HST.583: Functional Magnetic Resonance Imaging: Data Acquisition and Analysis
Harvard-MIT Division of Health Sciences and Technology
Dr. Mark Vangel

Selected Topics in Statistics for fMRI Data Analysis

Nuclear Magnetic Resonance Center
Massachusetts General Hospital
Charlestown, MA

HST 583

Outline

I. Adjusting for Multiple Comparisons

II. Modelling Data from Multiple Subjects

III. Some Thoughts on Model Validation

Multiple Comparisons

Ia. Bonferroni Approximation

Ib. Gaussian Random Fields

Ic. False Discovery Rate

A Hypothetical Hypothesis Test

Consider a hypothesis test for which you obtain the t-statistic

$$T = 4.62,$$

with 50 degrees of freedom. The corresponding p-value is

$$1 - \Pr(-4.62 \leq T_{50} \leq 4.62) = 0.000027.$$

Is this necessarily cause for celebration?

The Rest of the Story . . .

The t-statistic on the previous slide was obtained by choosing the maximum of $64 \times 64 \times 16 = 65,536$ random draws from the *null distribution* of the test statistic (i.e., the T_{50} distribution).

So you might typically expect to see a t-statistic this large or larger in a typical fMRI volume, even if what you're imaging is a bottle of water.

We need to adjust p-values for the number of tests performed, a process which statisticians call *adjusting for multiple comparisons*.

An Illustrative Example

In order to illustrate many of the basic ideas, it is sufficient to consider an example of confidence intervals (or hypothesis tests) on just two parameters.

Consider the simple linear regression model

$$y_i = \delta + \beta(x_i - \bar{x}) + e_i,$$

where $x_i = 0, 10, 20, \ldots, 100$, $\delta = 0$, $\beta = 1$, and the $e_i \sim \mathrm{N}(0, 10^2)$.

We are interested in testing the null hypothesis

$$H_0 : \delta = 0 \text{ and } \beta = 1,$$

against the alternative

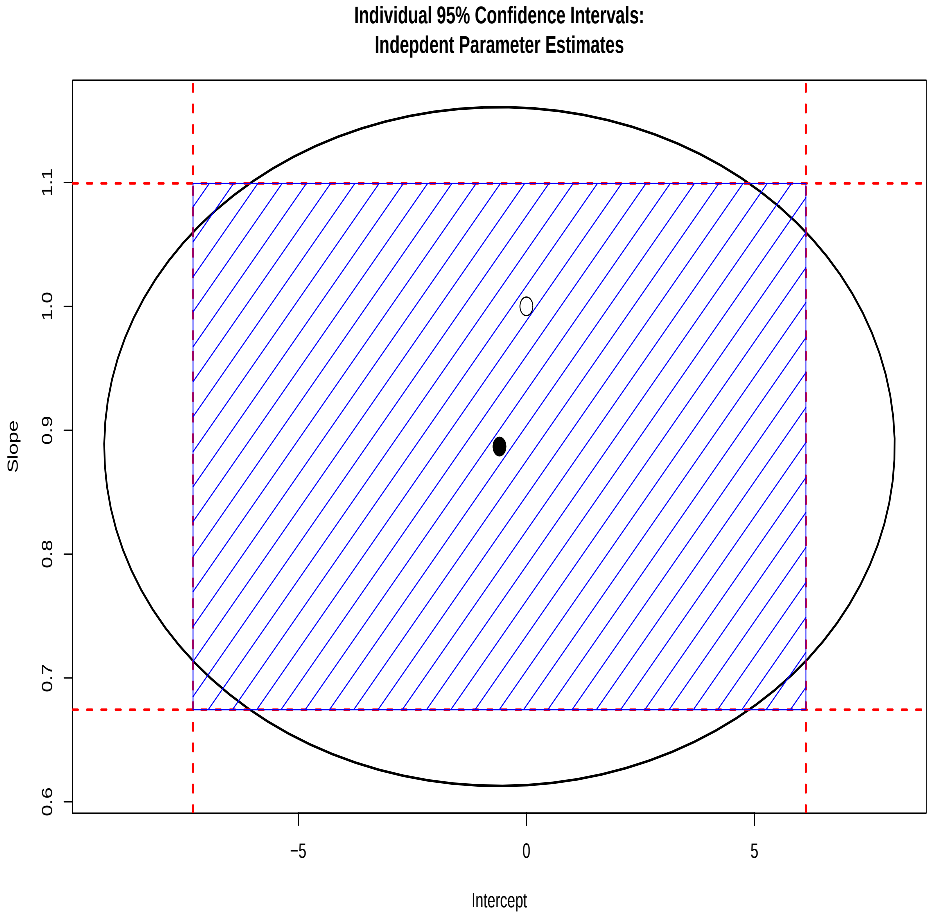
$$H_1 : \delta \neq 0 \text{ or } \beta \neq 1,$$

at the 0.05 significance level.

A joint 95% confidence region for (δ, β) would provide a critical region for this test.

(**Aside:** Note that the vectors $[1, 1, \ldots, 1]^T$ and $[x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}]^T$ are orthogonal.)

Confidence Region



Comments

- Note that the exact 95% elliptical contour is not contained withing the box corresponding to the joint confidence intervals.

- If we reject H_0 when $(\widehat{\alpha}, \widehat{\beta})$ falls outside of the rectangle, then we will reject H_0 too often.

- In other words, the corresponding hypothesis test overstates the significance of results.

Comments (Cont'd)

- Over repeated future data, the probability that an interval covers its parameter is $\alpha_* = 0.95$. Since the model has been set up so the the estimates are independent, the actual probability of rejecting H_0 for the pair of confidence intervals

  $$\begin{aligned}&\alpha = \Pr(|T_1| \geq t_1 \text{ or } |T_2| \geq t_2)\\ &= 1 - (1 - \alpha_*)^2 = 1 - (1 - 0.05)^2 = 0.0975.\end{aligned}$$

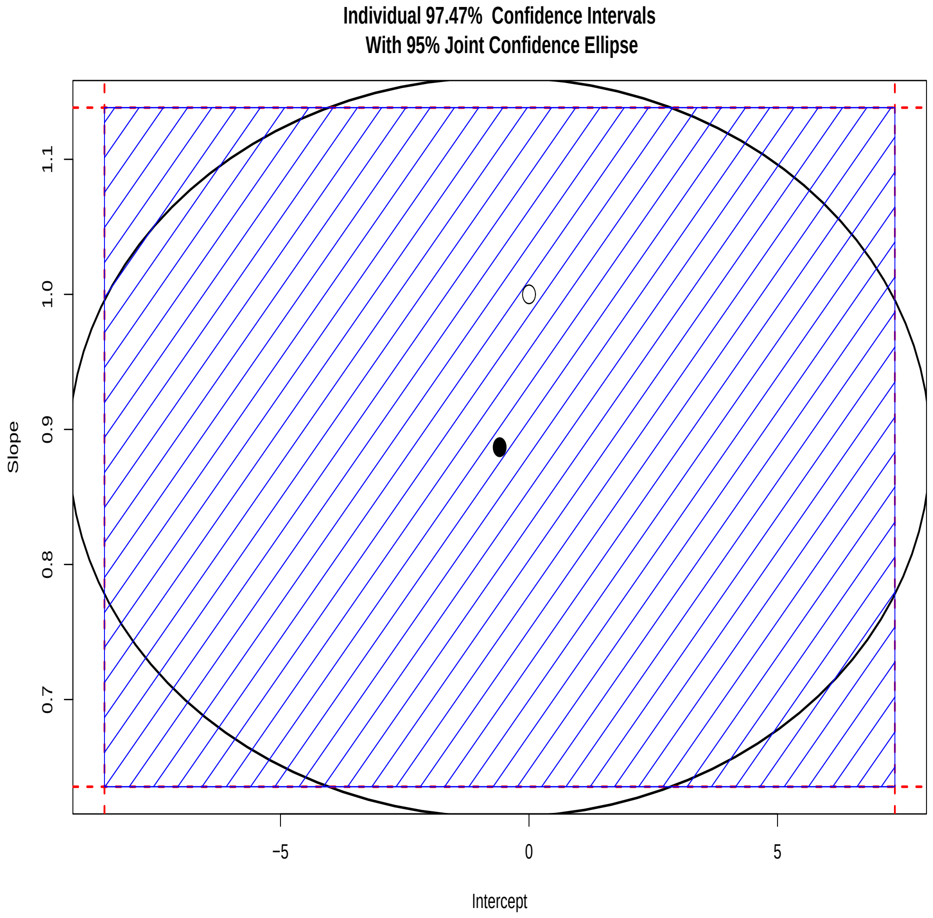
- Working backwards, if we choose α_* to be

  $$\alpha_* = 1 - \sqrt{1 - \alpha} \approx \alpha/2,$$

  then we will achieve our goal of an overall significance level of α.

- This approach achieves the desired significance if the test statistics are *independent*, but is conservative if the test statistics are *dependent*.

Bonferroni Intervals with 95% Confidence Ellipse



Bonferroni Correction

- **The Setup:** We have k independent test statistics $T_1, \ldots, T_k$, corresponding to parameters $\beta_1, \ldots, \beta_k$, respectively.

- For each test statistic, we reject the null hypothesis $H_i : \beta_i = 0$ when $|T_i| \geq t_i$, for constants $t_1, \ldots, t_k$.

- The probability of rejecting
$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
against the alternative that H_0 is not true is
$$\begin{aligned} \alpha &= \Pr_0(|T_1| \geq t_1 \text{ or } |T_2| \geq t_2 \text{ or } \ldots |T_k| \geq t_k) = \\ &= 1 - \prod_{i=1}^{k} \Pr(|T_i| \leq t_i) = 1 - (1 - \alpha_*)^k. \end{aligned}$$

- Hence, we choose
$$\alpha_* = 1 - (1 - \alpha)^{1/k} \approx 1 - (1 - \alpha/k) = \alpha/k.$$

Illustrative Example Revisited: An Alternative Parameterization

Next we see what happens in our simple linear regression example if we don't subtract off the mean of the xs:

$$y_i = \tilde{\delta} + \beta x_i + e_i,$$

where $x_i = 0, 10, 20, \ldots, 100$, $\delta = 0$, $\beta = 1$, and the $e_i \sim \mathrm{N}(0, 10^2)$. To relate this to the previous parameterization, note that

$$\tilde{\delta} = \delta - \bar{x}\beta.$$

We are interested in testing the null hypothesis

$$H_0 : \tilde{\delta} = -\bar{x} \text{ and } \beta = 1,$$

against the alternative

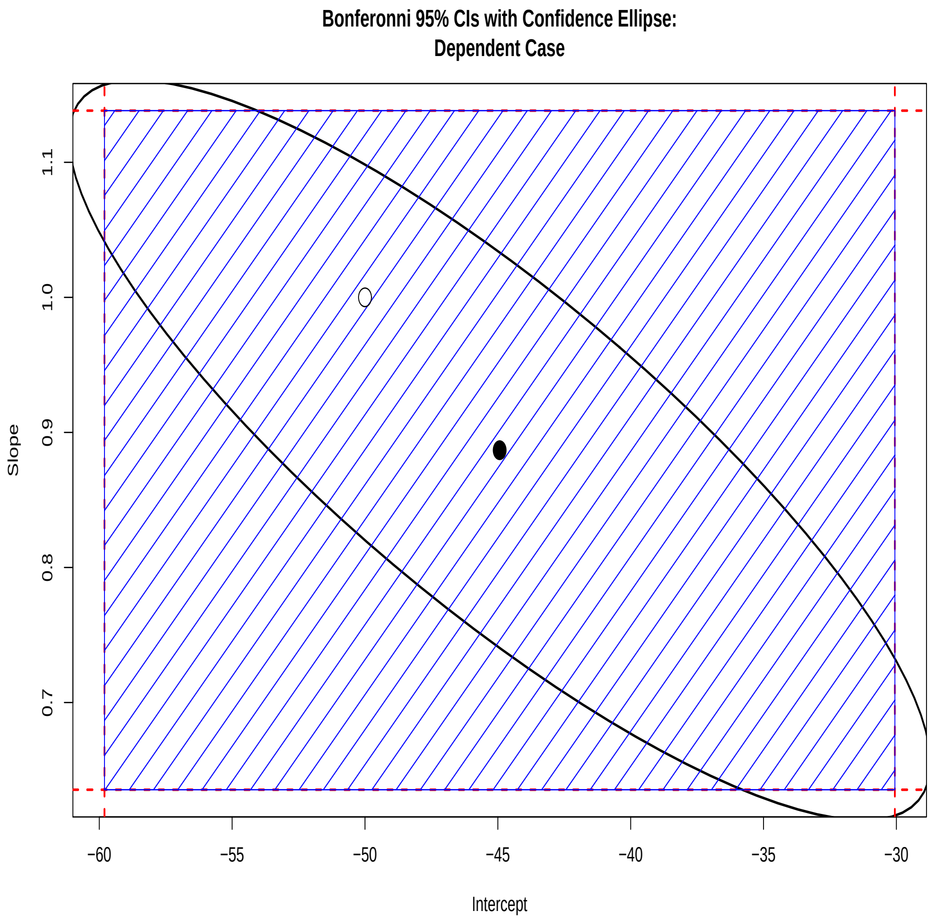
$$H_1 : \tilde{\delta} \neq -\bar{x} \text{ or } \beta \neq 1,$$

at the 0.05 significance level.

A joint 95% confidence region for $(\tilde{\delta}, \beta)$ would provide a critical region for this test.

(**Aside:** Note that the vectors $[1, 1, \ldots, 1]^T$ and $[x_1, x_2, \ldots, x_n]^T$ are *not* orthogonal! Consequently, the t-tests for $\tilde{\delta}$ and β will not be independent.)

Confidence Region for a Dependent Example, With Bonferroni Intervals



Conclusions on Bonferroni

- The Bonferroni correction for multiple comparisons is simple, and it can be reasonable to use if the test statistics are nearly spatially independent.

- When the statistics are spatially correlated, the Bonferroni approach can still be used, but will be conservative.

Gaussian Random Field

- A Gaussian random field is a stationary Gaussian stochastic process, usually in 2 or 3 dimensions.

- The one-dimensional case of GRF is Brownian motion (formally, a *Weiner process*).

- Unsmoothed BOLD activity is not well approximated as a GRF, so spatial smoothing is generally done if one is to use GRF theory.

- Smoothing is averaging, and averages of (almost) arbitrary random variables are approximately Gaussian. This is the essence of the *Central Limit Theorem*.

Euler Characteristic

- If one thresholds a continuous GRF, the the *Euler Characteristic* is

$$\text{EC} = (\#\ \text{Blobs}) - (\#\ \text{Holes}),$$

- if the threshold is sufficiently high, then this will essentially become the (# Blobs).

- If the threshold is higher still, then the EC will likely be zero or 1.

- If we threshold high enough, then we might be able to assume, at an appropriate significance level, that all blobs are due to activity.

Expected EC

- By definition,

$$E(\text{EC}) = \sum_k k \, \text{Pr}(\text{EC} = k)$$

- For high thresholds, the probability of more than one blob under H_0 is negligible, and we have

$$E(\text{EC}) \approx \text{Pr}(\text{EC} = 1)$$

- For large u, $E(\text{EC})$ will approximate

$$E(\text{EC}) \approx \text{Pr}(\max_i T_i > u).$$

- We can either attempt to approximate this expectation for a choice of u (adjusted p-value), or else select u so that $E(\text{EC})$ equals, say, 0.05 (adjusted hypothesis test).

Corrected p-Values via $E(\mathrm{EC})$

- We can obtain p-values by using

$$\Pr(\max_i T_i > u) \approx E(\mathrm{EC}_u)$$
$$= \frac{R(u^2-1)e^{-u^2/2}}{4\pi^2(2\log(2))^{3/2}}$$

- Where R is the number of *Resolution Elements*, defined to be a unit search volume, in terms of the FWHM of the kernel used for spatial smoothing.

- (So *now* you know why SPM requires that you do spatial smoothing!)

Resolution Elements

$$R = \frac{S}{f_x f_y f_z},$$

where

- S is the search volume, in *mm*³,

- and f_x, f_y, f_z are the FWHMs of the Gaussian spatial kernel in each coordinate direction, in *mm*.

Conclusions on Gaussian Random Fields

- The GRF approach to multiple comparisons is appropriate if the noise in the data approximates a Gaussian random field.

- This is generally not the case. The approach taken is to impose known Gaussian spatial correlation structure on the data through smoothing.

- If one is willing to accept such smoothing, then the GRF approach is reasonable.

False Discovery Rate

- The Bonferroni approach ensures that the probability of *in*correctly declaring *any* voxel active is small. If any voxels "survive," one can reasonably expect that *each one* is truly active.

- An alternative approach is to keep the *proportion* of voxels incorrectly declared active small. Among those voxels declared active, a predetermined proportion (e.g., 0.05), *on average*, will be declared active in error ("false discoveries").

Implementing FDR

- Order the N p-values from smallest to largest:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}.$$

- Declare as active voxels corresponding to ordered p-values for which

$$p_{(i)} \leq q\,c\,i/N,$$

where q is the selected FDR.

- The choice of c depends on the assumed correlation structure for the test statistics.

Values for c

- Two choices for c have been suggested in the literature

- For independent tests, or tests based on data for which the noise is Gaussian with non-negative correlation across voxels, use $c = 1$.

- For arbitrary correlation structure in the noise, use $c = 1/(\log(N) + \gamma)$, where $\gamma \doteq 0.577$ is Euler's constant.

A Simulated Example

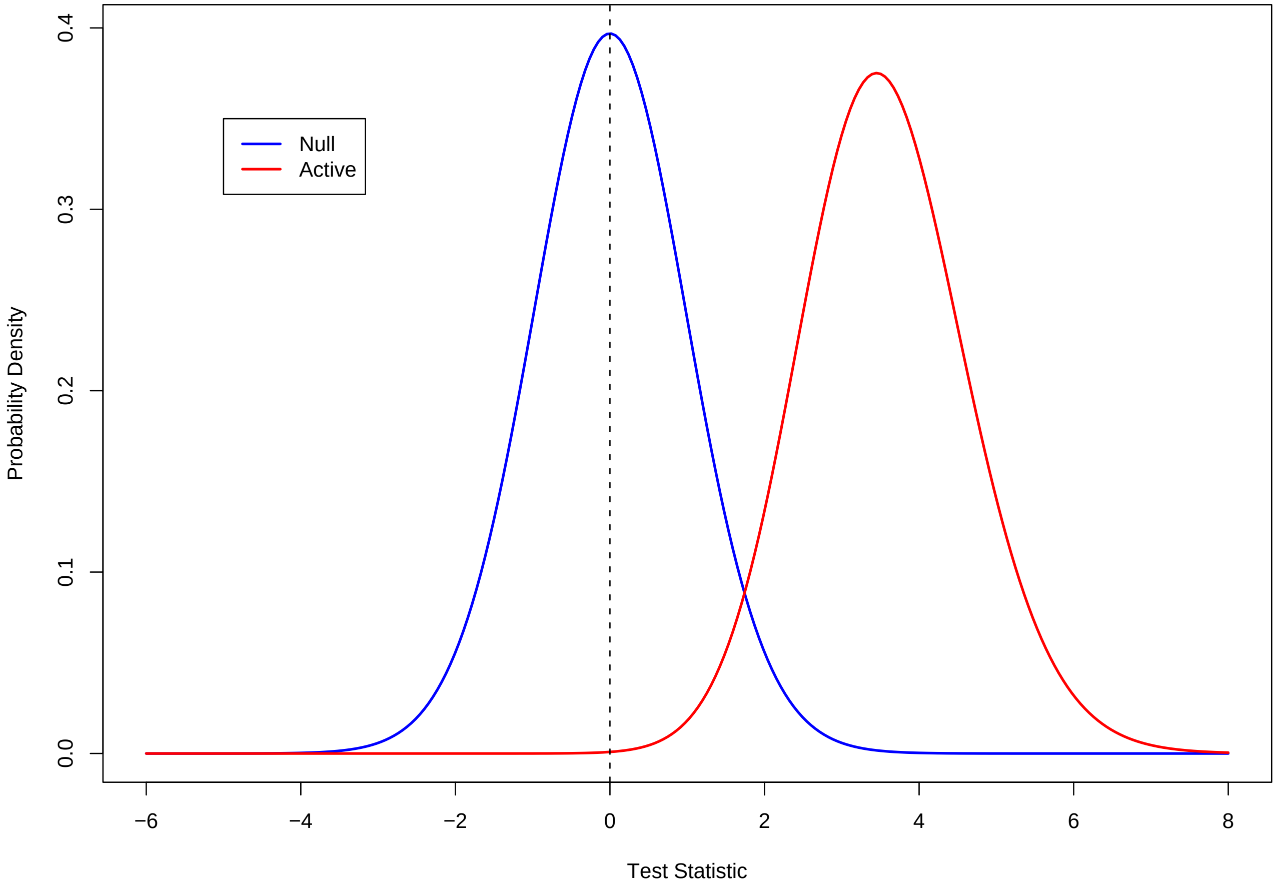
- **Number of Voxels:**
  $N = 64 \times 64 \times 16 = 65,536$

- **Number of Active Voxels:**
  $N_1 = 0.02N = 1,335$

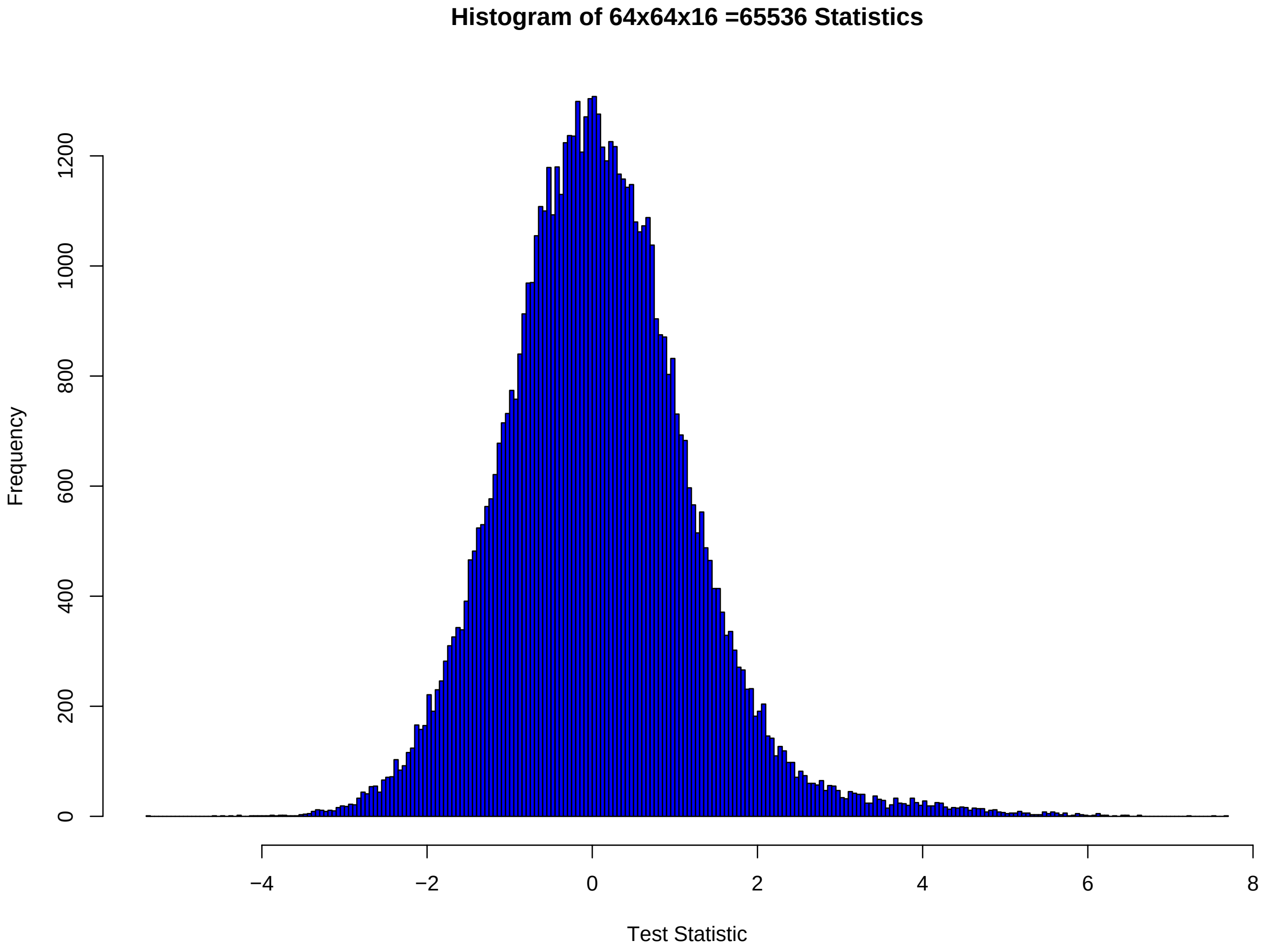
- "Inactive" statistics independently distributed t_{50}.

- "Active" statistics independently distributed *noncentral-t*, $t_{50}(\delta)$, where $\delta = 3.5$.

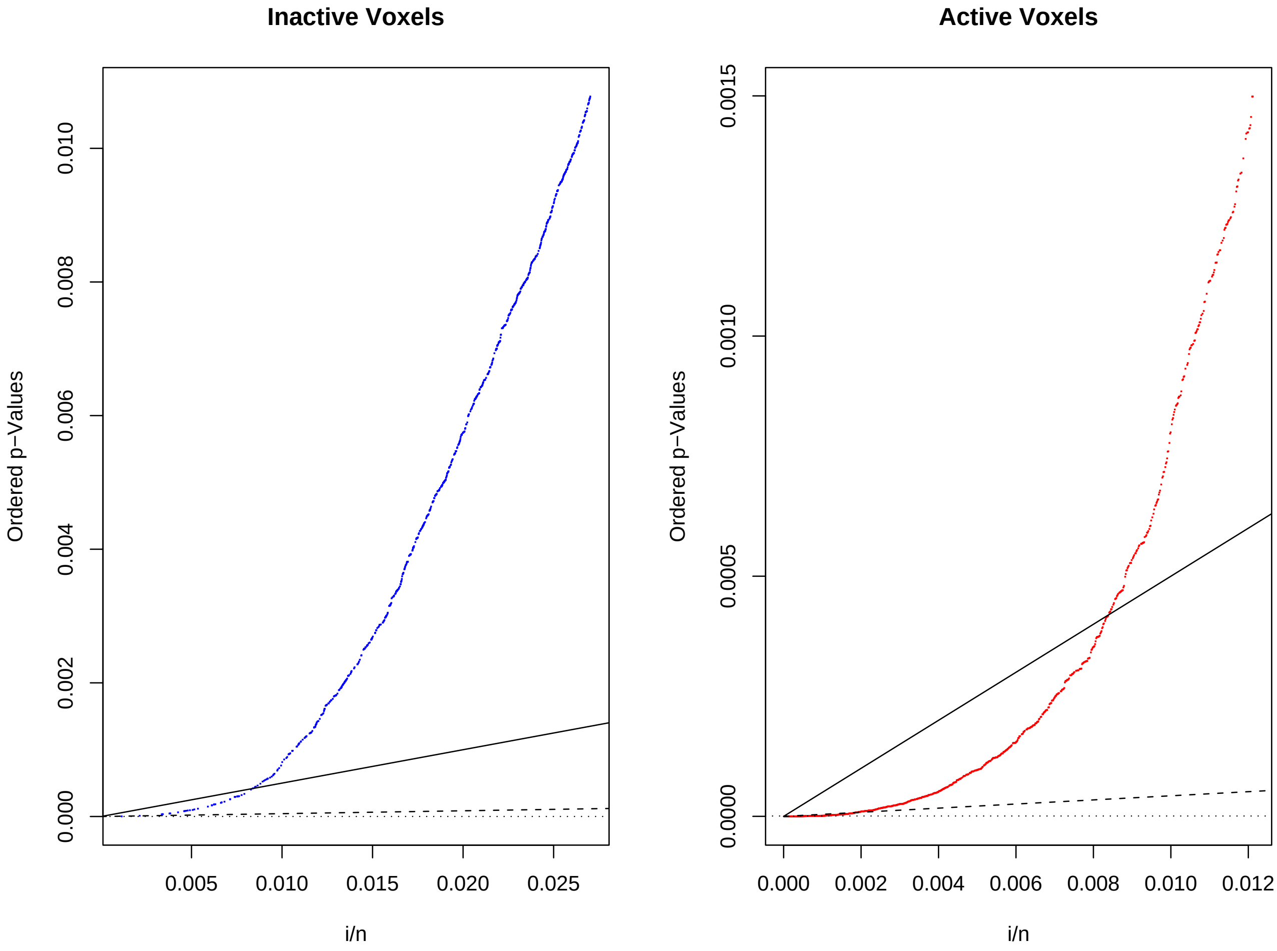
Densities for Active and Inactive Voxel Statistics



Histogram of the Voxel Statistics



Graphical Illustration of Results



Inactive Voxels

Active Voxels

Simulation Results

- FDR $= 35/549 \doteq 0.064$, $c = 1$:

	Discovered	
	Yes	No
Correct	514	64,166
Error	35	821

- FDR $= 1/123 \doteq 0.008$, $c = 1/(\log(N)+\gamma)$:

	Discovered	
	Yes	No
Correct	122	64,200
Error	1	1213

- Bonferroni (FDR $= 0$), $p = .05/N = 7.6 \times 10^{-7}$:

	Discovered	
	Yes	No
Correct	44	64,201
Error	0	1291

Conclusions on FDR

- FDR differs philosophically from both Bonferroni and GRF. These methods have a small probability of *any* false positive. FDR, on the other hand, seeks to control the *proportion* of false discoveries.

- Using FDR with $c = 1/(\log(N) + \gamma)$ is safe, but experience suggests that it can be very conservative. The choice $c = 1$ is probably better for general use.

Analyses for Groups of Subjects

IIa. Fixed Effects

IIb. Random Effects

IIc. Conjunction Analysis

Group Analyses

- We next consider approaches to data analyses which involve more than one subject.

- The first difficulty that one has to address in these situations is warping each subjects data onto a common template, such as Talaraich coordinates.

- This process can easily introduce and difficulties and distortions of its own, but these are beyond the scope of the present discussion.

Fixed Effects Analyses

- It is conceivable that one might want to make inference for only the subjects at hand, without any desire to extrapolate to a larger population.

- This might be the case for clinical applications of fMRI, for example, where the objective is to understand the subjects – patients – who are being studied or treated.

- *Fixed effects* models should be used in such cases.

- But since fMRI is presently a research tool, fixed effects analyses are usually less appropriate than *random effects* analyses, in which one is concerned with inferences valid for a population, or equivalently, for the "next" subject which one might obtain.

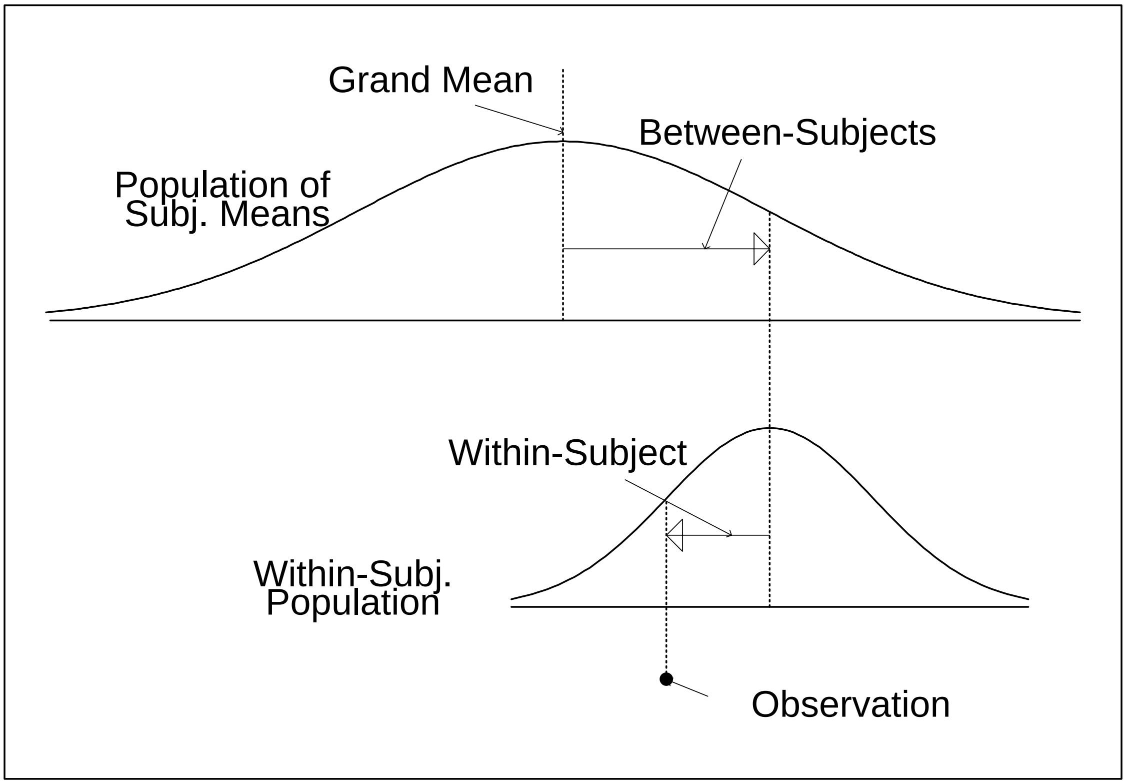
Fixed vs. Random Effects

- Assume that several machines are used in a production environment. To fix ideas, let's say these machines are for DNA sequencing.

- If I have several of these machines in my lab, I would presumably be interested in quantifying the relative performance of each of them. **Fixed effects** models would be appropriate.

- On the other hand, if I own the company that makes the machines, then I'd want to characterize the performance of *any one* of the machines, conceptually *drawn at random*. The machines would then constitute a population, and I'd use **random effects** analyses.

The Random-Effects Idea



Measurement

Fixed Effects Approach

- Stack together all the activity data from all the subjects (conceptually, at least) into one vector Y.

- Build a design matrix for all the data, with additional coefficients for the *group effect*. If the data are *subject-separable*, then the big X matrix will be block diagonal for within subject factors.

- Schematically (for two subjects):

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \delta \end{bmatrix}$$

Fixed Effects Approach (Cont'd)

- The coefficient δ models the shift in mean activity of subject 2 with respect to subject 1. (Why only one δ, though we have 2 subjects?)

- If $\beta_1 \neq \beta_2$, then we have an interaction between effect and subject. Hence, one hypothesis to test is

$$H_0 : \beta_1 = \beta_2$$

- If $\delta = 0$ then we don't have a significant subject effect; so another useful test is

$$H_0 : \delta = 0$$

Random Effects: Two-Stage Analyses

- The simple (and most common) way to do a random-effects group analysis is by a *two-stage* method; this is the approach which SPM uses. This requires the following assumptions:
  - **Balance:** The design matrix is the same for each subject.
  - **Equal Variance:** The error variances are the same for each subject.
  - **Separability:** The GLM model used must be subject-separable; i.e., the combined design matrix for all subjects is block diagonal, with one block for each subject.

A (Very) Simple Example

- To better understand the distinction between fixed- and random-effects, and to see how a two-stage analysis might work for the random case, it helps to consider the simplest possible model: One-way ANOVA, for a single voxel.

- Fixed Effects Model:

$$y_{ij} = \mu + \beta_i + \epsilon_{ij}$$

  where $i = 1, \ldots, s$ indexes subjects, $j = 1, \ldots, n$ indexes TRs within a subject. The noise term, ϵ_{ij}, is assumed independent N $(0, \sigma^2)$.

A (Very) Simple Example (Cont'd)

- Random Effects Model:

$$y_{ij} = \mu + b_i + \epsilon_{ij}$$

where $i = 1, \ldots, s$ indexes subjects, $j = 1, \ldots, n$ indexes TRs within a subject. The noise term, ϵ_{ij}, is assumed independent N $(0, \sigma^2)$. The subject effects b_i are independent N $(0, \sigma_b^2)$.

Fixed-Effects Estimates: One-Way ANOVA

- Grand Mean:

$$\widehat{\mu} = \bar{y}_{..} \equiv \sum_{i=1}^{s} \sum_{j=1}^{n} y_{ij}/(ns)$$

  Conventional notation: dot indicates summation over index, bar indicates average.

- Subject Effects:

$$\widehat{\beta}_i = \bar{y}_{i.} - \bar{y}_{..}$$

  (Note that $\sum_i \widehat{\beta}_i = 0$.)

- Variance:

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{\sum_{i=1}^{s} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2}{s(n-1)} \\ &\sim \sigma^2 \chi^2_{s(n-1)}/[s(n-1)].\end{aligned}$$

One-Way FE ANOVA (Cont'd)

- Inference for μ:

$$\widehat{\mu} \sim \mathrm{N}\left[\mu, \sigma^2/(ns)\right]$$

Under $H_0 : \mu = 0$

$$\frac{\widehat{\mu}}{\widehat{\sigma}/\sqrt{ns}} \sim T_{s(n-1)}$$

Random-Effects Estimates: One-Way ANOVA

- Grand Mean:

$$\hat{\mu} = \bar{y}_{..} \equiv \sum_{i=1}^{s} \sum_{j=1}^{n} y_{ij}/(ns)$$

- Variance Components:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum_{i=1}^{s} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2}{s(n-1)} \\ &\sim \sigma^2 \chi^2_{s(n-1)}/[s(n-1)]. \end{aligned}$$

$$\begin{aligned} \mathsf{MS}_b &= \sum_{i=1}^{s} n(\bar{y}_{i.} - \bar{y}_{..})^2/(s-1) \\ &\sim (n\sigma_b^2 + \sigma^2)\chi^2_{s-1}/(s-1) \end{aligned}$$

One-Way RE ANOVA (Cont'd)

- Inference for μ:

$$\widehat{\mu} \sim \mathsf{N}\left[\mu, \sigma_b^2/s + \sigma^2/(ns)\right]$$

$$\mathsf{MS}_b/(ns) \sim \left[\sigma_b^2/s + \sigma^2/(ns)\right] \chi_{s-1}^2/(s-1)$$

Under $H_0 : \mu = 0$

$$\frac{\widehat{\mu}}{\sqrt{\mathsf{MS}_b/(ns)}} \sim T_{s-1}.$$

Comments on One-Way ANOVA

- The *estimates* (here, $\widehat{\mu}$) are the same for both the random and fixed analyses.

- The *standard errors* are different. The standard error of $\widehat{\mu}$ under a FE model is $\sigma/\sqrt{ns}$. Under a RE model it is $\sqrt{\sigma_b^2/s + \sigma^2/(ns)}$. Note the between-subject component, which depends only on s, not on n.

- The estimated standard error of $\widehat{\mu}$ for the FE analysis makes use of all of the data:

$$\text{Var}\,(\widehat{\mu}) = \frac{\widehat{\sigma}^2}{ns} = \frac{\sum_{i=1}^{s}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2}{(ns)s(n-1)}$$

- The estimated RE standard error of $\widehat{\mu}$ uses *only the group means*:

$$\text{Var}\,(\widehat{\mu}) = \frac{\text{MS}_b}{ns} = \frac{\sum_{i=1}^{s} n(\bar{y}_{i.} - \bar{y}_{..})^2}{(ns)(s-1)}$$

Two-Stage Approach for Random Effects

- Stage 1: Obtain the a map of effects for each subject.

- Stage 2: Use these effect maps (by analogy, the estimates $y_{i.}$ in our one-way ANOVA example).

- Form the t-statistic for an overall test of significance of the effect or contrast.

- Note why the subject-separability, balance, and equal variance assumptions are required: we need to be able to estimate the individual subject maps independently, and these maps need to enter into the second stage on "equal footing".

The Problem of Not Enough Subjects

- RE models include variability between subjects into the standard errors of estimates.

- If you only have a few subjects (e.g., 5 or so), then there is not much information in the data to estimate this variability!

- So your standard errors are large, and it's much harder to establish significance than it is with FE analyses. (Note the degrees of freedom of the t-statistics in our example: $n(s-1)$ for FE; $s-1$ for RE. So the t-distribution is more diffuse, *and* the standard error has the extra σ_b^2/s term.)

Not Enough Subjects (Cont'd)

- It's important to realize that the large standard errors for RE analyses with few subjects is usually not a fault of the methodology. Rather, one is incorporating σ_b^2 in the standard errors of the estimates, and this is quantity which can't be well estimated except when either
  - You have lots of subjects, and so σ_b^2/s is reasonably small, and your t-test for effect significance has adequate degrees of freedom; or
  - You *regularize* the estimate of $\widehat{\sigma}_b^2$ by including information which isn't in the data. This can be done explicitly, via a prior distributions and a Bayesian analysis, or implicitly, as in Worsley's recent work (*NeuroImage*, 1-15, 2002)

Typicality

- Friston, Holmes and Worsley (*NeuroImage*, 1-5, 1999) introduce the concepts of *typicality* and *conjunction analysis* as a way to make inference with respect to a population in a fixed-effects context.

- If one has a small sample of subjects, and a certain feature is observed in several of these subjects (adjusting for multiple comparisons), then one can say, qualitatively, that this feature is "typical," and thus likely to be present in a population.

- This is to be contrasted from quantitative assessment of what the "average" effect is in a randomly selected subject from a population.

Conjunction Analysis

- In *conjunction analysis*, one attempts to find what activity is statistically significantly in all (or, perhaps, most) subjects.

- This feature can then be thought of as typical, i.e., more likely than not to be present in the population from which the subjects are drawn.

Model Validation

- The GLM is a very powerful tool, but like any modeling tool, it is only good to the extent that the modeling assumptions are valid.
- If assumptions are grossly violated, then inferences can be seriously misleading.
- The assumptions underlying the model include:
  - The form of the model for the mean.
  - The temporal correlation structure, and equal-variance assumptions.
  - Gaussian errors.
  - Separation of signal from noise (e.g., What part of the trend in a time course is a "nuisance effect" to be filtered out, and what part of it is slowly varying signal?)

The Form of the Model

- If your X matrix does not appropriately model the factors contributing to mean activity, then your estimates can be seriously biased.

- This bias can, in principle, be detected by looking at the residuals.

- Think of the example of a straight line fit to data for which a parabola would be much

- How would the residuals (deviations from the fit) tell you that your model is inappropriate?

Error Variance Assumptions

- Inappropriate modeling of temporal correlation can give you a biased estimate of the uncertainty in effects, and grossly incorrect estimates of degrees of freedom for voxel t- or F-statistics.

- In principle, one can test this by looking to see if the residuals at each time course are (at least approximately) white noise.

- How does the temporal autocorrelation vary from voxel to voxel? Is it adequate to use the same model for each voxel?

- Assuming equal within-voxel variances when these variances differ considerably is also something that one might want to look out for, though checking the correlation estimates is probably more important.

Gaussian Errors

- When doing inference, we assume that the noise in our data follows Gaussian distributions.

- (This assumption is necessary for determining standard errors of estimates; it is not required for the estimates themselves.)

- Fixed effects analysis are not very sensitive to violation of this assumption. The central limit theorem implies that averages tend to be Gaussian in many situations, and coefficient estimates are essentially weighted averages. The central limit theorem also tells us that standardized contrasts will have approximate t-distributions under fairly general conditions (provided the standard errors and degrees of freedom are appropriately estimated).

- This robustness, unfortunately, does not extend to random effects. Estimates of variances between subjects, for example, will likely be sensitive to to the assumption of Gaussianity. That being said, Gaussian random-effects models are very widely used, because there are not good alternatives.

Separation of Signal from Noise

- A necessary step in any fMRI analysis is to remove nuisance effects from the data.

- Usually these results are low-frequency trends, and they are removed either by high-pass filtering, or by explicit modeling via covariates in the GLM.

- Always keep in mind that if your have signal which looks like the trend being removed, then you might be "throwing the baby out with the bathwater."

- One example might be a nuisance physiological effect, which you'd like to model and remove. If this effect is, at least in part, associated with an experimental stimulus, then you could be discarding important signal with the noise.

Model Selection

- In any course in regression analysis, one learns how to choose a “best” model from within a family of interesting candidate models.

- Part of this approach involves examining candidate models for goodness-of-fit, mostly be examining residuals as discussed earlier.

- Another part of this approach is model comparison, which involves fitting a “large” model, with perhaps too many parameters, and then comparing this fit to a “smaller” model in which some of these parameters are constrained, either to equal zero, or else perhaps to equal each other.

Model Selection (Cont'd)

- Model comparison thus reduces to hypothesis testing, in the simplest textbook situations, to F-tests.

- This approach can be applied to fMRI, although instead of a single F-test, we will have F maps and associated p-value maps to interpret.

- More general model comparison tool compare the reduction in residual sum of squares between nested models, penalizing for complexity due to adding parameters. Two such criteria are AIC and BIC (Akaike Information Criterion; Bayesian Information Criterion).

Conclusions for Model Validation

- Basic assumptions of GLM: mean model (design matrix), noise model (Gaussian, correlation structure), and the partitioning of the model into noise, nuisance factors, and factors of interest.

- Residual analysis and model slection criteria (for a family of nested models of varying complexity) are useful for validating the mean model.

- Autocorrelation functions are the main tool for validating the correlation part of the noise model.

- The Gaussian assumption is not critical for fixed effects models. It is more important for random effects models, but no-Gaussian modelling alternatives are presently not available.

- The primary challenge in model validation in fMRI is the vast number of voxels.