical inference in fMRI analysis. A *null hypothesis* H_0 is an assertion about a parameter, some feature of the population from which we're sampling. H_0 is the default case, typically that of "no effect", and the *alternative hypothesis* H_1 corresponds to the scientific hypothesis of interest. A *test statistic* T is a function of the data that summarizes the evidence against the null hypothesis. We write T for the yet-to-be-observed (random valued) test statistic, and t for a particular observed value of T. (Note here T stands for a generic *Test* statistic, not t-test.) While there are many different possible types of test statistics with different units and interpretations (e.g., t-tests, F-tests, χ^2-tests), the *P-value* expresses the evidence against H_0 for any type of T: The P-value is $P(T > t|H_0)$, the chance under the null hypothesis of observing a test statistic as large or larger than actually observed. (Tests for decreases in T or two-sided changes, i.e., either positive or negative, are possible by redefining T.)

It is useful to dispense with two frequent misunderstandings about P-values. First, and crucially, the P-value *is not* the probability that the null is true given the data, $P(H_0|T)$. To determine this quantity, we must use Bayesian computations that are not part of Classical hypothesis testing (see Box 7.1). Roughly, the P-value expresses the surprise of observing the data if the null hypothesis was actually true. Second, a P-value can only be used to refute H_0 and doesn't provide evidence for the truth of H_0. The reason for this is that the P-value computation begins by assuming that the null hypothesis is true, and thus a P-value cannot be used to deduce that H_0 is true.

When P-values are used to decide whether to reject H_0 or not, there are two different types of errors that one can make, and we can quantify the likelihood of each. Rejecting H_0 when there is no effect is a Type I or false positive error. The desired tolerance of the chance of a false positive is the *Type I error level*, denoted α. Failing to reject H_0 when there truly is an effect is a Type II or false negative error. The chance that a testing procedure correctly rejects H_0 when there is a true effect is the *power* of the procedure (which is one minus the Type II error rate). Power varies as a function of the size of the true effect, the efficiency of the statistical procedure, and the sample size. This implies that a sample size that is sufficient to detect an effect in one study (which has a relatively large effect magnitude using a sensitive statistical test) may not be sufficient to find an effect in other studies where the true effect is smaller or the test is less sensitive. In Section 7.6 we consider power calculations in detail.

For any testing procedure used to make "Reject"/"Don't Reject" decisions, based either on T or on P, there are several ways to describe the performance of the test. A test is said to be *valid* if the chance of a Type I error is less than or equal to α; if this chance exceeds α, we say the test is *invalid* or *anticonservative*. A test is *exact* if the chance of a Type I error is precisely α, while if this probability is strictly less than α we say the test is *conservative*. In this terminology, we always seek to use valid tests, and among valid tests we seek those with the greatest power.

Box 7.1 Bayesian statistics and inference

Bayesian methods are growing in popularity, as they provide a means to express prior knowledge before we see the data. Thomas Bayes (1702–61) is remembered for the following theorem:

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

which says the chance of random event A occurring assuming or *given* that B occurs, can be computed from an expression involving the reverse statement, the chance of B given A. This expression gives a formal mechanism for combining

prior information with information in the data. In the context of the GLM with data y and parameters β, it allows us to write $f(\beta|y) \propto f(y|\beta)f(\beta)$, where $f(\beta)$ is the *prior* density, our beliefs about the parameters before we see the data (e.g., that BOLD percent changes generally range from -5% to $+5\%$), $f(y|\beta)$ is the traditional likelihood of the data given parameters, and $f(\beta|y)$ is the posterior, the distribution of the parameter after we observe the data. Crucially, it allows us to make probabilistic statements on the unknown parameter β, whereas classical (or *frequentist*) statistics assumes β is fixed and has no random variation. Bayesian inference is based entirely on the posterior: The posterior mean provides a point estimate, and the posterior standard deviation provides the equivalent of a standard error.

There are fundamental differences between the classical and Bayesian approaches. A classical method couches inference relative to infinite theoretical replications of your experiment: A 95% confidence interval means that if you were to repeat your experiment over and over, 19 out of 20 times (on average) the interval produced will cover the fixed, true, but unobservable parameter. The randomness of the data over hypothetical experimental replications drives frequentist inference. The Bayesian method casts inference based on belief about the *random* unobservable parameter: The prior expresses belief about the parameter before seeing the data, the posterior expresses belief about the parameter after seeing the data. There is no reference to infinite replications of your experiment, as the data are fixed (not random).

A true Bayesian thinks a classical statistician is absurd for referencing imaginary experiments that are never conducted. A true classical statistician thinks a Bayesian is irrational because different scientists (with different priors) could analyze the same data and come to different conclusions. Fortunately, in many settings the Bayesian and classical methods give similar answers, because with more and more data the influence of the prior diminishes and the posterior looks like the classical likelihood function.

7.2 Features of interest in images

For an image composed of V voxels, it might seem that there is only one way to decide where there is a signal, by testing each and every voxel individually. This approach is referred to as 'voxel-level' inference. Alternatively, we can take into account the spatial information available in the images, by finding connected clusters of activated voxels and testing the significance of each cluster, which is referred to as 'cluster-level' inference. (See Figure 7.1.) Finally, we might sometimes simply want to ask 'is there any significant activation anywhere?' which is referred to as a 'set-level' inference. First, we discuss what it means to have a significant voxel-level or cluster-level result,

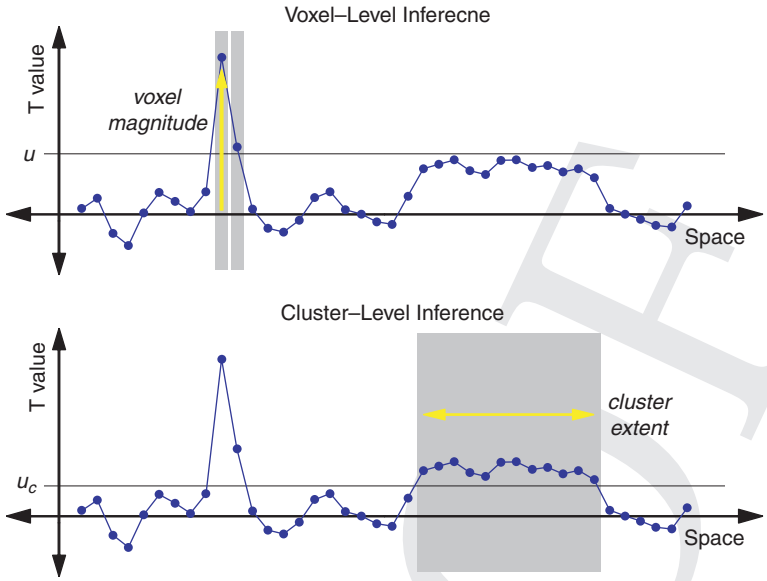

Figure 7.1. Illustration of voxel-level versus cluster-level inference. Both axes show the same one-dimensional section of a statistic image. In the top, voxel-level inference finds two voxels above a significance threshold, and thus both voxels are individual marked as significant. In the bottom, a cluster-forming threshold defines clusters, and cluster-wise inference finds a single cluster of 12 voxels significant; none of the 12 voxels are individually significant, but together they comprise a significant cluster.

and then how we actually compute significance (P-values) accounting for the search over the brain.

7.2.1 Voxel-level inference

In an image of test statistics, each voxel's value measures the evidence against the null hypothesis at that location. The most spatially specific inference that we can make is to determine whether there is a significant effect at each individual voxel. We do this by examining whether the statistic at each voxel exceeds a threshold u; if it does, then we mark that voxel as being "significant" (i.e., we reject the null hypothesis at that voxel. Such voxel-by-voxel inferences allow us to make very specific inferences if the threshold is chosen properly; in Section 7.3 we discuss how the threshold is chosen.

7.2.2 Cluster-level inference

Voxel-level inferences make no use of any spatial information in the image, such as the fact that activated voxels might be clustered together in space. However, we generally expect that the signals in fMRI will be spatially extended. One reason is that the brain regions that are activated in fMRI are often much larger than the size of a single voxel. The second reason is that fMRI data are often spatially smoothed

and then oversampled to small (e.g., 2 mm^3) voxels during spatial normalization, which results in a spreading of the signal across many voxels in the image.

To take advantage of this knowledge about the spatial structure of fMRI signals, it is most common to make inferences about clusters of activated voxels rather than about individual voxels, which is referred to as *cluster-level inference*. The most common approach to cluster-level inference involves a two-step procedure. First, a primary threshold (known as a *cluster-forming threshold*) u_c is applied to a statistic image, and the groups of contiguous voxels above u_c are defined as 'clusters'. What exactly constitutes 'contiguous' depends on the definition of a neighborhood. For example, in 2D, we certainly would consider two suprathreshold voxels connected if they share an edge (4-connectivity), but might also consider them connected if they share a corner (8-connectivity). In 3D, the choices are 6-connectivity (only faces), 18-connectivity (also edges) or 26-connectivity (also corners).[1] Second, the significance of each cluster is determined by measuring its size (in voxels) and comparing this to a critical cluster size threshold k. Methods for choosing this threshold k are discussed below in Section 7.3.

Cluster size inference is generally more sensitive than voxel-level inference for standard MRI data (Friston et al., 1996a). In rough terms, cluster size inference should be better at detecting a signal when that signal is larger in scale than the smoothness of the noise. To see this, consider an example where our fMRI noise has smoothness of 10 mm FWHM and the true effects are also 10 mm in scale. In this instance, the true signal clusters will be similar in size to noise-only clusters, and it will be difficult for cluster-level inference to detect the signal. In contrast, if the scale of the effect is larger than 10 mm, cluster-level inference should detect the effects more often than voxel-level inference. Assigning signifiance to clusters based on their extent ignores the statistic values within a cluster. It would seem that using such intensity information would improve the sensitivity of cluster inference, and indeed some authors have found this result. Poline & Mazoyer (1993) proposed inference using the minimum of the cluster size P-value and cluster peak P-value, and Bullmore et al. (1999) suggested *cluster mass* inference based on the sum of all voxel-level statistic values in a cluster. For the mass statistic in particular, Hayasaka & Nichols (2004) found that it has equal or greater power than the size statistic.

There are two drawbacks to cluster-level inference: The arbitrary cluster-forming threshold and the lack of spatial specificity. The cluster-forming threshold u_c can be set to any value in principle, though if set too low, focal signal may be lost in the gigantic clusters that are formed, and if set to high it may exclude voxels with weak signal intensity (see Figure 7.2). Also, random field theory results break down for thresholds more generous than $\alpha = 0.01$ (see Section 7.3.1). Most troubling, if one

[1] The SPM software uses 18-connectivity while the FSL uses 26-connectivity; in practice you find very similar clusters with either connectivity unless the data has very low smoothness.

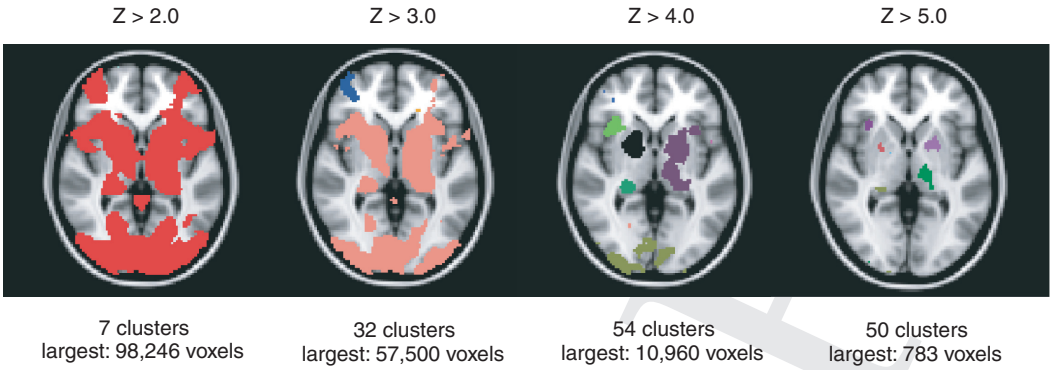

Figure 7.2. Effects of cluster-forming threshold on cluster size. The same data were thresholded using increasing cluster-size thresholds; the resulting clusters are randomly color-coded to show which voxels belong to each cluster. At the lowest threshold, there is one large cluster that encompasses much of the brain, whereas higher thresholds break up this cluster, at the expense of excluding many regions that do not survive the higher threshold.

adjusts u_c up or down just slightly, some clusters may merge or split, and significant clusters disappear. In practice, most users take u_c to correspond to either $\alpha = 0.01$ or $\alpha = 0.001$ (FSL users must set a statistic value threshold rather than a P-value, usually $t = 2.3$ or 3.1).

Cluster inference's greater power comes at the expense of spatial specificity, or precision. When a 1,000 voxel cluster is marked as statistically significant, we cannot point to a single voxel in the cluster and say "The signal is here." All we can conclude is that one or more voxels within that cluster have evidence against the null. This isn't a problem, though, when cluster sizes are small. If you get a cluster that covers half the brain, however, this can be quite unsatisfying. The only remedy is to resort to raising u_c to get smaller clusters, but this further compounds the multiple testing problem because one is searching across multiple thresholds.

A recently developed method attempts to address these two problems. Threshold Free Cluster Enhancement (TFCE) (Smith & Nichols, 2009) uses all possible u_c, and then integrates over u_c to provide a voxel-level map that indicates cluster-level significance. By eliminating one parameter it does introduce two new parameters, specifically how to weight u_c versus cluster size, but these are set to fixed values inspired by theory and empirical simulations. While not an established approach, it has shown promise as a sensitive approach to cluster-level inference that removes the dependence on the cluster-forming threshold.

7.2.3 Set-level inference

Although rare, there may be some cases when one simply wants to know if there is any significant activation for a particular contrast, with no concern for where the activation is. In SPM, there is an inference method known as *set-level inference* that

Box 7.2.3 Inference on location vs. inference on magnitude

When we apply a threshold to a statistic image and search the brain for activations, the end result is an inference on location. We answer the question: "*Where* in the brain is there a response to my experiment?" Once a significant region is identified as active, one would like to characterize the nature of the effect, in particular the effect magnitude. However, due to a problem of *circularity* (discussed in greater detail in Chapter 10; see Box 10.4.2), we cannot subsequently answer the question of *how large* the identified effect is. The reason is that of all the possible true positive voxels we will detect, we are more likely to find the voxels that are randomly higher than the true effect and will miss those that are randomly smaller. In genetics this is known as the "winner's curse," as the first group to find a gene will often report an effect size that is greater than any subsequent replication.

At the present time, there is no way to correct for the bias in effect sizes found by searching the brain for activations. One simply must recognize that effect size bias is present and note this when discussing the result. If unbiased effect size estimates are required, one must sacrifice inference on location and instead assume a fixed and known location for the effect. Specifically, one must use a priori specified regions of interest (ROIs) and average the data within those regions. For more on the topic of circularity, see Kriegeskorte et al. (2009).

is a overall test of whether there exists any significant signals anywhere in the brain. The test statistic is the number of clusters for an arbitrary cluster defining threshold u_c that are larger than an arbitrary cluster size threshold k. A significant set-level P-value indicates that there are an unusually large number of clusters present, but it doesn't indicate *which* clusters are significant. For this reason it is referred to an *omnibus* test and has no localizing power whatsoever.

7.3 The multiple testing problem and solutions

As previously reviewed, classical statistical methods provide a straightforward means to control the level of false positive risk through by appropriate selection of α. However, this guarantee is a made only on a test-by-test basis. If a statistic image has 100,000 voxels, and we declare all voxels with $P < 0.05$ to be "significant," then on average 5% of the 100,000 voxels – 5,000 voxels – will be significant as false positives! This problem is referred to as the *multiple testing problem* and is an critical issue for fMRI analysis.

Standard hypothesis tests are designed only to control the 'per comparison rate' and are not meant to be used repetitively for a set of related tests. To account for the multiplicity, we have to measure false positive risk over an entire image. We define,

in turn, two measures of false positive risk – the familywise error rate and the false discovery rate.

7.3.1 Familywise error rate

The most common measure of Type I error over multiple tests is the 'familywise error rate', abbreviated FWER or FWE. FWE is the chance of one or more false positives anywhere in the image. When we use a valid procedure with $\alpha_{\mathrm{FWE}} = 0.05$, there is at most a 5% chance of *any* false positives anywhere in the map. Equivalently, after thresholding with a valid $\alpha_{\mathrm{FWE}} = 0.05$ threshold, we have 95% confidence that there are no false positive voxels (or clusters) in the thresholded map. For a particular voxel (or cluster), we can refer to its "corrected FWE *P*-value" or just "corrected *P*-value," which is the smallest α_{FWE} that allows detection of that voxel (or cluster).

Several procedures that can provide valid corrected P-values for fMRI data are available.

7.3.1.1 Bonferroni correction

Perhaps the most widely known method for controlling FWE is the 'Bonferroni correction.' By using a threshold of $\alpha = \alpha_{\mathrm{FWE}}/V$, where V is the number of tests, we will have a valid FWE procedure for any type of data. However, even though it will control FWE for any dataset, the Bonferroni procedure becomes conservative when there is strong correlation between tests. Because of the smoothness of fMRI data, Bonferroni corrections are usually very strongly conservative. Instead, we need a method that accounts for the spatial dependence between voxels. The two main methods that do this are random field theory (RFT) and permutation methods.

7.3.1.2 Random field theory

Random field theory uses an elegant mathematical theory on the topology of thresholded images. The details of this method require mathematics beyond the scope of this book, but an approachable overview can be found in Nichols & Hayasaka (2003); treatments with more mathematical detail can be found in Cao & Worsley (2001) and Adler & Taylor (2007).

A crucial aspect of RFT is how it accounts for the degree of smoothness in the data. Smoothness is measured by FWHM $= [\mathrm{FWHM}_x, \mathrm{FWHM}_y, \mathrm{FWHM}_z]$. This smoothness is *not* the size of the Gaussian smoothing kernel applied to the data, but it is important to point out that the smoothness of the data is *not* equivalent to the size of a Gaussian smoothing kernel applied to the real data, but rather the intrinsic smoothness of the data. That is, even before any smoothing, there is some spatial correlation present in all imaging data, and the RFT smoothness parameter relates to the combination of the intrinsic and applied smoothing.

The definition of RFT's FWHM is somewhat convoluted: It is the size of a Gaussian kernel that, when applied to spatially independent "white noise" data, induces the

degree of smoothness present in the noise of the data at hand. See previous citations for a more precise definition in terms of the variability of the partial derivatives of the noise.

A related concept is 'RESEL' or RESolution ELement, a virtual voxel of size $\text{FWHM}_x \times \text{FWHM}_y \times \text{FWHM}_z$. The analysis volume expressed in units of RESELs is denoted R, the RESEL count.

We present one formula to gain intuition on how RFT results work, the expression for the corrected P-value for a voxel value t in a three-dimensional Gaussian statistic image

$$P_{\text{FWE}}^{\text{vox}}(t) \approx R \times \frac{(4\ln(2))^{3/2}}{(2\pi)^2} e^{-t^2/2}(t^2 - 1) \tag{7.1}$$

where $R = V/(\text{FWHM}_x\text{FWHM}_y\text{FWHM}_z)$ is the RESEL count for the image. This demonstrates the essential role of the RESEL count and shows that, for a given statistic value t and search volume V, as the product of FWHM's increase, the RESEL count decreases and so does the corrected P-value, producing increased significance. The intuition is that greater smoothness means there is a less severe multiple testing problem, and a less stringent correction is necessary. Conversely, as the search volume in RESELs grows, so does the corrected P-value, producing decreased significance for the same statistical value. This should also make sense, as the larger the search volume, the more severe the multiple testing problem.

These observations illustrate how RFT inference adapts to the smoothness in the data, and how the RESEL count is related to the number of 'independent observations' in the image. This loose interpretation, however, is far as it goes, and RFT should never be misunderstood to be equivalent to a 'RESEL-based Bonferroni correction'. This is not the case, and there is no equivalent voxel count that you can feed into Bonferroni correction that will match RFT inferences (Nichols & Hayasaka, 2003).

RFT can also be used to obtain P-values for the clusters based on cluster-size (Friston et al., 1994b). Again, the details are mathematically involved, but Gaussian random field theory provides results for the expected size and number of clusters, and these results adapt to the smoothness of the search volume. RFT P-values have also been developed for the alternate cluster statistics mentioned earlier, combined cluster size and peak height (Poline & Mazoyer, 1993) and cluster mass (Zhang et al., 2009).

Limitations of RFT. Even though RFT methods form the core of fMRI inference, they have a number of shortcomings. First, they require a multitude of distributional assumptions and approximations. In particular, they require that the random field be sufficiently smooth, which practically means that one needs to smooth the data with a Gaussian filter whose FWHM is at least twice the voxel dimensions. In fact, the RFT methods are overly conservative for smoothness less than three- to four-voxel FWHM (Nichols & Hayasaka, 2003; Hayasaka & Nichols, 2003). In addition,

RFT methods are overly conservative for sample sizes less than about 20 (Nichols & Hayasaka, 2003; Hayasaka & Nichols, 2003).

7.3.1.3 Parametric simulations

Another approach to voxel-level and cluster-level inference is Monte Carlo simulation, from which we can find a threshold that controls the FWE. For example, Forman et al. (1995) proposed a Monte Carlo cluster-level inference method. Gaussian data are simulated and smoothed based on the estimated smoothness of the real data, creating surrogate statistic images under the null hypothesis. These surrogate images are thresholded, and an empirical cluster size distribution is derived. These methods have an underlying model that is similar to RFT's model (i.e., smooth Gaussian data), but they do not rely on an asymptotic or approximate results. They are, however, much more computationally intensive than RFT.

This method is implemented in AFNI's alphasim program. Users of this approach must take care that the smoothness parameter, which, as in RFT, is not the size of the applied smoothing kernel but the estimated intrinsic smoothness of the data. In addition, the analysis mask used for the simulation must be exactly the same as the mask used for analysis of the real data.

7.3.1.4 Nonparametric approaches

Instead of making parametric assumptions about the data to approximate *P*-values, an alternative approach is to use the data themselves to obtain empirical null distributions of the test statistic of interest. The two most widely used resampling methods are permutation tests and the bootstrap. While the bootstrap is perhaps better known, it is an asymptotic method (meaning that it is only provably correct in the large-sample limit), and in particular has been shown to have poor performance for estimating FWE-corrected *P*-values (Troendle et al., 2004). In contrast the permutation test, which has exact control of false positive risk, is a useful alternative to RFT methods for small samples.

A permutation test is easy to understand when comparing two groups. Considering just a single voxel, suppose you have two groups of ten subjects, high performers (H) and low performers (L), each of whose BOLD response data you wish to compare. Under the null hypothesis of no group difference, the group labels are arbitrary, and one could randomly select ten subjects to be the H group, reanalyze the data, and expect similar results. This is the principle of the permutation test: repeatedly shuffling the assignment of experimental labels to the data, and analyzing the data for each shuffle to create a distribution of statistic values that would be expected under the null hypothesis. Just as a parametric *P*-value is found by integrating the tails of the null distribution that are more extreme than the actual data observed, the nonparametric permutation *P*-value is the proportion of actually observed permuted statistic values that are as or more extreme than the value that was actually observed. See Figure 7.3 for an illustration with this example with three subjects per group.

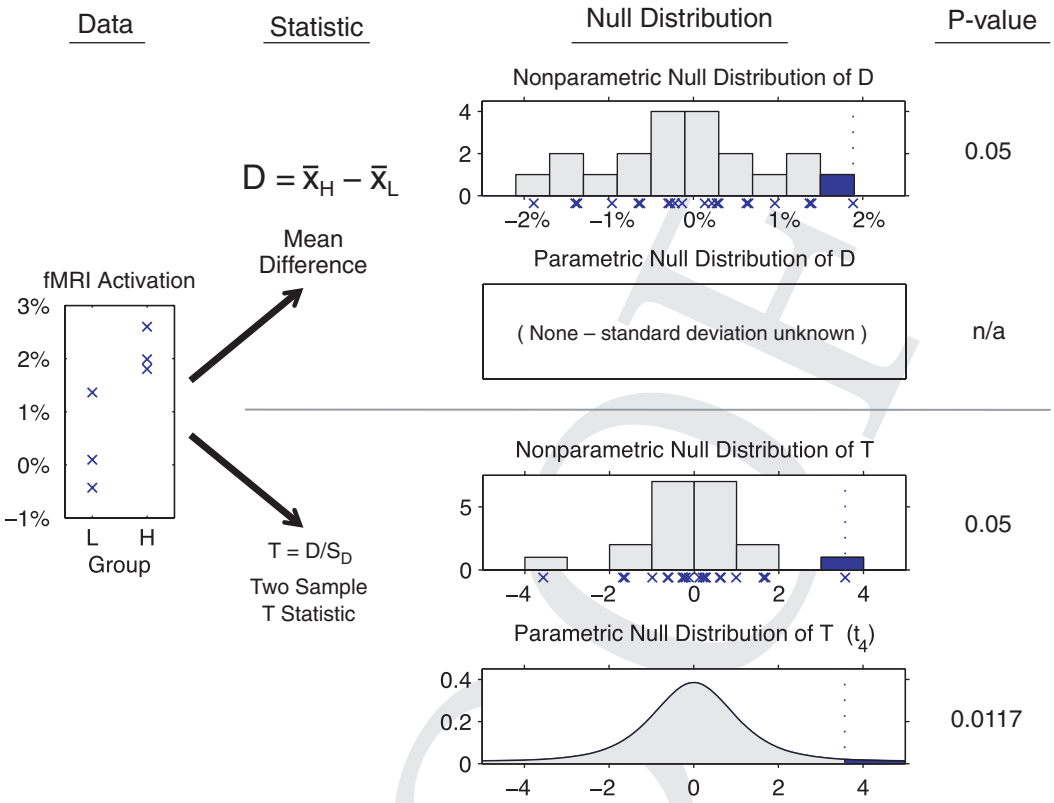

Figure 7.3. Illustration of parametric and nonparametric inference at the group level, comparing two groups of three subjects. Parametric methods use assumptions about the data to find the null distribution of the test statistic. Nonparametric methods use the data itself to find the null distribution, allowing the consideration of nonstandard test statistics. Under the null hypothesis the group labels are irrelevant, and thus we can reanalyze the data over and over with different permutations of the labels. Here, there are 20 possible ways to assign three subjects to the Low-performers group (and the other three must be high performers), and thus the permutation distribution consists of 20 test statistic values. With either parametric or nonparametric methods, the *P*-value is the proportion of the null distribution as large or larger than the statistic actually observed. However, there is no parametric test for the difference, as the standard deviation (S_D) is unknown.

For a single subject's fMRI data, the permutation test is difficult to apply. Drift and temporal autocorrelation make the timeseries autocorrelated and thus not "exchangeable" under the null hypothesis (since randomly reordering the data points would disrupt the temporal autocorrelation). Even though there are methods to decorrelate the data (Bullmore et al., 2001) as part of the permutation procedure, such *semiparametric* methods are very computationally demanding and depend on accurate modelling of the correlation.

At the group level, on the other hand, the permutation approach is easy to apply (this correction is implemented in the *randomise* tool in FSL and in the *SnPM*

toolbox for SPM). Each subject is analyzed with a standard GLM model, and for each contrast of interest, an effect image (a Contrast of Parameter Estimates or COPE image in FSL; a con image in SPM) is created. If there is just a single group of subjects, it might seem that a permutation test is impossible, as there are no group labels to permute. If we instead assume that the COPE images have a symmetric distribution (about zero under H_0), a permutation test can made by randomly multiplying each subject's COPE by 1 or -1. The assumption of symmetry is much weaker than a Gaussian assumption and can be justified by the first-level errors having a symmetric distribution.

So far, we have discussed the use of permutation tests to obtain null distributions at each voxel, but this does not solve the multiple testing problem. Importantly, permutation can also be used to obtain FWE-corrected *P*-values. An FWE-corrected *P*-value is found by comparing a particular statistic value to the distribution of the maximal statistic across the whole image. In the previous High and Low performers example, this means that for each random labeling of Hs and Ls, the entire brain volume is analyzed, and the maximum statistic value across the whole brain is noted. In the case of voxel-level inference, this is the largest intensity in the statistic image, whereas for cluster-level inference, this is size of the largest cluster in the image. With repeated permutation a distribution of the maximum statistic is constructed, and the FWE corrected *P*-value is the proportion of maxima in the permutation distribution that as large or large than the observed statistic value.

The primary drawback of permutation methods is that they are computationally intensive. Whereas RFT computations take seconds at most, a typical permutation analysis can take anywhere from 10 minutes to an hour on modern computing hardware. However, given the great amount of time spent to perform other aspects of fMRI processing, this seems like a relatively small price to pay for the accuracy that comes from using permutation tests. In general, when FWE-corrected results are desired, we recommend the use of permutation tests for all inferences on group fMRI data.

7.3.2 False discovery rate

While FWE-corrected voxel-level tests were the first methods available for neuroimaging, practitioners often found the procedures to be quite insensitive, leaving them with no results that survived correction. While sometimes FWE inferences are conservative due to inaccurate RFT methods, even with exact permutation FWE methods, many experiments will produce no positive results (especially with small sample sizes). A more lenient alternative to FWE correction is the false discovery rate (Benjamini & Hochberg, 1995; Genovese et al., 2002). The false discovery proportion (FDP) is the fraction of detected voxels (or clusters) that are false positives (defined as 0 if there are no detected voxels). FDP is unobservable, but FDR procedures guarantee that the average FDP is controlled. Put another way, where a level 0.05 FWE procedure is correct 95% of the time – no more than 5% of experiments

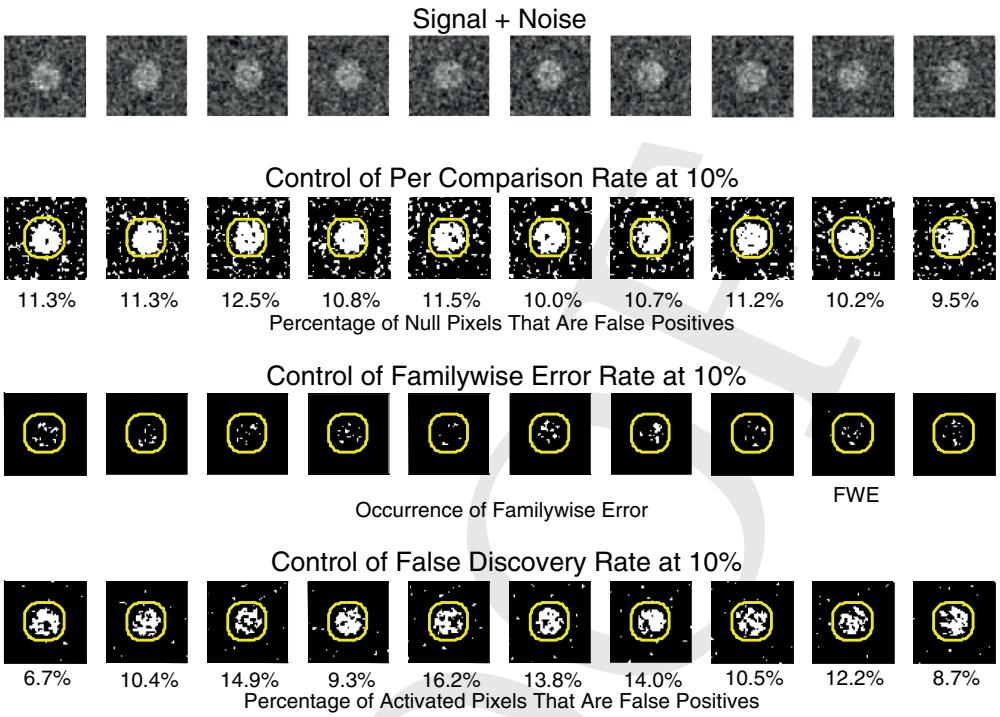

Figure 7.4. Illustration of three different multiple comparison procedures. Each column corresponds to a different realization of signal plus noise as illustrated in the simulated data in the top row, and can be thought of as your next ten experiments. The top row shows the statistic image without any thresholding. The second row illustrates the control of the per comparison rate at 10%, that is, no special account of multiplicity. The third row shows control of the familywise error rate at 10%, say with RFT or Bonferroni. The bottom row shows control of the false discovery rate. With no adjustment for multiple testing (second row) there is excellent sensitivity, but very poor specificity – there are false positives everywhere. Controlling FWE (third row) gives excellent specificity – only 1 out of 10 experiments have *any* false positives– but poor sensitivity. Controlling FDR (bottom row) is a compromise between no correction and FWE correction, giving greater sensitivity at the expense of some false positives, even though it is still controlled as a fraction of all voxels detected. Note that, just as the emperical per comparison error rate for each experiment is never exactly 10%, likewise the emperical false discovery rate is never exactly 10%; in both instances, we're guaranteed only that, in the long run, the average rate will not exceed 10%.

examined can have *any* false positives – a level 0.05 FDR procedure produces results that are 95% correct – in the long run the average FDP will be no more than 5%. (See Figure 7.4.)

FDR's greater sensitivity comes at the cost of greater false positive risk. That risk is still measured in an objective way that accounts for features of the data, which is in contrast to, say, an uncorrected $\alpha = 0.001$ threshold, which will give varying false positive risk depending on smoothness and the size of the search region. Standard

FDR, as applied voxel-level, lacks any kind of spatial specificity, similar to cluster inference. Given a map of FDR-significant voxels, one cannot point to a single voxel and conclude that it is significant. One can only assert that, on average, no more than 5% (or the FDR level used) of the voxels present are false positives. Even if some significant voxels form a large cluster, and others are scattered as tiny clusters, voxel-level FDR has no spatial aspect and does not factor cluster size into significance.

This lack of spatial precision has led some to criticise FDR, with Chumbley & Friston (2009) even recommending that voxel-level FDR should not be used at all. They propose, instead, that FDR should be applied in a cluster-level fashion. On balance, both voxel-level and cluster-level FDR are reasonable procedures, and each needs to be interpreted with care, accounting for the presence of false positives in the map of significant voxels or clusters.

7.3.3 Inference example

To illustrate these inference methods just discussed, we consider data from a gambling task (Tom et al., 2007); these data are available from the book Web site. In this experiment, 16 subjects were offered 50/50 gambles where the size of the potential gain and loss varied parametrically from trial to trial. Here we just consider the negative parametric effect of potential loss on BOLD response (which identifies regions whose activity goes down as the size of the potential loss goes up). Using a cluster-forming threshold of $Z = 2.3$, 154 clusters were found (Figure 7.5a). Based on the search volume and estimated smoothness, RFT finds 5% FWE critical cluster size threshold to be 570 voxels, and only four clusters are larger (Figure 7.5b and Table 7.1).

Note the difficulty in visualizing 3D clusters with orthogonal slices. In Figure 7.5b, in the coronal (middle) slice, there appears to be perhaps six or more separate clusters, yet in fact there are only three clusters shown this panel. Contiguity of clusters is measured in 3D and is very difficult to gauge visually from 2-D slices.

Another challenge arises with large clusters, such as the first cluster in Table 7.1 and as seen in the lower portions of the sagittal and coronal (left and middle) slices. This cluster covers a number of anatomical regions, yet calling this cluster significant only tells us there is some signal somewhere in these 6,041 voxels. If we had known this would be a problem a prori, we could have used voxel-level inference instead to improve spatial specificity. For this data, though, no voxels are found significant with either FWE or FDR (max voxel $Z = 4.00$ has $P^{\text{vox}}_{\text{FWE}} = 1.0$, $P^{\text{vox}}_{\text{FDR}} = 0.1251$). This is typical of cluster inference greater sensitivity over voxel-level inference.

7.4 Combining inferences: masking and conjunctions

A single fMRI experiment will usually produce a number of different contrasts, and fully understanding the outcome of the study may require combining the statistic

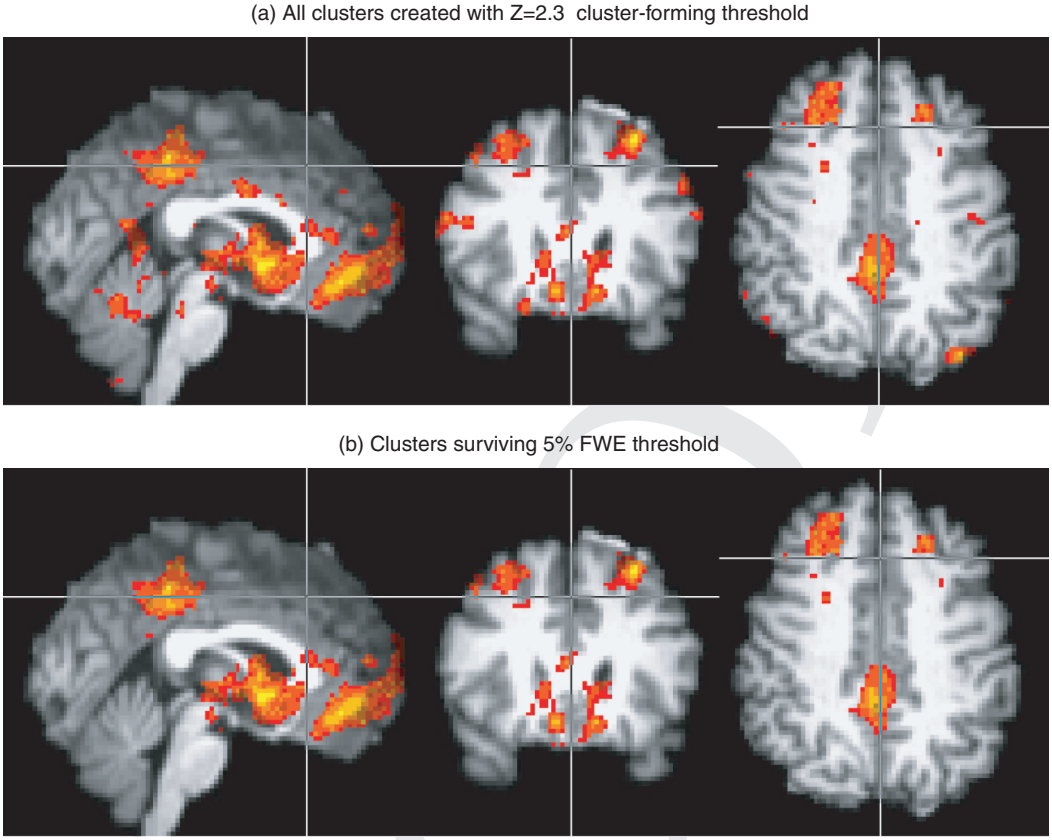

Figure 7.5. Thresholded maps from the gambling experiment, parametric effect of the size of potential loss on BOLD response. Top (a) shows clusters created with $Z = 2.3$ cluster-forming threshold and no cluster-size threshold, while bottom (b) shows the 3 clusters that survive a critical cluster size threshold of 570 voxels.

Table 7.1. Significant clusters from the gambling experiment.

Region	Cluster Size (voxels)	Corrected P-value $P_{\text{FWE}}^{\text{clus}}$	X	Y	Z
Striatum, ventromedial prefrontal cortex, ventral anterior cingulate cortex, medial orbitofrontal cortex	6,041	<0.0001	0	4	−4
Right superior frontal gyrus	1,102	0.0010	22	42	38
Posterior cingulate	901	0.0040	4	−38	40
Left superior frontal gyrus	738	0.0133	−30	24	54

Notes: Search volume: 236,516 $2 \times 2 \times 2$ mm^3 voxels, 1.89 liters, 1,719.1 RESELs, FWHM 5.1 mm
Cluster forming threshold $Z = 2.3$, 0.05 FWE cluster size threshold $k = 570$.
[a] Of the 154 clusters found (see Figure 7.5) with a cluster-forming threshold of $Z = 2.3$, only the four listed here are FWE significant at 0.05. X, Y, Z coordinates listed are the location of the peak Z value in each cluster.

images in different ways. To make these issues concrete, consider a 2×2 factorial design, where there are two factors with two levels each. Henson et al. (2002) uses such a design for a face recognition and implicit memory. That study has two factors, "Fame" indicating whether a presented face is famous or nonfamous and "Repetition" indicating whether this is the first or second presentation of a face (each face was presented exactly twice). Among the contrasts of interest are: $\mathbf{c}_{Famous>Nonfamous}$, the positive effect of famousness, averaged over both presentations; $\mathbf{c}_{Famous:Rep1>Nonfamous:Rep1}$, the famousness effect on the first presentation; $\mathbf{c}_{Famous:Rep2>Nonfamous:Rep2}$, the famousness effect on the second presentation; and $\mathbf{c}_{Fame\times Repetition} = \mathbf{c}_{Famous:Rep1>Nonfamous:Rep1} - \mathbf{c}_{Famous:Rep2>Nonfamous:Rep2}$, a one-sided test of the interaction, repitition-dependent effect of famousness.

The interaction contrast $\mathbf{c}_{Fame\times Repetition}$ is perhaps the most interesting effect, but it detects voxels both where the $\mathbf{c}_{Famous:Rep1>Nonfamous:Rep1}$ effect is positive and greater than $\mathbf{c}_{Famous:Rep2>Nonfamous:Rep2}$ *and* where decreases in the $\mathbf{c}_{Famous:Rep1>Nonfamous:Rep1}$ effect are less negative than decreases in $\mathbf{c}_{Famous:Rep2>Nonfamous:Rep2}$. Assume we are only interested in the interaction when the effects of famousness are positive. We can address this by first creating the statistic image for $\mathbf{c}_{Famous:Rep1>Nonfamous:Rep1}$ and thresholding at 0 to create a binary mask indicating where $\mathbf{c}_{Famous:Rep1>Nonfamous:Rep1}$ is positive. We then create the statistic image for $\mathbf{c}_{Fame\times Repetition}$, apply significance thresholding as usual, and finally apply the binary mask. The resulting map will show significant effects for $\mathbf{c}_{Fame\times Repetition}$ *masked* for positive effects of famousness. Note that here we are using masking as an image processing manipulation, eliminating voxels that satisfy an arbitrary condition on an supplemental contrast. That is, the statistical threshold is uninformed about the nature of the mask, and, in general, the false positive rate will be only lower after application of such a mask. See the next section for use of regions of interest to change the search region and affect the multiple testing correction.

Whereas an interaction looks for differences in effects, a *conjunction* looks for similarities (Nichols et al., 2005). For example, we may wish to find regions where there is a Fame effect for both the first and second face presentation. A conjunction of the tests specified by contrasts $\mathbf{c}_{Famous:Rep1>Nonfamous:Rep1}$ and $\mathbf{c}_{Famous:Rep2>Nonfamous:Rep2}$ will provide this inference. Note that this conjunction is *not* the same as the main effect of Fame, $\mathbf{c}_{Famous>Nonfamous}$, which could be significant if there was a positive Fame effect in just either the first or second presentation.

Valid conjunction inference is obtained by thresholding each statistic image separately and then taking the voxel-level intersection of above-threshold voxels. There is no assumption of independence between each contrast tested, and the voxel-level significance level of the conjunction is that of each of the combined tests; for example, if a 5% FWE voxel-level threshold is applied to each statistic image, the conjunction inference has level 5% FWE. Alternatively, the voxel-wise minimum can be computed, and this minimum image can be thresholded as if it were a single statistic image. The precise definition of conjunction inference is that it

measures the evidence against the *conjunction null hypothesis* that one or more effects are null.

Note there is often low power to detect a conjunction, simply because it is a stringent requirement that each and every tests must demonstrate a significant effect. Friston et al. (2005) proposed a weakened form of conjunction inference that also uses the minimum statistic of K effects. Instead of making inference on the conjunction null hypothesis, which has an alternative hypothesis that all K effects are true, they make inference on an intermediate null whose alternative holds that at least $k < K$ of the effects are true. This alternative approach, however, requires an assumption of independence between the tested effects and, as stated, cannot provide an inference that all effects are true.

7.5 Use of region of interest masks

If a study is focused on a particular region of the brain, then it is possible to limit the search for activations to a region of interest, which reduces the stringency of the correction for multiple testing. In Chapter 10 we discuss the issue of ROI analysis in more detail; here we focus on the use of ROIs with voxel-level or cluster-level inference to reduce the volume of brain searched for activations, often known as a 'small volume correction'. The advantage of this strategy is that the ROI definitions do not have to be very precise, as they are only used to define regions of the brain that are of interest or not. As mentioned before, it is crucial that the ROI is defined independently of the statistical analysis of interest.

The only practical concerns to be aware of is that not all multiple testing procedures work equally well for very small ROIs. Cluster-level inference based on RFT, for example, assumes that the search region is large relative to the smoothness of the noise. Clusters that touch the edge of the search can have their significance underestimated, with either RFT or permutation, thus cluster-level inference is not ideal when using ROIs smaller than about 25 RESELs. For example, if FWHM *in voxel units* is $[3, 3, 3]$ voxels^3, a 1,000-voxel ROI has RESEL count $1,000/(3 \times 3 \times 3) = 37.0$, and thus is sufficiently large. Similarly voxel-level inference with FDR correction can work poorly when ROIs are very small. In essence, FDR has to *learn* the distribution of nonnull P-values to distinguish them from the background of null P-values.

7.6 Computing statistical power

One of the most common question asked of statisticians is "How many subjects do I need in my study in order to detect a hypothesized effect.? To answer this question, we need to compute the *statistical power* of the test. As mentioned at the start of the chapter, power is the probability of correctly rejecting the null hypothesis when it is false (i.e., when there is a true signal). Figure 7.6 illustrates how power is calculated for a simple univariate test. The red distribution is the null distribution of a Z test

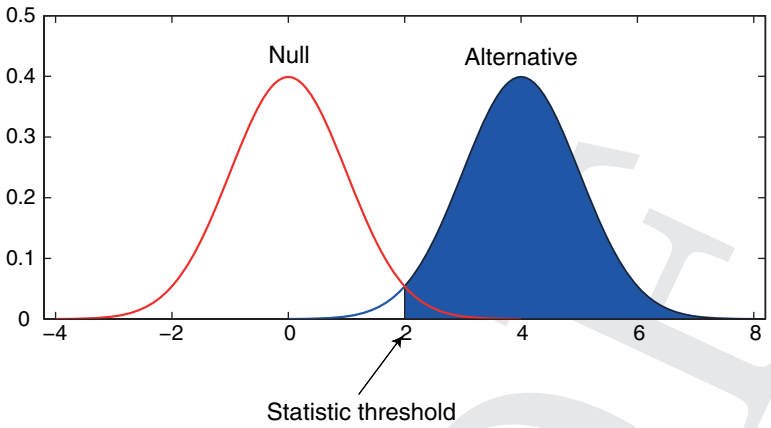

Figure 7.6. Illustration of how power is calculated. The red distribution is the null distribution, which is centered about 0, and the blue distribution is the alternative distribution centered about a value determined by the expected mean and variance of the activation. The statistic threshold indicates the threshold that is used to assess whether or not a statistic value is significant or not. The area under the null distribution to the right of this threshold is the type I error rate, α, and the area under the alternative distribution to the right of this threshold (blue shaded region) is power.

statistic, and the blue distribution is the alternative distribution. The mean of the alternative distribution is a function of the size of the activation you expect to have in your study, its variance, and the sample size. For a given α level (e.g., $\alpha = 0.05$ for a single test), you find the corresponding null distribution threshold such that the area to the right of this threshold under the null distribution is α (the Type I error rate) and then the area to the right of this threshold under the alternative distribution is the power. If your test has 80% power, it means that you will have, with many possible replicate experiments, an 80% chance of detecting the specified signal.

Power analyses must be carried out *prior* to data collection to plan how many subjects are necessary for a study. The power calculation itself is a function of the number of subjects in the study; the Type I error rate, α; the size of the effect that you wish to be able to detect, δ; and the variance of this effect, σ^2. Power is also impacted by the number of runs of data that will be collected and the length of the runs because those factors affect the variance of the effect (see Mumford & Nichols, 2008, for details). Using this calculation, one can compute the number of subjects necessary to find the desired effect with 80% power, which is the generally accepted threshold for reasonable power. The size of the effect and its variance are often based on pilot data or data from a similar previous study. As discussed in Chapter 6, the variance of the effect takes on a complicated form, including a within-subject and between-subject component, and so it must be carefully estimated to reflect this structure.

In fMRI, we are of course faced with thousands of tests, and thus a comprehensive power analysis would require specifying the effect size of every voxel. Further, the probability calculations would have to account for spatial correlation and the multiple testing problem. In practice this isn't done (though see Hayasaka et al., 2007), and to simplify power analyses we consider only an a priori ROI, and predict the power for the mean percent BOLD change in that ROI based on a simple single-group ordinary least squares (OLS) mode. While our aims are rarely so simple, if one doesn't have sufficient power for this setting, any other analysis will surely be underpowered. In this case, the power analysis is simply that of a one-sample t-test. From pilot data, if $\hat{\mu}$ is the ROI mean (over space and subjects) and $\hat{\sigma}$ is the ROI standard deviation (over subjects, of the ROI mean), then the power for a sample size of N and a type I error rate of α would be

$$\text{Power} = P(T_{NCP,N-1} > t_{1-\alpha,N-1}) \tag{7.2}$$

where $T_{NCP,N-1}$ corresponds to a noncentral T random variable where NCP is the noncentrality parameter and is set to $NCP = \frac{\sqrt{N}\hat{\mu}}{\hat{\sigma}}$ and $t_{1-\alpha,N-1}$ is the $1-\alpha$ quantile of a central t distribution with $N-1$ degrees of freedom. For other group models such as a two-sample t-test or ANOVA, models estimated using OLS examples can be found in Cohen (1988), and estimation techniques for full mixed effects models can be found in Mumford & Nichols (2008). A tool for computing power estimates based on previous studies is also available at http://www.fmripower.org.

As an example, say you are planning a new study using a stop signal task and want to ensure you have sufficient subjects to distinguish between successfully stopping versus not successfully stopping in the putamen. You have a previous study with data on 16 subjects for this very sort of experiment and contrast; by using this data we make the assumption that our future study will use a similar scanner and acquisition parameters, preprocessing options, number of trials per run, and runs per subject. Using an anatomical atlas to create a mask for the putamen, we measure the mean BOLD signal change in for each subject (see Section 10.4.3.3 for instructions on converting to percent signal change units). We find that the mean over subjects is 0.8% BOLD signal change units, and the standard deviation across subjects is 2% BOLD. Based on these two numbers and α-level 0.05 using a range of sample sizes with Equation 7.2, the power curve in Figure 7.7 is generated. This curve crosses the 80% mark between 40 and 41 subjects and so a sample of at least 41 subjects will yield at least 80% power, if the given effect is 0.8% and standard deviation is 2%. Note, if you are working on a grant application the power calculations will often not be what you had hoped and you will need to refigure your budget. Because of this, carrying out your power analyses well in advance of your grant deadline is highly recommended.

Several limitations of power analyses are worth considering. First and foremost, appreciate that power computations are quite speculative enterprises. The whole

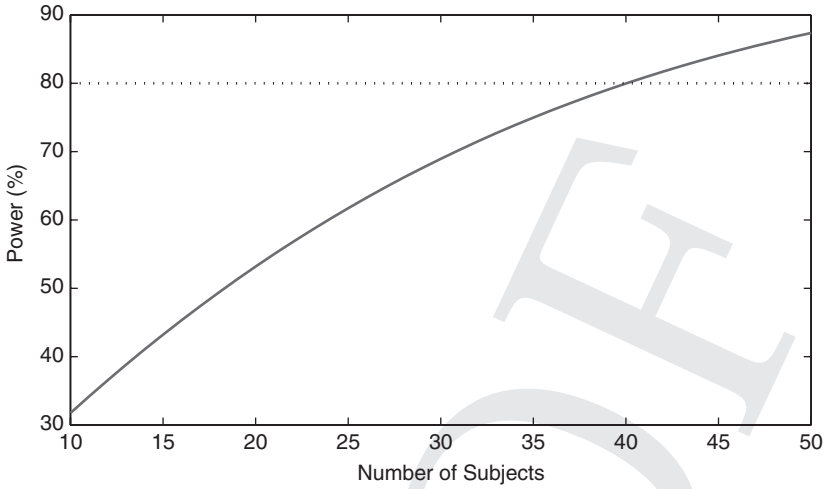

Figure 7.7. Power curve. The curve was generated using an estimated mean effect of 0.8% signal change units with standard deviation of 2% signal change units and a type I error rate of 0.05 using Equation 7.2. Since the graph crosses 80% between 40 and 41 subjects, a sample size of 41 will yield at least 80% power.

point of planning an experiment is to study an effect, yet a power analysis assumes you know the true effect magnitude and standard deviation. Thus, it is a good idea to consider a range of "what if" scenarios: What if true effect is 10% smaller? 20% smaller? What if standard deviation is off, by 10%? and so on. If it appears you still have good power over a range of alternative scenarios, you should be in good shape.

Second, never compute the power of a study post hoc. That is, it is pointless to assess the power of a study that has already been performed: If the effect is there and you detect it, you have 100% power; if it is there and you missed it, you have 0% power. Another way to see this is to consider a series of failed experiments, where the null hypothesis is always true. If $\alpha = 0.05$ is used, we will reject the null hypothesis and declare a significant result on 5% of these tests. Further, say the observed test statistic t is just equal to the statistic threshold, and we use t to compute an effect size and power (t could be higher, but let's be pessimistic). In this case, you will compute the power to be 50% (as it can be inferred from Figure 7.6, if you shift the mean of the alternative distribution left to equal the statistic threshold). Thus, a series of failed experiments will tell you that you have at least 50% power whenever they detect something, when in fact you have 0% power.

Finally, best practice dictates that you base your power analysis on studies that are as similar to your planned study as possible. From those studies, calculate the typical mean and standard deviation of the relevant effect and use independently determined ROIs to avoid circular estimates of effect size (see Box 10.4.2). For more details on the limitations of power analysis, see Hoenig & Heisey (2001).