Statistics Review for fMRI Data Analysis

These notes review the basics of statistical inference, with a special emphasis on concepts necessary for understanding the primary statistical methods for fMRI data analysis. For further information, DeGroot and Schervish *Probability and Statistics* (hereafter *DS)* is a good source. These notes complement separate sets of notes on linear regression analysis and on probability.

1 Random Samples, Realizations and Statistics

Let *Xi,* i = 1,. . ., *n* denote a collection of random variables, mutually independent with a common distribution *F.* We will often say that the *Xi* are *iid* , which is short for "independent and identically distributed." We can model data as *realizations* of these random variables, i.e., observed values Xl,. .. ,*Xn,* for which each *Xi* is an independent observed value from *F.* We sometimes refer to the observed *Xi* as a *random sample* from *F.*

A statistic is any function of the data, for example the sample mean

$$\bar{x} \equiv \sum_{i=1}^{n} x_i/n.$$

If each $Xi \overset{rv}{} F,$ then x is itself a realization of a random variable $X,$ where

$$\bar{X} \equiv \sum_{i=1}^{n} X_i/n.$$

The random variable $\bar{X}$ will have a probability distribution (induced by F), and we can, for example, use $\bar{X}$ to test hypotheses about the *population mean* $E(X_i) = \mu$.

This illustrates some fundamentals underlying much of statistical inference. We model our data as being a random sample from a population. This population is described by a probability distribution, which usually involves unknown parameters. We may or may not have prior beliefs about the values of these parameters. Having observed the data, we perhaps want to make some statement about the parameters, or maybe a prediction regarding future observations. There is some hope of doing these things because functions of the data will themselves have probability distributions, induced by our population model. And we can incorporate prior beliefs regarding parameters, using a basic result of probability theory known as *Bayes Theorem* (*DS*, in particular Section 6.2).

2 Types of Statistical Problems

Statistics addresses questions of the following types (*DS*, Section 6.1). We illustrate each of these with a simple example.

- *Estimation:* What is the mean difference in BOLD activation in a certain brain region between two different cognitive tasks, and what is the uncertainty in this difference?
- *Hypothesis Testing:* Is it plausible that the mean difference in the above example is equal to zero?
- *Experimental Design:* How can one choose a sequence of event types in an event-related fMRI design so as to obtain as much information as possible relevant to a scientific hypothesis of interest?
- *Decision-Making:* This is one area of statistics about which we will have little, if anything, to say in this class. See *DS* p. 326 for a very brief discussion.

3 Sampling Distributions

In this section, we discuss the distributions of some important statistics.

3.1 The Sample Mean

Let X_i be *iid* F, with $E(X_i) = \mu$ and $\mathrm{Var}\,(X_i) = \sigma^2$, and denote our data by $x_1, \ldots, x_n$, a random sample from this population. The sample mean (or average) $\bar{x}$ is perhaps the first statistic that one would calculate. The exact probability distribution of $\bar{X}$ will in general be very complicated, but we can say quite a bit about $\bar{X}$ without knowing its exact distribution. Using results on the expectation and variance of a linear combination of random variables, we see that

$$E(\bar{X}) = \mu, \tag{1}$$

and

$$\mathrm{Var}\,(\bar{X}) = \sigma^2/n. \tag{2}$$

So $\bar{X}$ estimates $E(X)$. One measure of the uncertainty in this estimate is the standard deviation of $\bar{X}$, which equals $\sqrt{\mathrm{Var}\,(\bar{X})} = \sigma/\sqrt{n}$. As n increases, this uncertainty decreases at a rate of order $1/\sqrt{n}$. This is the main reason why experimenters prefer averages over single measurements.

Although we don't know the precise distribution of F, an extremely important result of probability theory states that under very general conditions, so long as n is not "too small," $\bar{X}$ is approximately distributed $\mathrm{N}\,(\mu, \sigma^2/n)$. This result is the "Central Limit Theorem," and it is one reason why simple statistical inference based on Gaussian assumptions can be so useful in wide range of situations (*DS*, Section 5.7).

If the X_i are correlated, this result usually still holds, though one needs to replace n with an *effective sample size* in order for the approximation to be a good one. Intuitively, highly correlated values convey less information than uncorrelated ones (think of the limiting case of perfect correlation, for which all of the x_i will be equal, for any n). We define the effective sample size, for the case of $\bar{X}$ to be

$$n' \equiv \frac{\mathrm{Var}\,(X_i)}{\mathrm{Var}\,(\bar{X})} \tag{3}$$

we will see in this course that this idea can be generalized in obvious ways for more complicated situations.

3.2 The Sample Variance and Standard Deviation

Another statistic often calculated is a measure of the scatter in the data, the sample standard deviation:

$$s \equiv \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

It can be easily shown (*DS*, p. 429) that

$$E(s^2) = \sigma^2; \tag{4}$$

it is for this reason that we divide by $n-1$ in the denominator of s^2. It is *not* true that $E(s) = \sigma$, but usually $E(s) \approx \sigma$, at least for moderate to large n.

The central limit theorem doesn't help much for inference about σ or σ^2. If the X_i are Gaussian, then $s^2 \sim \sigma^2 \chi^2_{n-1}/(n-1)$, where χ^2_ν denotes a *chi-square random variable with ν degrees of freedom* (*DS*, Sections 7.2-7.3). If Y has a χ^2_ν distribution, then

$$f_Y(y) = \frac{y^{\nu/2-1}e^{-y/2}}{2^{\nu/2}\Gamma(\nu/2)},$$

where $\Gamma(t)$ denotes the *gamma function*

$$\Gamma(t) \equiv \int_0^\infty x^{t-1}e^{-t}dx.$$

It's a simple exercise to show that $E(Y) = \nu$ and $\mathrm{Var}\,(Y) = 2\nu$ (*DS*, p. 394, and Section 5.9). In contrast to inference on means, statistical inference on variances based on this χ^2_ν distribution depends rather strongly on the data being at least approximately Gaussian.

4 Estimation

Assume now that our data are a random sample from a population, with X_i distributed according to a probability density $f_X(x|\theta)$. We have modified our notation for a probability density to indicate explicitly that this density depends on a parameter. We would like to estimate θ using some statistic $U = U(X_1, X_2, \ldots, X_n)$. We call U an *estimator* of θ, and we refer to a

realization of U, i.e., $u(x_1, \ldots, x_n)$ as an *estimate* of θ. Often we indicate an estimate of a parameter by putting a "hat" on the quantity being estimated; e.g., we might use the notation $\hat{\theta}$ for both u and U above.

A useful way of characterizing the quality of an estimator is its *mean square error*, or the square root of this quantity: *root mean square error*. Mean square error is the expected squared deviation of an estimate from the thing that it is estimating:

$$\text{MSE}\,(U, \theta) \equiv E[(U - \theta)^2].$$

It is simple to show that it is *always* the case that mean square error can be decomposed into the sum of the variance and squared *bias* of an estimator:

$$\text{MSE}\,(U, \theta) = \text{Var}\,(U) + E[(U - \theta)^2].$$

Bias is the difference between the expected value of an estimator and the value of the thing that it's supposed to be estimating.

This simple decomposition of MSE into squared error and bias can provide insight in diverse real applications. For example, one might observe a noisy signal in time, which one can think of as an estimate of some smooth underlying function. An improved estimate of this underlying function might be obtained by smoothing the signal, since this reduces the variance. But smooth it too much, and though the variance may become quite small, the bias can become very large. Optimal choice of degree of smoothing of signals and images is in general a difficult problem, in which one attempts to arrive at a reasonable compromise between bias and variance, leading to improved estimates.

4.1 Standard Error, Confidence Interval

We usually characterize uncertainty in an estimate by the *standard error*, the standard deviation of the estimator. In the notation of the above example,

$$\text{SE}\,(U) \equiv \sqrt{\text{Var}\,U(X_1, \ldots, X_n)}.$$

For example, the standard error of $\bar{X}$ as an estimator of $E(X_i) = \mu$ is $\sigma/\sqrt{n}$.

Another way of quantifying uncertainty in a parameter estimate is a *confidence interval* (*DS*, Section 7.5). A confidence interval is an interval which, informally, includes the "true" value of the parameter of interest with a

prescribed probability γ, called the *confidence level*. Strictly speaking, a confidence interval is constructed so that if one were to repeatedly obtain random samples, of the same size as the data, from the same population, and calculate a confidence interval for each of these hypothetical future data sets, a proportion γ of these intervals would contain the true value. It is one advantage of a Bayesian formulation of statistics that one can simply say that the interval contains θ with probability γ. Even if one is not taking a Bayesian approach, there is little harm in practice of adopting, at least informally, this simpler interpretation.

4.2 Maximum Likelihood Estimation

The *likelihood* is the conditional probability density of the data given the parameters. If $X_1, \ldots, X_n$ are *iid* $f_X(x|\theta)$, then the likelihood function is

$$L = \prod_{i=1}^{n} f_X(x_i|\theta).$$

A useful approach to finding estimators with reasonably nice properties is to maximize the likelihood function with respect to the parameters. Estimates obtained in this way are called *maximum likelihood estimates* or *MLE*s (*DS*, Section 6.5a).

4.2.1 The MLE of μ for Gaussian Data

If $X_i \sim \mathrm{N}(\mu, \sigma^2)$, then the likelihood is

$$L = (2\pi)^{-n/2} \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Except for a constant which does not involve the parameters, the *log-likelihood* is

$$\ln(L) = \sum_{i=1}^{n} (x_i - \mu)^2/(2\sigma^2).$$

Differentiating with respect to μ and setting this derivative equal to zero, we see that the maximum likelihood estimate is $\hat{\mu} = \bar{x}$. (Note that strictly speaking one needs to show that $\hat{\mu}$ is indeed a maximum [as opposed to a minimum or a saddle point]. But this is obvious for this example.)

Maximum likelihood estimates are asymptotically Gaussian, with variance (or, in the multi-parameter case, covariance matrix) determined from the curvature of the likelihood at the MLE (the *Fisher Information*; see *DS* Sections 6.6, 7.8, and pp. 442-443 for details).

5 Hypothesis Testing

The set-up for statistical hypothesis testing takes some getting used to: it's probably fair to say that it's not very intuitive. A *hypothesis* in statistics is a statement in terms of the parameters of a model. Hypotheses usually come in pairs: a *null hypothesis* which you hope to reject, and an *alternative hypothesis* which you usually hope is true (*DS*, Chapter 8).

For example, suppose that you've done a cognitive experiment to test the (scientific) hypothesis that BOLD activity in some region of the brain is higher during a certain cognitive task than it is during rest. If the parameter in your model which represents the expected value of the difference in BOLD response for this brain region is δ, then your null hypothesis might be

$$H_0 : \delta \leq 0,$$

with corresponding alternative hypothesis

$$H_1 : \delta > 0.$$

From your data, you calculate an estimate of δ and divide this estimate by its standard error to form a *test statistic*

$$T = \frac{\hat{\delta}}{\text{SE}\,(\hat{\delta})}.$$

To perform a hypothesis test, one begins by tentatively assuming that H_0 is true. Then one checks if there is enough evidence in the data in support of the alternative H_1 to enable one to safely reject H_0 as implausible. In our example, the larger T is, the stronger the evidence in the data in support of the alternative hypothesis. If T exceeds some constant, say c, then one *rejects* H_0 in favor of H_1. Otherwise, one *fails to reject* H_0.

5.1 Type I and Type II Errors

The rub, of course, is in the choice of c. One can make two kinds of errors in a hypothesis test. A *Type I* error occurs when one rejects H_0 incorrectly; a *Type II* error occurs when one rejects H_1 incorrectly. The probabilities of these errors are universally denoted α and β. The probability of a Type I error (α) is also called the *significance level* or *size* of the test; one minus the probability of a Type II error ($1 - \beta$) is called the *power* of the test.

By making c large enough, you can make it as hard as you like for H_0 to be rejected; i.e., you can make α arbitrarily small. But as you change c so as to decrease α, you will be *increasing* β, since you will be requiring larger and larger values of T in order to reject H_0. So one cannot make both α and β arbitrarily small. The approach which is generally taken is to fix α at a small value, conventionally usually taken to be 0.05, but sometimes 0.10 or 0.01. Having done so, one can then determine $c = c(\alpha)$. The probability of a Type II error will depend on the state of nature under H_1, i.e., the closer the true δ is to zero, the closer together H_0 and H_1 are, the larger β will be (equivalently, the smaller the power will be) for a given sample size. (In *experimental design*, one selects a small significance level (say, $\alpha = 0.05$), a reasonably high power (say, $1 - \beta = 0.80$), and a plausible value of δ which you'd like to be able to detect with this power and statistical significance. One can then solve for the required sample size.)

6 Inference on a Population Mean: the t-Distribution

We now illustrate concepts introduced in the above sections for a simple example. Although this example is very simple, it is closely related to the problems of inference in the general linear model, which, as we will learn in this course, are central to many conventional approaches to fMRI data analysis.

Assume that our data are a random sample from a Gaussian distribution with unknown mean and variance. We estimate μ and σ^2 by $\bar{x}$ and s^2, the sample mean and variance, respectively. We know that (to a good approximation, even without the Gaussian assumption) that

$$\bar{X} \sim \mathrm{N}\,(\mu, \sigma^2),$$

and (here we cannot be so cavalier about the Gaussian assumption)

$$S^2 \sim \sigma^2 \chi^2_{n-1}/(n-1).$$

It can be shown that

$$T_\nu = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

has a t-distribution with $\nu = n-1$ degrees of freedom (*DS*, Sections 7.4 and 8.5). The density of T_ν is

$$f_T(t|\nu) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

The corresponding cumulative distribution function is tabulated in all elementary statistics textbooks. The t-density is bell-shaped, with tails which are heavier than those of a Gaussian. The t-density can be reasonably well approximated by a Gaussian, however, for many applications, once n exceeds about 20 or so.

When the data come from a non-Gaussian population, the statistic T_ν still very often has an *approximate* t-distribution. This is an important fact, which is often under-emphasized in elementary texts. It is one of the main reasons why tests and estimates based on the t-distribution are so useful in real applications, where idealized assumptions often do not hold. If the data are correlated, however, the degrees of freedom of the approximate t-distribution may have to be adjusted, using estimated effective degrees of freedom, in order to improve the approximation.

6.1 Confidence Interval for μ (*DS*, pp.410-411)

From a table, we determine t_ν, the value such that the probability that a T_ν random variable exceeds t_ν is 0.025. Since the T_ν density is symmetric about 0, we have that

$$\Pr\left(-t_\nu \leq \sqrt{n}\frac{\bar{X} - \mu}{S} \leq t_\nu\right) = 0.95.$$

This can be rewritten as

$$\Pr\left(\bar{X} - t_\nu S/\sqrt{n} \leq \mu \leq \bar{X} + t_\nu S/\sqrt{n}\right) = 0.95.$$

Hence, $\bar{X} \pm t_\nu S/\sqrt{n}$ is a two-sided 95% confidence interval for μ.

6.1.1 Test of $H_0 : \mu = 0$

To test the hypothesis

$$H_0 : \mu = 0$$

against the alternative

$$H_1 : \mu \neq 0$$

We form the test statistic

$$t = \frac{\bar{x}}{s/\sqrt{n}},$$

and reject H_0 when $|t|$ is greater than t_ν (you should convince yourself that this is equivalent to rejecting H_0 when the 95% confidence interval for μ does not contain 0). This is a two-sided hypothesis test is at the $\alpha = 0.05$ significance level.

7 The Generalized Linear Model

A separate set of notes (A Review of Statistics Part 3: The Generalized Linear Model, Emery N. Brown) reviews the basics of the linear model, otherwise known as *linear regression analysis* (see also *DS*, Chapter 10). One aspect of the linear model which is not covered in these notes is the testing of hypotheses on linear contrasts in parameters. This topic is central to a commonly-used approach to fMRI data analysis, so we review it briefly here.

After motion correction, de-trending, and other data-preparation steps, imagine that the time course data for each voxel can be described by a linear model of the form

$$Y = X\beta + \epsilon,$$

where the BOLD signal vector, Y, is $n \times 1$, the (known) design matrix X is $n \times p$, and the unknown coefficient vector β is $p \times 1$. We assume that the residual error vector, ϵ, is Gaussian with zero mean. For simplicity, in the present discussion we will also assume that the ϵ_i are *iid*. The coefficient vector, β will vary from voxel to voxel, but it is estimated separately for each voxel.

The method of least squares, which is also the maximum likelihood method for this problem, estimates β by minimizing the sum of squared residuals

$$Q \equiv (Y - X\beta)^T (Y - X\beta),$$

Differentiating with respect to the vector β, and setting the derivative equal to zero, we see that

$$X^T(Y - X\hat{\beta}) = 0 \Rightarrow \hat{\beta} = (X^T X)^- X^T Y, \tag{5}$$

where $(X^T X)^-$ is $(X^T X)^{-1}$ if X is of full rank; otherwise the Moore-Penrose generalized inverse is used. (When X is of less then full rank, there are infinitely many $\hat{\beta}$ estimates which minimize the sum of squares, although all of these estimates lead to the same predicted values $\hat{Y}$. The Moore-Penrose inverse leads to the unique $\hat{\beta}$ vector of shortest length.)

Using the formula for the variance of a linear transformation of random variables, we see that

$$\begin{aligned} \text{Var}\,(\hat{\beta}) &= \left[(X^T X)^- X^T\right] \text{Var}\,(Y) \left[(X^T X)^- X^T\right]^T \\ &= (X^T X)^- X^T (\sigma^2 I) X (X^T X)^- \\ &= \sigma^2 (X^T X)^- (X^T X)(X^T X)^- \\ &= \sigma^2 (X^T X)^-. \end{aligned} \tag{6}$$

(The last step follows immediately from the definition of the Moore-Penrose inverse. If you're not familiar with genralized inverses, you can easily redo the above calculation for the simpler case where X is of full rank, and hence $(X^T X)^{-1}$ exists.)

Assume that this study is comparing three different tasks with baseline fixation, and that we are interested in determining those voxels for which at least one of these three activities differs from the fixation control. Ignoring for simplicity any columns of the design matrix which might be associated with "nuisance" effects such as motion correction, we have that X is a $n \times 4$ matrix with rows $[1, 0, 0, 0]$ for fixation, $[0, 1, 0, 0]$ for task #1, $[0, 0, 1, 0]$ for task #2, and $[0, 0, 0, 1]$ for task #3. We write the β vector as $[\beta_0, \beta_1, \beta_2, \beta_3]^T$. We would like to test the null hypothesis that

$$H_0 : \beta_0 = \beta_1 \text{ and } \beta_0 = \beta_2 \text{ and } \beta_0 = \beta_3$$

against the alternative that at least one of these 3 equations doesn't hold. These constraints on the parameters under the null hypothesis can be written in matrix form using a *contrast matrix*

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \equiv C\beta = 0.$$

To test this hypothesis, we obtain the usual (unconstrained) least squares coefficient estimate $\hat{\beta}$, and the *constrained* estimate $\hat{\hat{\beta}}$, for which $C\hat{\hat{\beta}} = 0$. One simple way to find $\hat{\hat{\beta}}$ is to use $C\beta = 0$ to substitute for some of the βs in terms of others. There is also an explicit formula for $\hat{\hat{\beta}}$. For X of full rank, Seber (1977, pp. 84-86) shows that

$$\hat{\hat{\beta}} = \hat{\beta} + (X^T X)^{-1} C^T \left[C(X^T X)^{-1} C^T \right]^{-1} C\hat{\beta}. \tag{7}$$

Let the number of parameters be p, and the number of contrasts $q < p$. The error variance σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p},$$

which is distributed as $\sigma^2 \chi^2_{n-p}/(n-p)$, indepdendently of $\hat{\beta}$. It is *always* the case that

$$Q_1 = (y - X\hat{\beta})^T (y - X\hat{\beta})$$

is less than

$$Q_0 = (y - X\hat{\hat{\beta}})^T (y - X\hat{\hat{\beta}}),$$

since the $\hat{\beta}$ are unconstrained. It turns out tha the increase in sums of squares due to relaxing the constraints $C\beta = 0$, divided by the number of contrasts q, provides an *independent* estimate of σ^2, when H_0 is *true*. Specifically,

$$\hat{\hat{\sigma}}^2 \equiv \frac{Q_1 - Q_0}{q} \sim \sigma^2 \chi^2_q / q.$$

If H_0 is true, then the ratio

$$F \equiv \frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2} = \frac{(Q_1 - Q_0)/q}{Q_0/(n-p)}$$

will be distributed according to an F-distribution with q and $n-p$ degrees of freedom. (When $q = 1$, F reduces to the square of a t random variable with $n - p$ degrees of freedom; thus the F-test for a single contrast is equivalent to a t-test.)

The F distribution is tabulated in many statistics textbooks. If H_0 is true, then the numerator and denominator are both estimating σ^2, so F will tend to be near 1, or smaller than 1. It can be shown that under H_1 F will

be larger than under H_0. Hence one proceeds as usual for hypothesis tests: one tentatively assumes that H_0 is true, one calculates F, and compares this calculated value with an appropriate tabulated upper percentile for an F random variable with q and $n-p$ degrees of freedom. If the calculated value is sufficiently large, one rejects H_0, and concludes that *at least one* of the contrasts is nonzero.

7.1 Satterthwaite Approximation

It is convienient when quadratic forms, such as $\hat{\sigma}^2$ and $\hat{\hat{\sigma}}^2$ above, have distributions proportional to χ^2 random variables, since one can then form at least approximate t and F statistics for various tests of hypotheses and confidence intervals.

However, there are often situations where the distribution of a quadratic form is more complicated. In particular, this happens in the case of the linear model when the noise terms ϵ_i either have unequal variances, are not Gaussian, or are correlated. Often one doesn't even have a closed-form expression for the distribution of a quadratic form. It is common in such situations to approximate the distribution of a quadratic form Q by $a\chi^2_b$, where a and b are constants to be determined. We know the mean and variance of a quadratic form, at least for the Gaussian case (see the probability review notes), and we know that $E(\chi^2_b) = b$, and $\mathrm{Var}\,(\chi^2_b) = 2b$. So a simple thing to do is to equate the mean and variance of Q to the mean and variance of $a\chi^2_b$, and solve for a and b.

Denote the mean and variance of a quadratic form Q by μ_Q and σ^2_Q, repectively. We approximate Q by $a\chi^2_b$, where a and b are determined from

$$\mu_Q = ab$$

and

$$\sigma^2_Q = 2a^2b.$$

This approximation was proposed by Satterthwaite in a statistics journal in the early 1950's, so it's usually referred to as the *Satterthwaite approximation* (at least by statisticians). One use of this approximation in fMRI data analysis is to adjust the degrees of freedom of t-statistics in the generalized linear model for temporal correlation of voxel time-course data.

7.2 Further Reading

For a discussion of elementary results in linear regression, see *DS*, Chapter 10. The F-distribution is introduced in *DS*, Section 8.7. The proof that the statistic F is distributed according to the indicated F-distribution is beyond the scope of most elementary statistics texts. One notable exception is Hoel, Port and Stone, 1971, Chapter 5. For a more detailed treatment, see any text on the theory of linear models in statistics, such as Seber (1977), Chapter 3.

8 The Multiple Comparisons Problem

It is common for hypothesis tests to be performed at the $\alpha = 0.05$ significance level. This means that one would only reject H_0 when it's true one time out of 20. A standard approach to fMRI data analysis is to perform a hypothesis test of the form introduced in the previous sections for *each voxel*. With several thousand voxels in a typical fMRI image, the "only" in the previous section no longer applies!

In statistics, this is referred to as the problem of *multiple comparisons*. A standard approach to resolving this difficulty is to reduce the nominal α level to something sufficiently small that one can claim an overall significance level of 0.05 or 0.01. How one does this reduction in α depends on what assumptions one is willing to make about the correlation among the voxel time courses: that is, there's more than one way to do it.

If one assumes that the voxels are independent, the the *Bonferoni correction* is appropriate. In this approach, one would use the nominal significance level of α/N for each voxel, where N is the number of voxels, in order to assure an overall significance level of α.

However, N is very large, and the voxel time courses tend to be spatially correlated. So the Bonferoni approach is usually not used in fMRI data analysis. It is usually too conservative. Instead, there are at least three approaches that one can take. The most widely adopted approach at present is to select a threshold for the statistics at each voxel based on the assumption that these voxels, perhaps after some spatial smoothing, can be regarded as a realization of a *Gaussian random field* (e.g., Worsley (2001)). This will be less conservative than the Bonferoni approach, since it makes some allowance for spatial correlation. There is some controversy, however, over the validity of the Gaussian field assumption for fMRI data. A second approach is to do a

Bayesian or Empirical Bayes analysis (Genovese (2000), Friston (2002)). For such an analysis, one can avoid the hypothesis testing problem altogether. A third approach is to control the *false discovery rate* instead of the overall significance level. The FDR is the expected proportion of false rejections of H_0 in a volume; this approach can provide less conservative results than the Bonferoni approach, without requiring assuming a Gaussian random field as a model for the data (Geovese et al., 2002).

9 Need More Review?

With the exception of much of Sections 7 and 8 of these "Statistics" notes, and Section 8 of the "Probability" notes, this material should be review. But the presentation here is necessarily concise. For a more detailed discussion of the basic concepts of probability and statistics, you can of course consult any elementary textbook. In particular, we've put a copy of *DS* on reserve. Pages of this text which are particularly useful for review include 268-277, 282-283, 355-369, 393-396, 404-412, 435-444, 449-462, 485-493, 506-508, and 599-664.

References

1 DeGroot, M.H. and Schervish, M.J. (2002). *Probability and Statistics*, 3rd Edition, Addison-Wesley.

2 Friston, K.J. (2002). "Bayesian Estimation of Dynamical Systems: An Application to fMRI," *NeuroImage*, 16, 513-530.

3 Genovese C.R. (2000). "A Bayesian Time-Course Model for Functional Magnetic Resonance Imaging Data," *Journal of the American Statistical Association*, 95, 451.

4 Genovese C.R.; Lazar, N.A.; Nichols, T. (2002). "Thresholding in Functional Neuroimagining Using the False Discovery Rate," 15, 870-878.

5 Hoel, P.G.; Port S.C.; Stone C.J. (1971). *Introduction to Statistical Theory*, Houghton-Mifflin, Boston.

6 Seber, G.A.F.(1977). *Linear Regression Analysis*. John Wiley and Sons, New York.

7 Worsley, K. (2001). "Statistical Analysis of Activation Images," in *Functional MRI: An Introduction to Methods*, Oxford University Press.