Selected Topics in Statistics for fMRI Data Analysis

Mark Vangel

vangel@nmr.mgh.harvard.edu

mvangel@mit.edu

General Clinical Research Center

MGH & MIT

Outline

I. Adjusting for Multiple Comparisons
II. Permutation Tests
III. Modelling Data from Multiple Subjects
IV. Some Thoughts on Model Validation

I. Multiple Comparisons

Ia. Bonferroni Approximation

Ib. Gaussian Random Field Assumption

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 one 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 (Model)

- 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 $e_i \sim \mathrm{N}\left(0, 10^2\right)$.

- (**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.)

Illustrative Example (Hypothesis)

- We are interested in testing, at the $\alpha = 0.05$ level, the null hypothesis

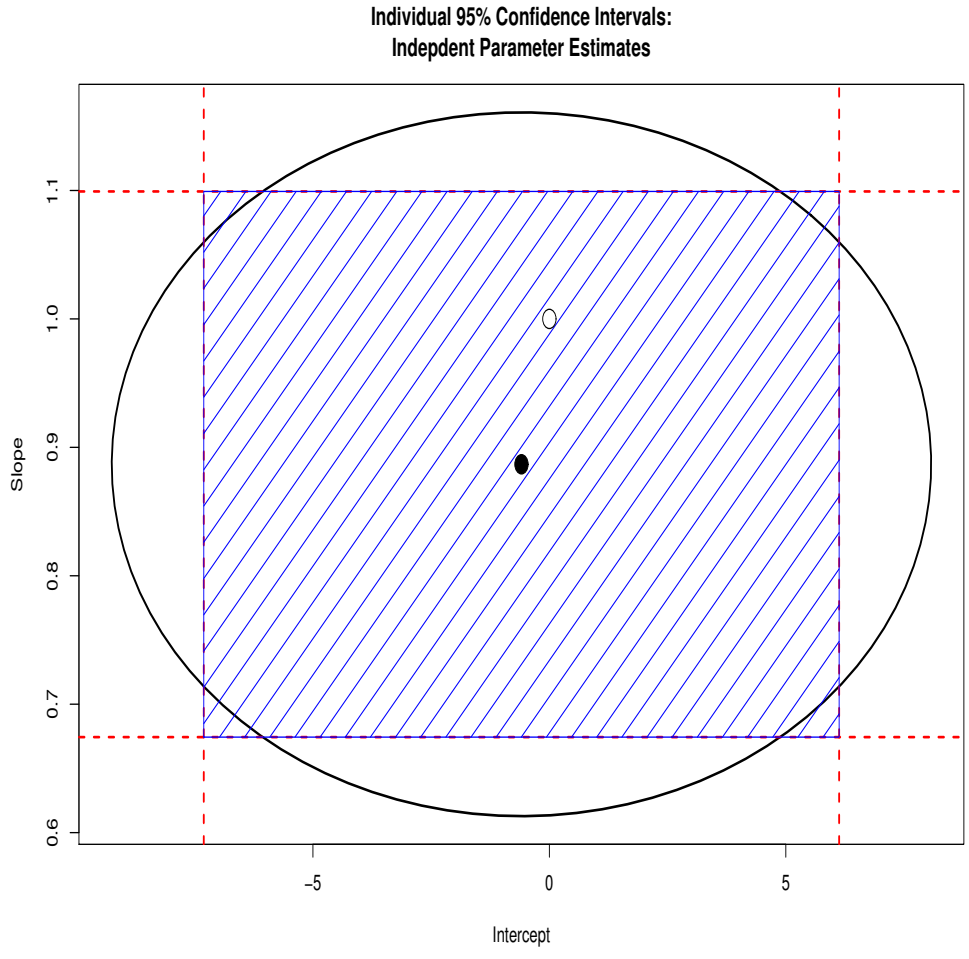
$$H_0 : \delta = 0 \text{ and } \beta = 1,$$

against the alternative

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

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

Confidence Region



Comments

- The box formed by the two individual confidence intervals is considerably smaller than the actual bivariate confidence region.
- Each confidence interval for a parameter has confidence level 0.95, so the region formed by the intersection of these intervals has confidence $0.95^2 = 0.9025 < 0.95$.

Comments (Cont'd)

- Over repeated future data, the probability that either parameter falls in it's interval is $1 - \alpha_* = 0.95$. Since the model has been set up so the the estimates $(\hat{\delta}, \hat{\beta})$ are independent, the actual probability of rejecting H_0 for the pair of confidence intervals

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

Comments (Cont'd)

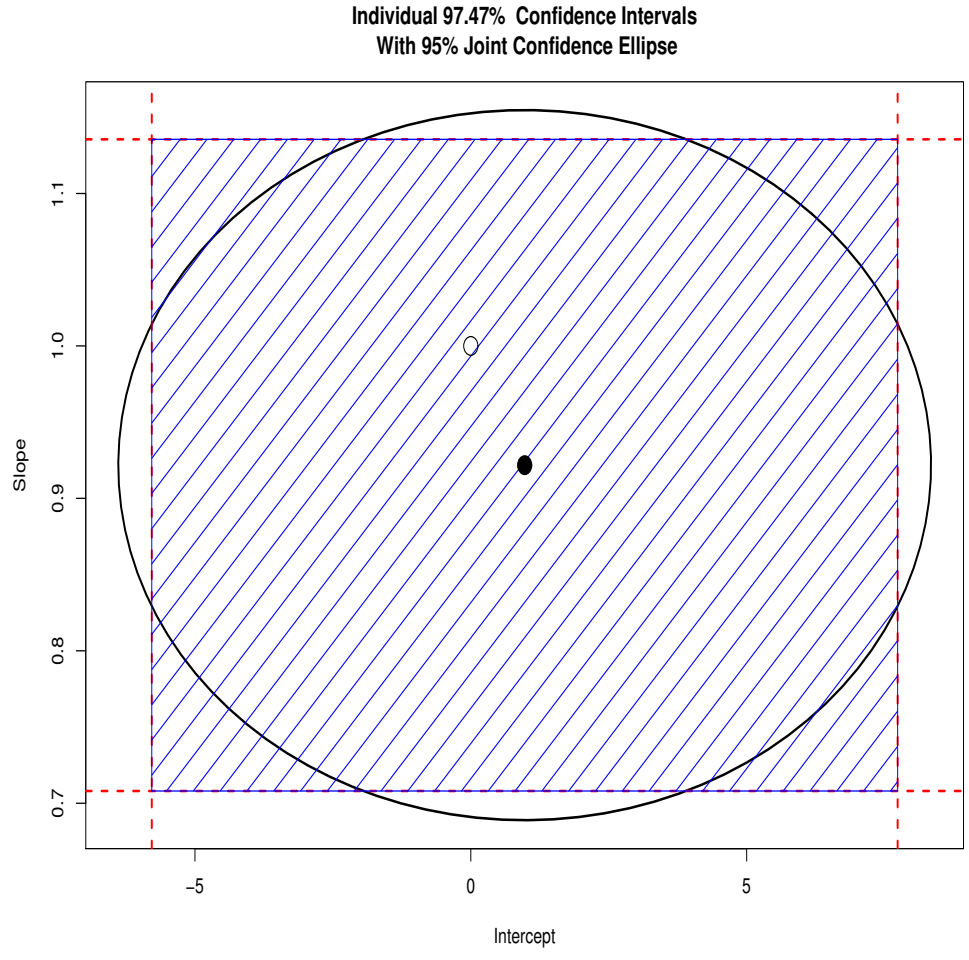
- Working backwards, if we choose α_* to be

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

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

- This is approach achieves exactly 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 $t_1, \ldots, t_k$.
- We would like to calculate the probability of rejecting the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

against the alternative that H_0 is not true.

Bonferroni Correction (Cont'd)

- This probability of rejecting H_0 is

$$\begin{aligned}\alpha &= P_0(|T_1| \geq t_1 \text{ or } |T_2| \geq t_2 \text{ or } \ldots |T_k| \geq t_k) \\ &= 1 - \prod_{i=1}^{k} \Pr(|T_i| \leq t_i) = 1 - (1-\alpha_*)^k.\end{aligned}$$

- Hence, we choose

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

Example Revisited: Alternative Parameterization

- Next we see what happens in our simple linear regression example if we don't subtract of 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 $e_i \sim \mathrm{N}\left(0, 10^2\right)$. To relate this to the previous parameterization, note that

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

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

Alternative Parametrization (Cont'd)

- We are interested in testing the null hypothesis

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

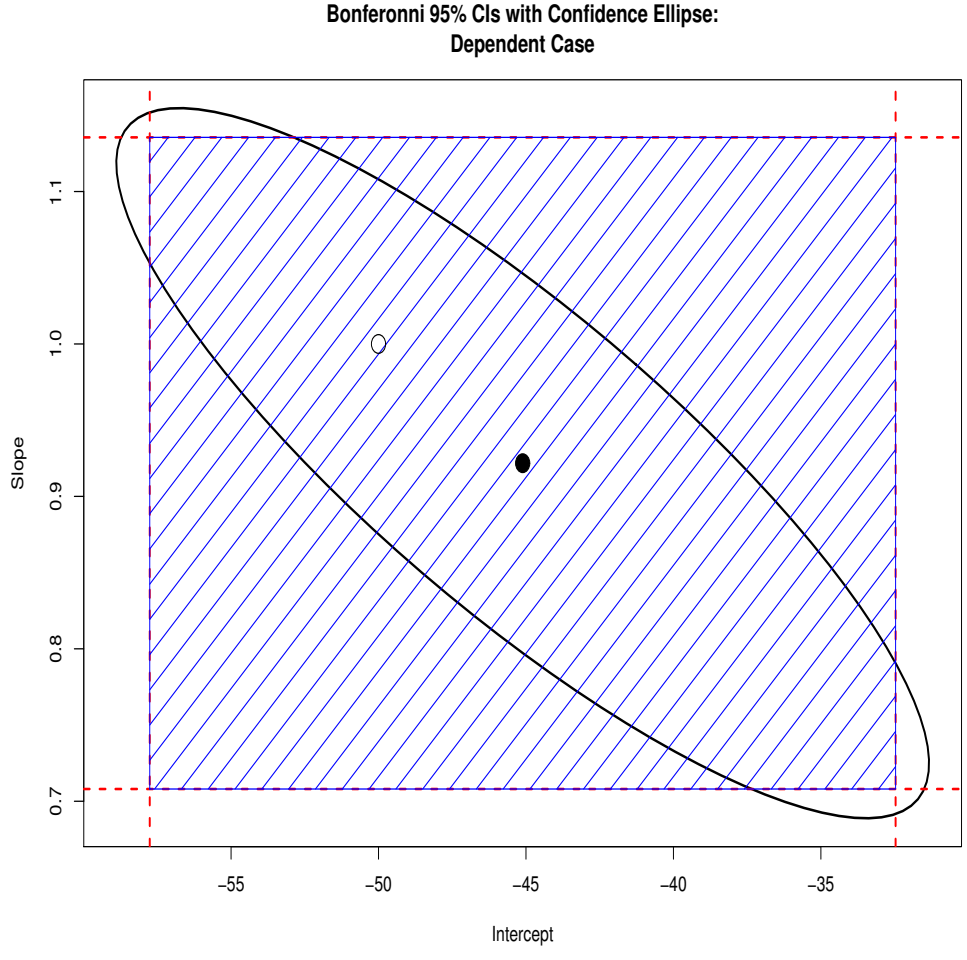
against the alternative

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

at the 0.05 significance level.

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

Confidence Region for a Dependent Example, With Bonferroni Interval



Bonferroni and Activation Clusters

- In addition to requiring that p values be below a threshold, one can impose as an additional requirement that there be a minimum number of voxels *clustered* at any "active" location.
- There are obviously many ways to pair critical p-values with minimum cluster sizes.
- There is a stand-alone C program, AlphaSim that can determine cluster significance levels by simulation.
- AlphaSim is part of the AFNI distribution (Bob Cox, NIH, afni.nimh.nih.gov)

Example AlphaSim Command Line

- A typical run of AlphaSim :

```
AlphaSim  -nx 46 -ny 55 -nz 46 \
          -dx 4.0 -dy 4.0 -dz 4.0 \
          -sigma 0.65 \
          -rmm 6.93 \
          -pthr 0.05 -iter 10000
```

AlphaSim Command Line (Cont'd)

- -nx -ny -nz : Dimension of brain in voxels
- -dx -dy -dz : Voxel size in *mm*.
- -sigma : SD of Gaussian smoothing kernel
- -rmm : Two active voxels $\leq$ rmm *mm* apart are considered to be in the same cluster.
- -pthr : Threshold p-value
- -iter : Number of simulations.
- (See AlphaSim documentation for other options.)

Example AlphaSim Output

```
Data set dimensions:
nx =    46    ny =    55    nz =    46    (voxels)
dx =  4.00    dy =  4.00    dz =  4.00    (mm)

Gaussian filter widths:
sigmax  =  0.65    FWHMx  =  1.53
sigmay  =  0.65    FWHMy  =  1.53
sigmaz  =  0.65    FWHMz  =  1.53

Cluster connection radius:  rmm =  6.93
Threshold probability:    pthr = 5.000000e-02
Number of Monte Carlo iterations   = 10000
```

Example AlphaSim Output (Cont'd)

Cl	Size	Frequency	Max Freq	Alpha
	1	15616950	0	1.000000
	2	5123184	0	1.000000
	3	2397672	0	1.000000
	4	1320445	0	1.000000
	38	228	210	0.113100
	39	190	175	0.092100
	40	140	134	0.074600
	41	114	108	0.061200
	42	91	87	0.050400
	43	60	57	0.041700

Interpretation of AlphaSim Results

- Maximum active clusters of 42 or more below threshold $p = 0.05$ occur about 5% of the time under the null hypothesis of no activation.
- Note the following:
  - For a higher p-value threshold, the minimum significant cluster size will be larger.
  - This approach accounts for spatial correlation induced by smoothing, but not for and spatial correlation present in the unsmoothed data.

Summary: Bonferroni

- For an overall test at the α significance level, select individual voxels among N total as active if $p \leq \alpha/N$.
- Not a bad approximation if voxels are nearly independent.
- Can be *very* conservative if there is considerable spatial correlation among voxels.
- Using both a p-value threshold *and* a minimum cluster size via AlphaSim is one way to partially overcome this conservatism.

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 activation 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

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

- 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 activation.

Expected EC

- By definition,

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

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

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

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

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

Expected EC (Cont'd)

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

- Either
  - Attempt to approximate this expectation for a choice of u (adjusted p-value), or
  - Select u so that $E(\mathrm{EC})$ 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 full width at half maximum (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*.

Summary: Gaussian Random Fields

- GRF theory requires that we know the spatial correlation, at least approximately.
- In order to meet this requirement, we must do fairly hefty spatial smoothing (i.e., precoloring).
- This has the obvious disadvantage of blurring together brain structures with different functions, particularly if the smoothing is not done on the cortical surface.
- Compare with `AlphaSim` , another way for accounting for spatial correlation due to smoothing.

False Discovery Rate

- The Bonferroni and GRF approaches ensure 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

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

  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:**

  $N = 64 \times 64 \times 16 = 65,536$

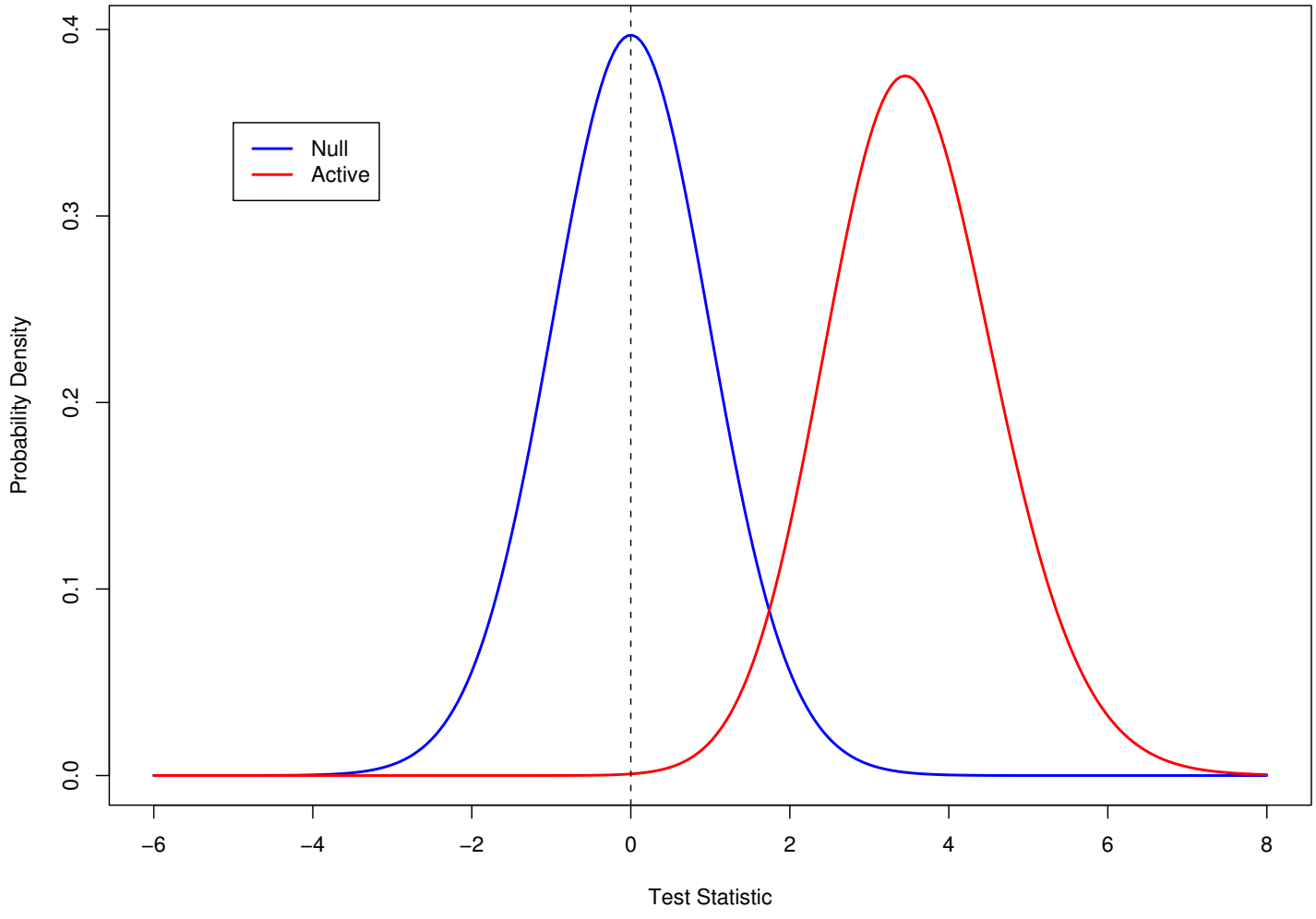
- **Number of Active Voxels:**

  $N_1 = 0.02N = 1,335$

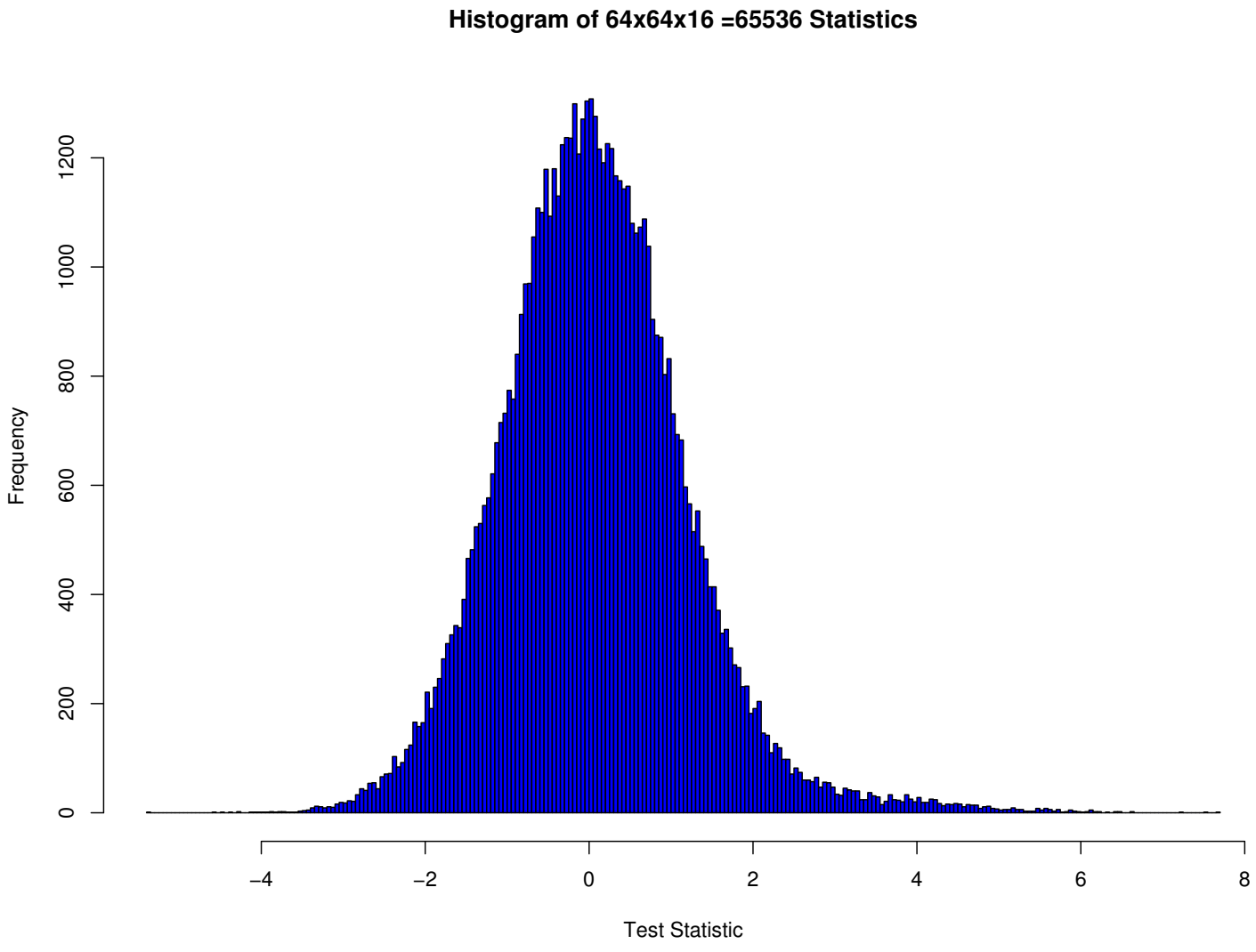
- "Inactive" statistics independently distributed t_{50}.

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

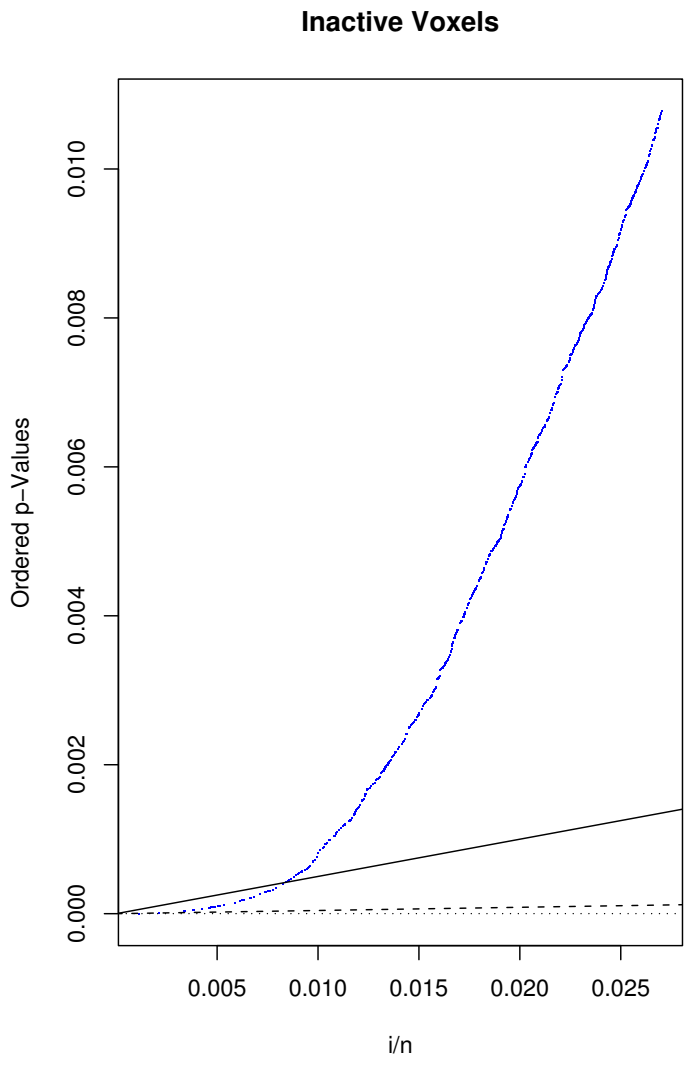
Densities for Active and Inactive Voxel Statistics

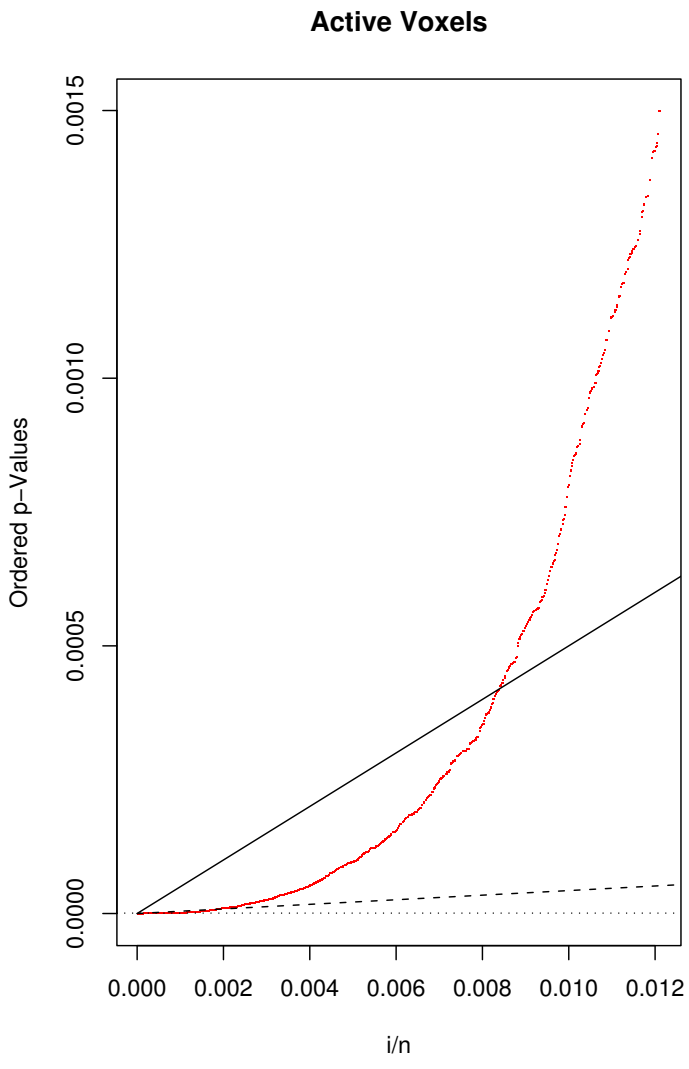


Histogram of the Voxel Statistics



Graphical Illustration of Results





Simulation Results

- $\mathrm{FDR} = 35/549 \doteq 0.064$, $c = 1$:
  (Solid line in preceeding figure)

	Discovered	
	Yes	No
Correct	514	64,166
Error	35	821
Total	549	64,987

Simulation Results

- $\mathrm{FDR} = 1/123 \doteq 0.008$, $c = 1/(\log(N) + \gamma)$:
  (Broken line in preceeding figure)

	Discovered	
	Yes	No
Correct	122	64,200
Error	1	1213
Total	123	65,413

Simulation Results

- Bonferroni $(\mathrm{FDR} = 0)$, $p = .05/N = 7.6 \times 10^{-7}$
  (Not shown in preceeding figure)

	Discovered	
	Yes	No
Correct	44	64,201
Error	0	1291
Total	44	65,492

Summary: False Discovery Rate

- Can be more sensitive at detecting true activation than Bonferroni without requiring the heavy spatial smoothing of GRF theory.
- But a change in philosophy is required: instead of making the likelihood of *any* voxel being falsely declared active small, one is willing to accept that a small proportion of voxels will *likely* be false discoveries, and instead attempt to control the size of this proportion.

II. Permutation Tests

IIa. Introduction and illustrative example (Strauss et al., NeuroImage 2005). .

IIb. Heart Damage and Stroke (Ay et al., Neurology 2006)

Permutation Tests: Introduction

- Permutation tests are useful for comparing groups or conditions without distributional assumptions:
- **Ref:** Nichols, TE and Holmes, AP (2001). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, **15**, 1-25.

Illustrative Example

Data from Strauss et al. (2005). FMRI of sensitization to angry faces. *NeuroImage*, **26(2)**, 389-413. Left anterior cingulate (LaCG) activation to angry faces in first and second half of a session, for eight subjects.

Subject	AS	BG	CS	GK	JT	ML	MP	RL
First	0.02	0.06	0.00	0.33	-0.07	0.01	-0.17	0.18
Second	0.36	0.22	0.19	0.26	0.47	0.16	0.46	0.09

Illustrative Example (Cont'd)

- A paired t-test with $8 - 1 = 7$ degrees of freedom leads to the t-statistic 2.515.
- Comparing this value to the *theoretical* reference null distribution (T_7), one determines a two-sided p-value of 0.040.
- The t-test is quite robust to modest departures from assumptions, even for $N = 8$, so using the T_7 as a reference distribution for the p-value is probably OK.
- However, what if one did not want to make the assumptions necessary for the validity of this theoretical null distribution?
- (Note that for some complicated test statistics, or for very messy data, one often doesn't know a reasonable approximation to the null distribution.)

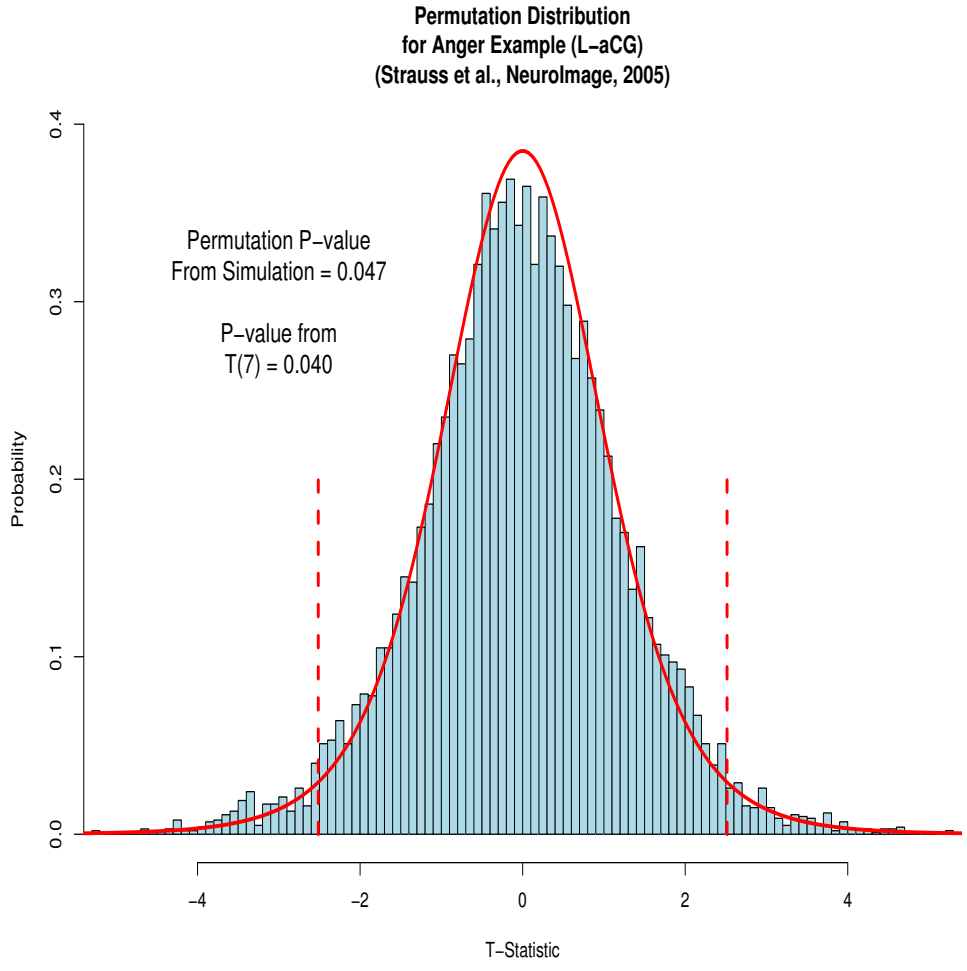
Example: The Permutation Distribution

- The set of 16 numbers can be divided into 2 ordered pairs (first, last) 518,918,400 ways – of which only 1 will correspond to the *correct* pairing.
- The basic idea of a permutation test is to randomly permute the "labeling" of the data (i.e., the assignment of values to pairs, and the ordering of these pairs) many times.
- For each labeling, a test statistic of interest is calculated (here a paired t-statistic).
- One then compares that statistic obtained from the correctly labeled data (here, $T = 2.515$) with the *empirical* reference distribution of the same statistic calculated for many permuted labellings.

Example: Remarks

- Note that one calculated a t-statistic, but never needed to use the theoretical t-distribution to get a p-value.
- Note also that this approach can be applied in a very wide range of practical situations.

Example: Permutation Test Result



Troponin, Stroke, and Myocardial Injury

- Ay, H. et al. (2006). Neuroanatomic correlates of stroke-related myocardial injury. *Neurology*, to appear.
- Hypothesis:
  - A High level of troponin is a sensitive marker of heart damage.
  - Heart damage could result from strokes in certain locations.
  - Can we determine where these locations might be by comparing stroke patients with high troponin with matched low-troponin controls?

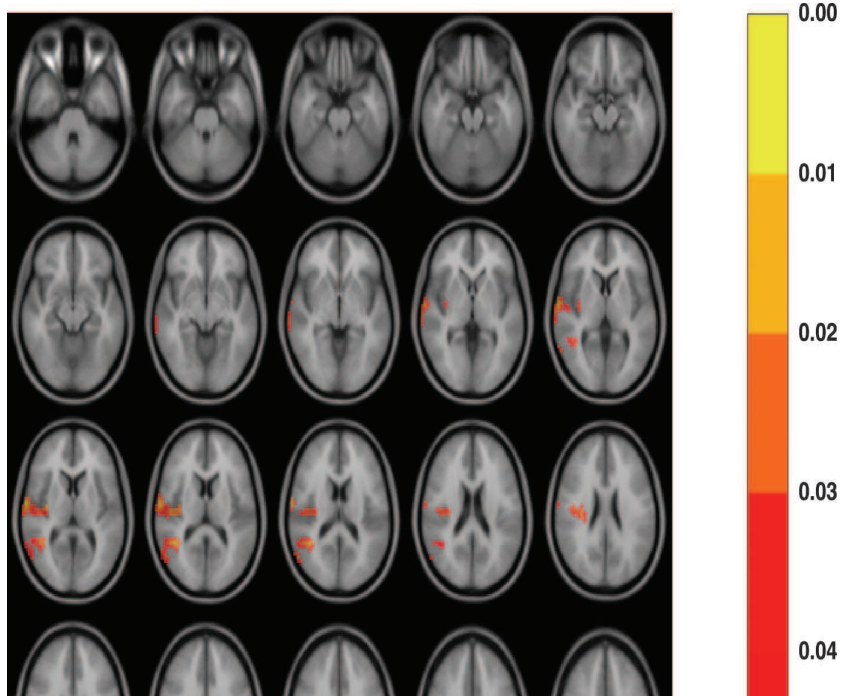
Troponin: Data

- Data: For 50 consecutive stroke patients with high troponin and 50 stroke controls with very low troponin, we have a mask of zeros and ones indicating which voxels are infarcted in each stroke lesion.
- We could compare these with voxel-wise t-tests, except that the masks are *very* non-Gaussian.

Troponin: Permutation Test

- Permute the labeling of high (cases) and low (controls) troponin and calculate voxel-wise t-statistics.
- Use AlphaSim to determine a suitable threshold and cluster constraint (threshold of 0.05, minimum cluster of 43 voxels).
- Result: Patients with strokes in the right insula and right inferior parietal lobule tended to more frequently have high troponin than other stroke patients.

Example: Permutation Test Result



Summary

The concept of a permutation test is extraordinarily powerful and useful. These tests are easy to understand, and, in principle, easy to apply. They are useful in situations where one wishes to employ a statistic with unknown distribution under the null hypothesis, or perhaps a well-known test statistic in situations where the assumptions for the usual null distribution are not satisfied.

III. Analyses for Groups of Subjects

IIIa. Fixed Effects

- Analysis on average maps.

IIIb. Random Effects

- Usual two-stage approach
- Worsley et al. (*NeuroImage*, 2002)
- A Bayesian approach

IIIc. Examples of Bayesian Two-Stage Random Effects Modelling

- Spatial visual cueing
- Passive viewing of angry faces

IIId. 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.

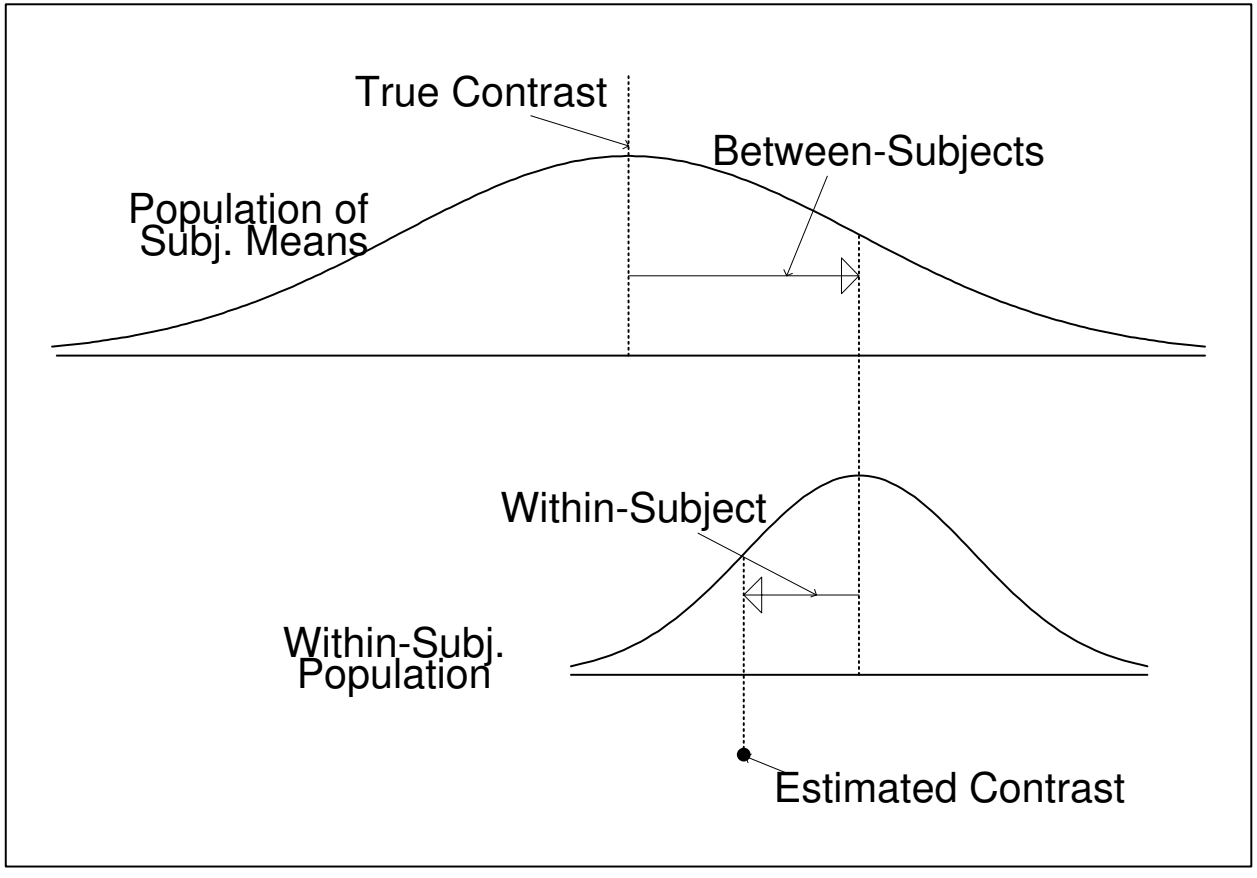
Fixed vs. Random Effects

- 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 owned 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

- A contrast at any given voxel is regarded as a sum of three components:
  1. The true (but unknown) contrast
  2. A random shift from the truth which depends **only** on the subject.
  3. A second random shift from the truth due to measurement uncertainty **within** a subject.
- In the limit of many subjects, (2) can be made arbitrarily small; in the limit of long scans, (3) can be made arbitrarily small (except perhaps for a measurement bias).

The Random-Effects Idea: Schematic



Two Approaches to Data Analysis

- **Fixed-Effects Analysis:** Average data over subjects, look at p-values for contrast on average map. (Degrees of freedom $\approx$ number of **time points** in scan.)
- **Random-Effects Analysis:** Estimate contrast map for each subject. Use these maps as "data" for a second-level analysis. (Degrees of freedom $\approx$ number of **subjects.**)

“Standard” Two-Stage Approach for Random Effects

- Stage 1: Obtain the a map of effects for each subject.
- Stage 2: Use these effect maps as “data” in the second stage of the analysis.
- Form the t-statistic for an overall test of significance of the effect or contrast.
- Note that these maps enter into the second stage on “equal footing”.

Critique of Usual Two-Stage Approach

- The usual two-stage approach to multi-subject analyses treats the contrast estimate maps from each subject as given data, without consideration of the uncertainty in these values, which may be considerable and which may differ from subject to subject.
- A better approach is two summarize a contrast of interest by *two* maps: a contrast estimate map, and a corresponding standard error map. This is the approach advocated by Worsley (*NeuroImage* (2002)), for example.

Worsley et al. *NeuroImage*, 2002, 1-15

- **Within-run analysis:** Fit linear model with cubic regression spline terms for trend, assuming AR (p) error structure. Prewhiten using estimated covariance matrix, and refit.
- Covariance matrix is estimated by implicitly solving Yule-Walker equations; correlations are corrected for bias and spatially smoothed.
- For a contrast of interest, summarize each run with a contrast map and a SE map.

Worsley et al. *NeuroImage*, 2002, 1-15 (Cont'd)

- **Between-Subject Analysis:** Fit a second-level model, fixing the "within" errors at their estimates, and estimating (EM/REML) "between" variance σ^2, and possible second-level fixed-effect covariates.
- Regularize σ^2 by spatially smoothing between/within ratio. Estimate approximate degrees of freedom of smoothed σ^2 using Gaussian random field theory, form $T-$ or $F-$statistic map for second-level covariates.

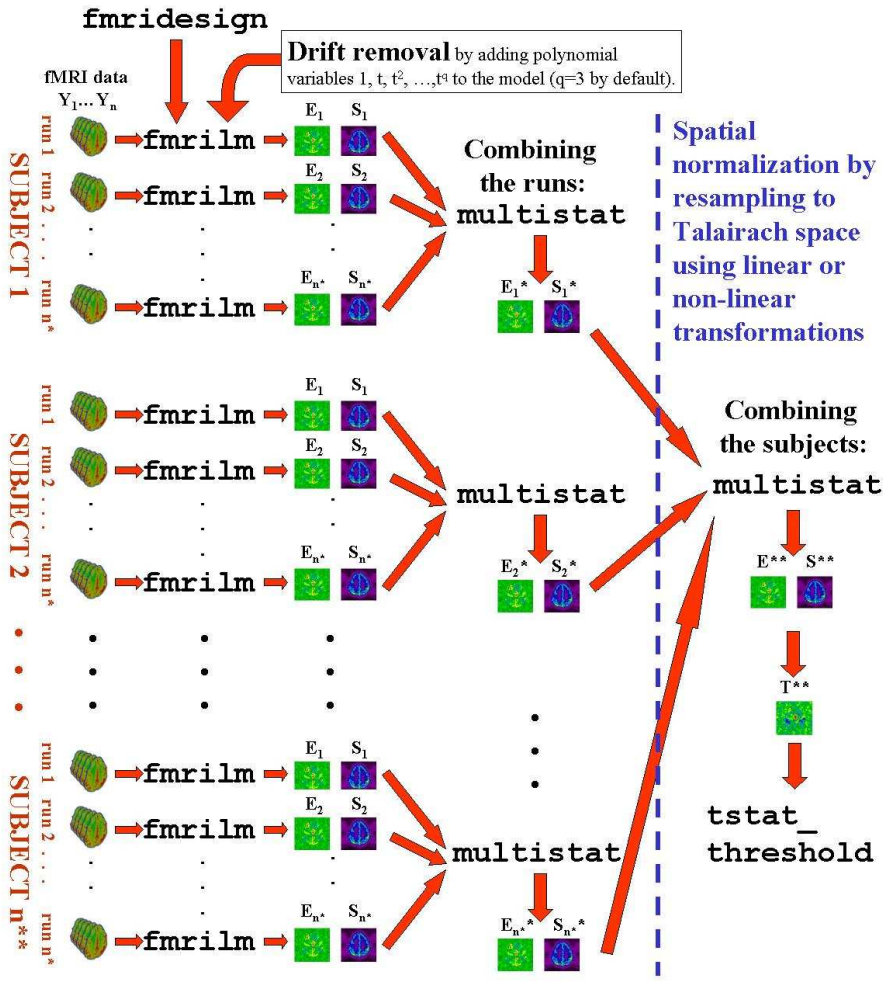


Figure 1: `Fmristat` flow chart for the analysis of several runs (only one session per subject); E = effect, S = standard deviation of effect, $T = E/S = T$ statistic.

A Bayesian Approach

- Assume χ^2 and normal contributions to the likelihood for the within-subject variances and contrast estimates, respectively.
- Model the betwen-subject effects as normally distributed with mean zero and unknown variance.
- Use non-informative prior distributions for within-subject standard deviations, contrast estimates, and usually (but not necessarily) for the between-subject standard deviation.

Bayesian Approach (Cont'd)

- Calculation of posterior distribution of contrast is straightforward by numerical integration.
- Introducing subject-level covariates (e.g., age, treatment) is easy in principle, though simulation ("Gibbs Sampler") will have to replace exact integration.

Bayesian Hierarchical Model for RE Analysis

$i = 1, \ldots, k$ indexes subjects

$j = 1, \ldots, n_i$ indexes time points

$$
\begin{aligned}
p(x_{ij}|\delta_i, \sigma_i^2) &= \mathrm{N}\left(\delta_i, \sigma_i^2\right) \\
p(\sigma_i) &\propto 1/\sigma_i \\
p(\delta_i|\mu, \sigma^2) &= \mathrm{N}\left(\mu, \sigma^2\right) \\
p(\mu) &\propto 1 \\
p(\sigma) &\propto 1
\end{aligned}
$$

Posterior for μ given $\sigma = 0$, $k \geq 1$

Given $\sigma = 0$, then the posterior distribution of the consensus mean μ is proportional to a product of scaled t-densities:

$$p(\mu|\{x_{ij}\}|\sigma = 0) \propto \prod_{i=1}^{k} \frac{1}{t_i} T'_{n_i - 1} \left(\frac{x_i - \mu}{t_i} \right)$$

The General Case: $\sigma \geq 0$

In general, $p(\mu|\sigma, \{x_{ij}\})$ is proportional to a *product* of the distributions of the random variables

$$U_i = x_i + \frac{s_i}{\sqrt{n_i}} T_{n_i-1} + \sigma Z,$$

- where T_{n_i-1} is a t-distributed random variable with $n_i - 1$ degrees of freedom, Z is distributed $\mathrm{N}(0, 1)$, and T_{n_i-1} and Z are independent.
- $t_i = s_i/\sqrt{n_i}$ is within-subject SE; x_i is within subject mean.

A Useful Probability Density

Let T_ν and Z denote independent Student-t and standard normal random variables, and assume that $\psi \geq 0$ and $\nu > 0$. Then

$$U = T_\nu + Z\sqrt{\frac{\psi}{2}}$$

has density

$$f_\nu\left(u;\psi\right) \equiv \frac{1}{\Gamma_{\nu/2}\sqrt{\pi}} \int_0^\infty \frac{y^{(\nu+1)/2-1} e^{-y\left[1+\frac{u^2}{\psi y+\nu}\right]}}{\sqrt{\psi y+\nu}} dy.$$

Posterior of (μ, σ)

- Assume $\delta_i \sim \mathrm{N}(\mu, \sigma^2)$, $\sigma \sim p(\sigma)$, $p(\mu) \propto 1$, $p(\sigma_i) \propto 1/\sigma_i$.
- Then the posterior of (μ, σ) is

$$p(\mu, \sigma | \{x_{ij}\}) \propto p(\sigma) \prod_{i=1}^{p} \frac{1}{t_i} f_\nu \left[\frac{x_i - \mu}{t_i}; \frac{2\sigma^2}{t_i^2} \right].$$

- The posterior of μ given $\sigma = 0$ is a product of scaled t-densities centered at the x_i, since

$$\frac{1}{t_i} f_\nu \left[\frac{x_i - \mu}{t_i}; 0 \right] = \frac{1}{t_i} T'_\nu \left(\frac{x_i - \mu}{t_i} \right).$$

- We will take $p(\sigma) = 1$, though an arbitrary proper prior does not introduce additional difficulties.

Example 1: Spatial Visual Cueing

Pollmann, S. and Morillo, M. (2003). “Left and Right Occipital Cortices Differ in Their Response to Spatial Cueing,” *NeuroImage*,**18**, 273-283.

Neumann, J. and Lohmann, M. (2003). “Bayesian Second-Level Analysis of Functional Magnetic Resonance Images,” *NeuroImage*, **20**, 1346-1355.

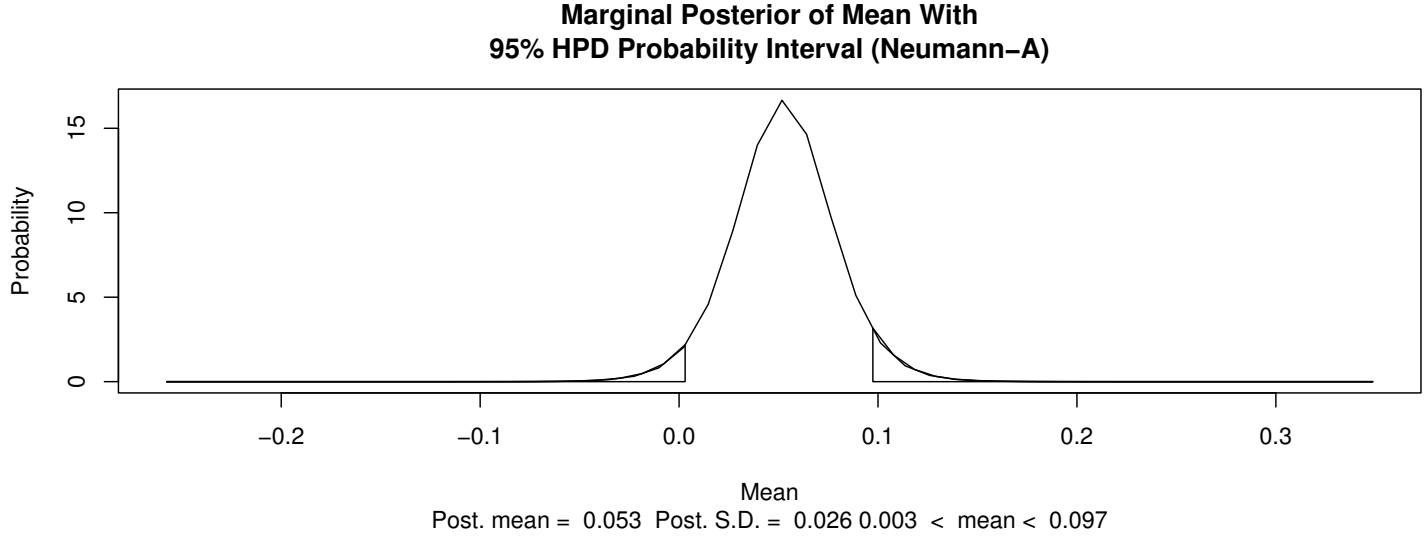
Occipital Cortex and Spatial Cueing

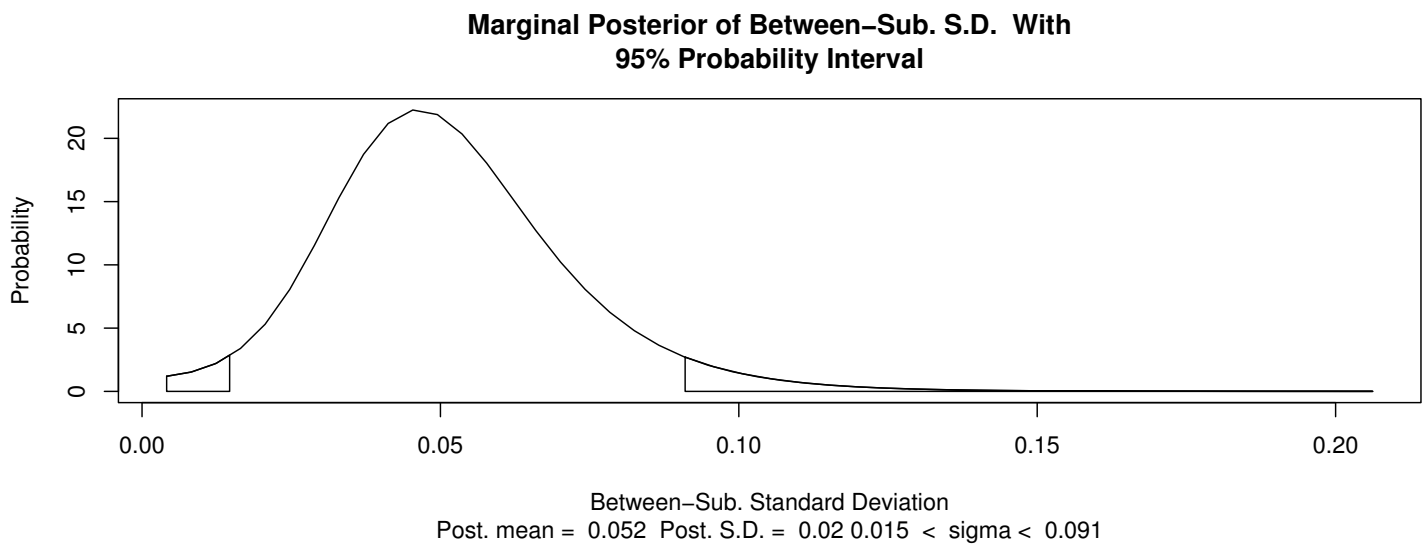
- Visual cue (large or small) on one side of screen (left or right).
- Subject told to fixate on center of screen, but pay attention to side where cue appeared.
- Target appeard either on same side as cue (valid trial) or opposite side (invalid trial)

Pollman and Marillo, Results

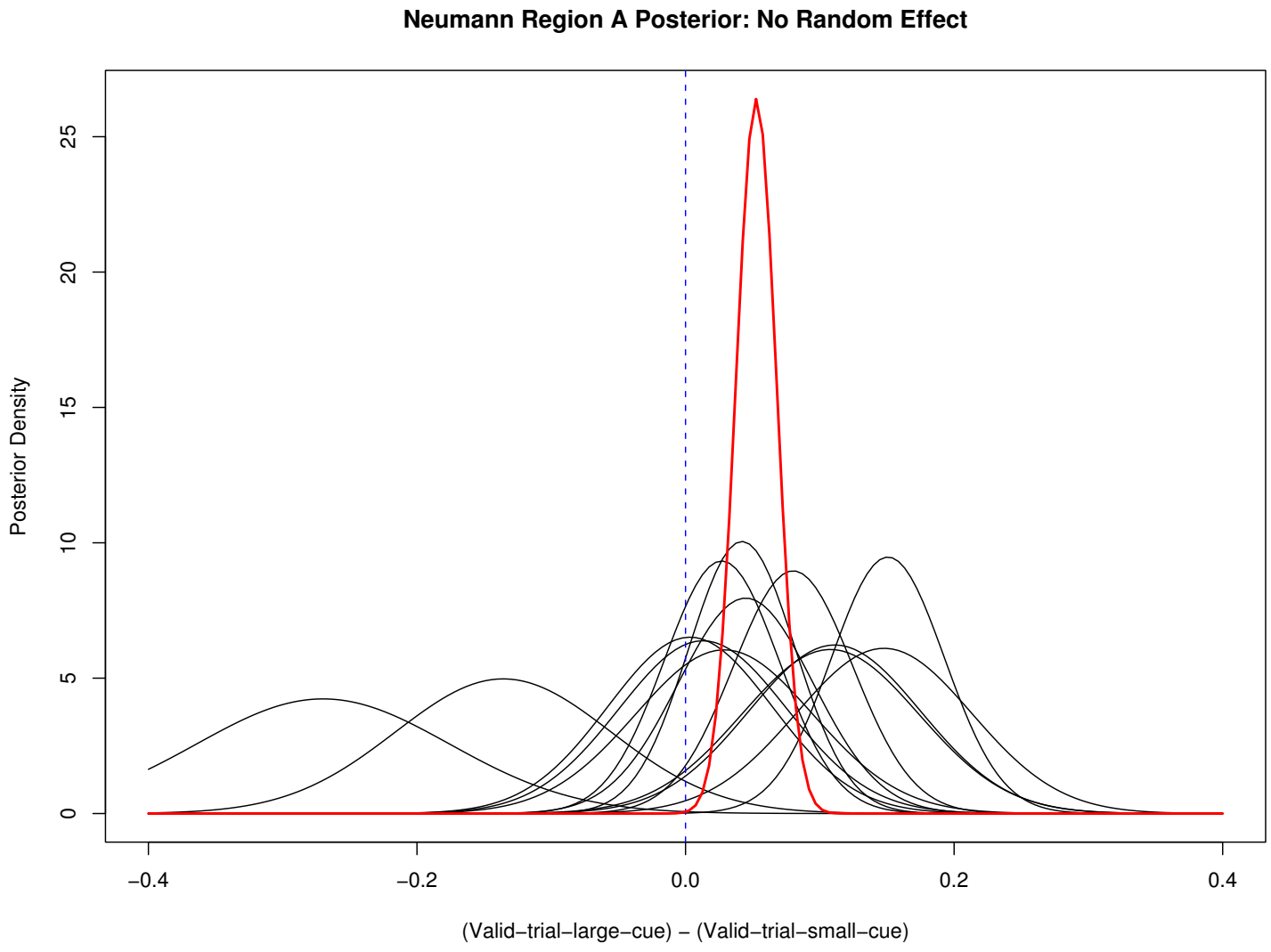
- Main results: Contrast of valid-trial LHS with valid trial RHS showed significant differences in bilateral lingual gyrus and lateral occipital gyrus, and IPS/TOS.
- Second contrast: **valid-trial-small-cue** with **valid-trial-big-cue** significant in three regions from Bayesian analysis of Neumann and Lohmann (2003).

Region A: *valid-small-cue* vs *valid-large-cue*





Posterior A: *valid-small-cue* vs *valid-large-cue*



Example 2: Sensitization to Angry Faces

Vangel, MG and Strauss, MM (2005). “Bayesian and Frequentist Approaches to Two-Stage Inference in Multi-Subject fMRI With an Application to Sensitization to Angry Faces,” Poster at Organization for Human Brain Mapping annual meeting.

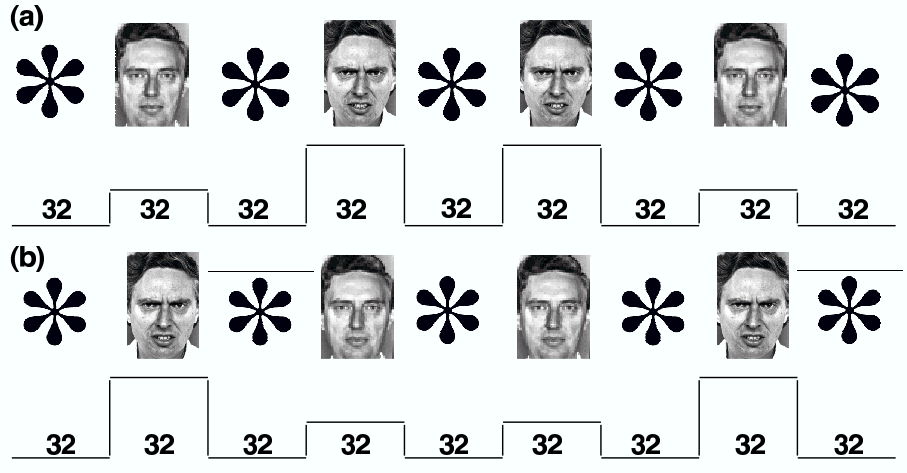
Strauss M., Makris N., Kennedy D., Etcoff N., Breiter H. (2000). “Sensitization of Subcortical and Paralimbic Circuitry to Angry Faces: An fMRI Study,” *NeuroImage* **11**, S255.

Strauss, M.M. (2003). “A Cognitive Neuroscience Study of Stress and Motivation,” Phd Dissertation, Department of Psychology, Boston Univeristy.

Sensitization to Angry Faces

Eight participants passively viewed alternating blocks of angry and neutral Ekman faces, with fixations in between.

Figure 1: Experimental Paradigms



Angry Faces: Design

Subject	Sequence			
A	1	2	1	2
B	1	2	1	2
C	2	1	2	1
D	2	1	2	1
E	1	2	2	1
F	1	2	2	1
G	1	2	2	1
H	2	1	1	2

. . . where NAAN $= 1$ and ANNA $= 2$.

Habituation vs. Sensitization

- One typical aspect of block designs (such as the “angry faces” study) is that subjects tend to *habituate* to the stimulus, with consequent decreased BOLD activation.
- An interesting aspect of the present data is that, in many regions subjects tended to have a *stronger* BOLD response in the second half as compared to the first. This is called *sensitization*.

A Regression Model

- For "representative" voxels in each subject:

$$\begin{aligned} \log(y_t) &= \beta_0 + \beta_{\text{half}} + \beta_{\text{type}} + \\ & \quad \beta_{\text{half}} \times \beta_{\text{type}} + \epsilon_t \end{aligned}$$

- where β_{type} is a 3-level factor for face type (Angry, Neutral, Fixation); β_{half} (levels 1 and 2) compares the first and second half of the experiment, and ϵ_t is (for simplicity) here modeled as white noise.

Habituation/Sensitiziation Contrast

- For models of log of the data, contrasts become dimensionless ratios. (Only BOLD changes have real meaning.)
- The following contrast is useful for testing for sensitization/habituation:

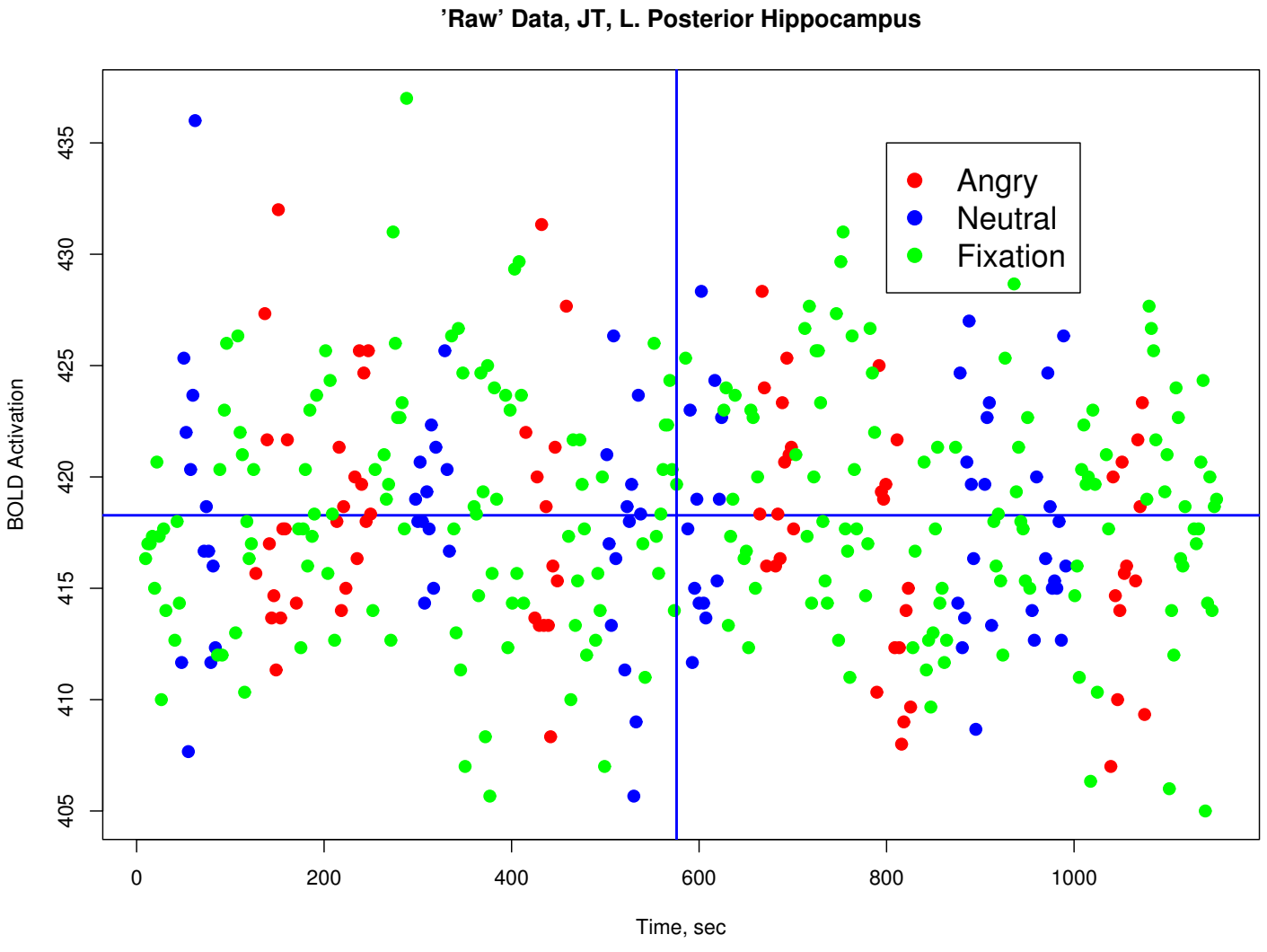
$$c_S = \exp[(\beta_{A,2} - \beta_{N,2}) - (\beta_{A,1} - \beta_{N,1})$$

- We also looked at

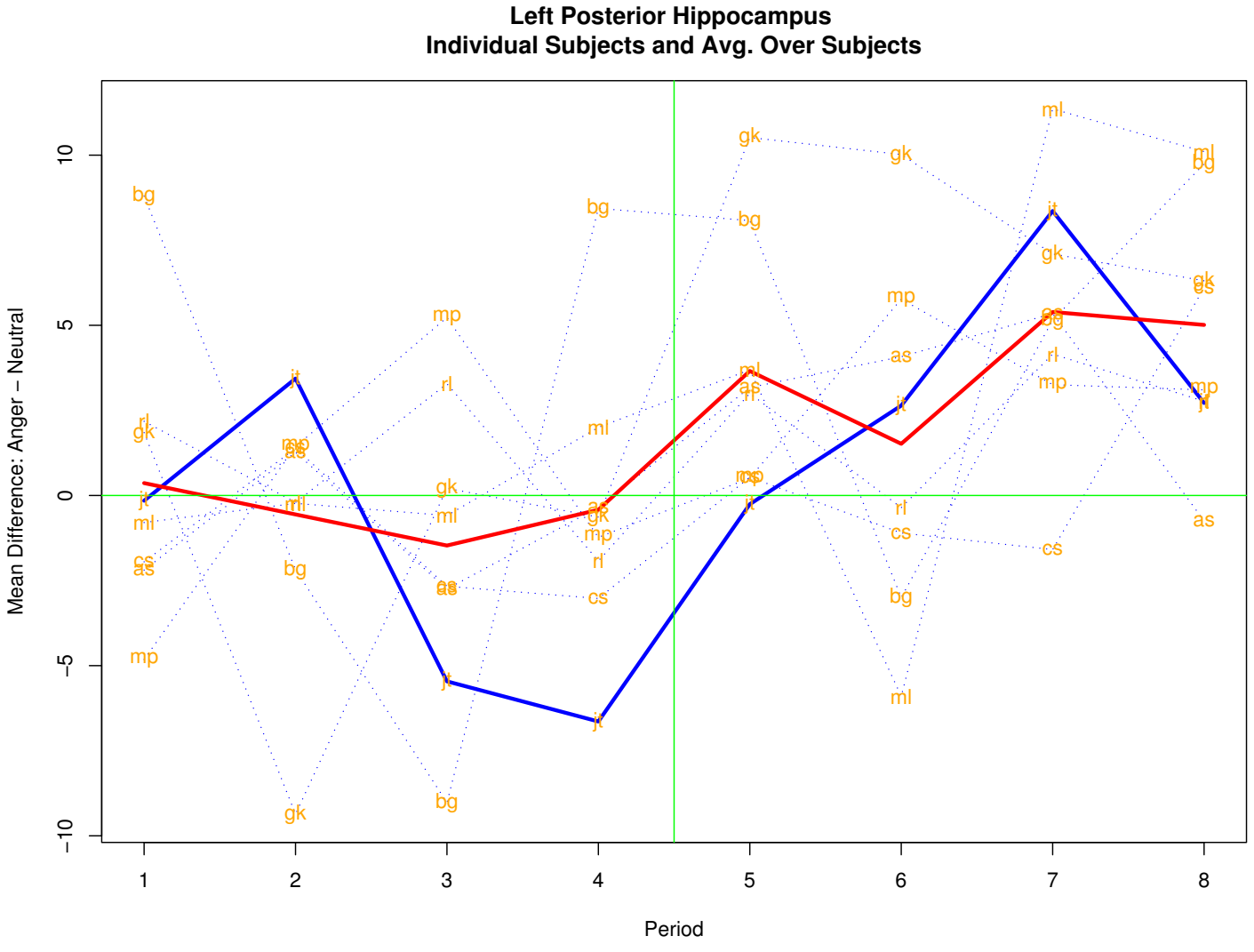
$$c_H = \exp(\beta_{N,2} - \beta_{N,1})$$

- Data from each subject are summarized by contrasts estimates and standard errors, which are used as input to a second-level Bayesian analysis.

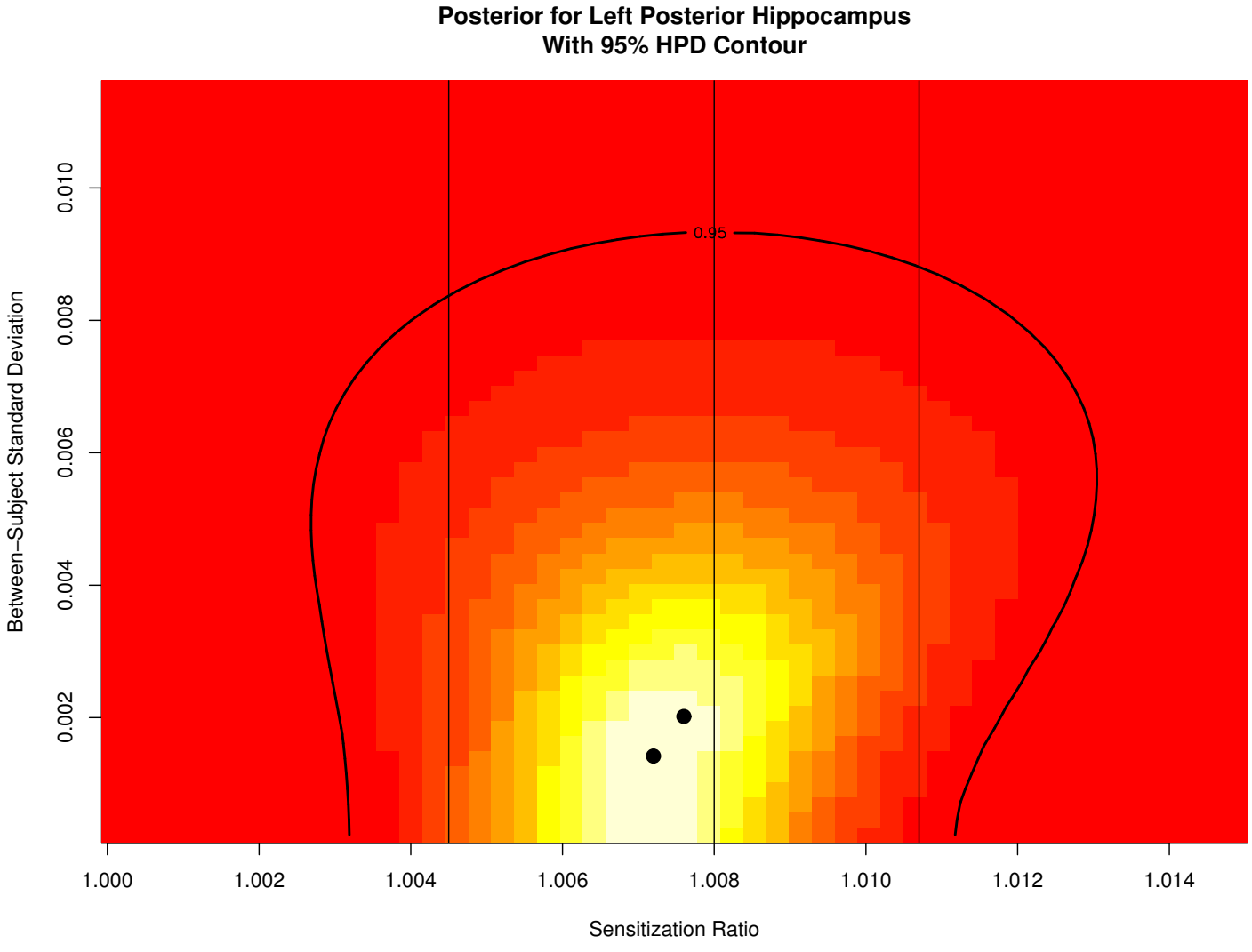
Typical ‘Raw’ BOLD Timecourse



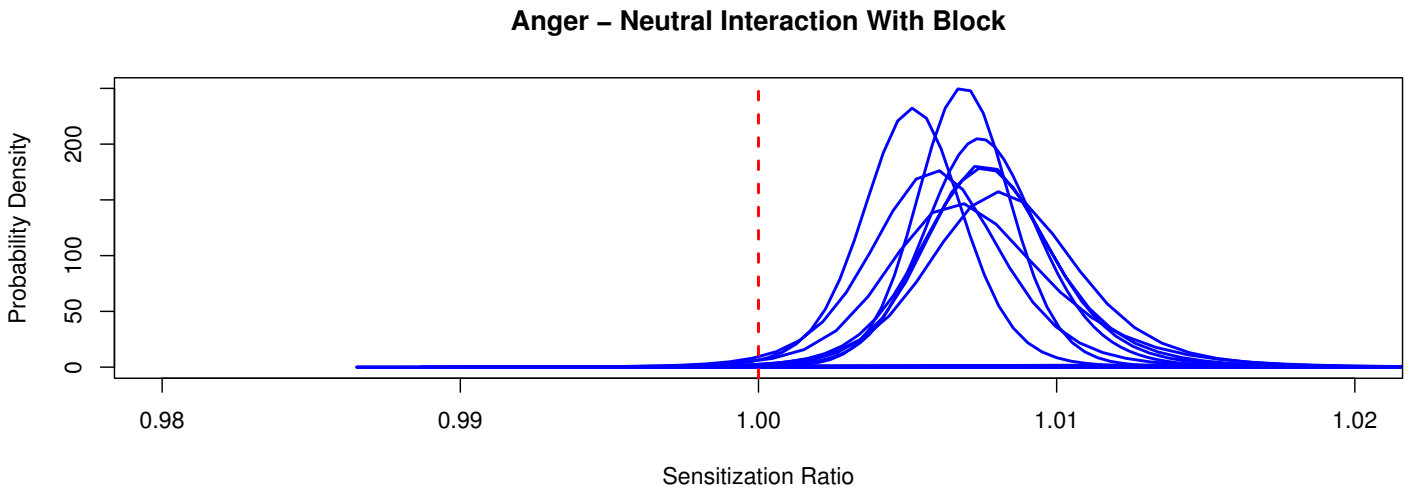
Block Averages For All Subjects

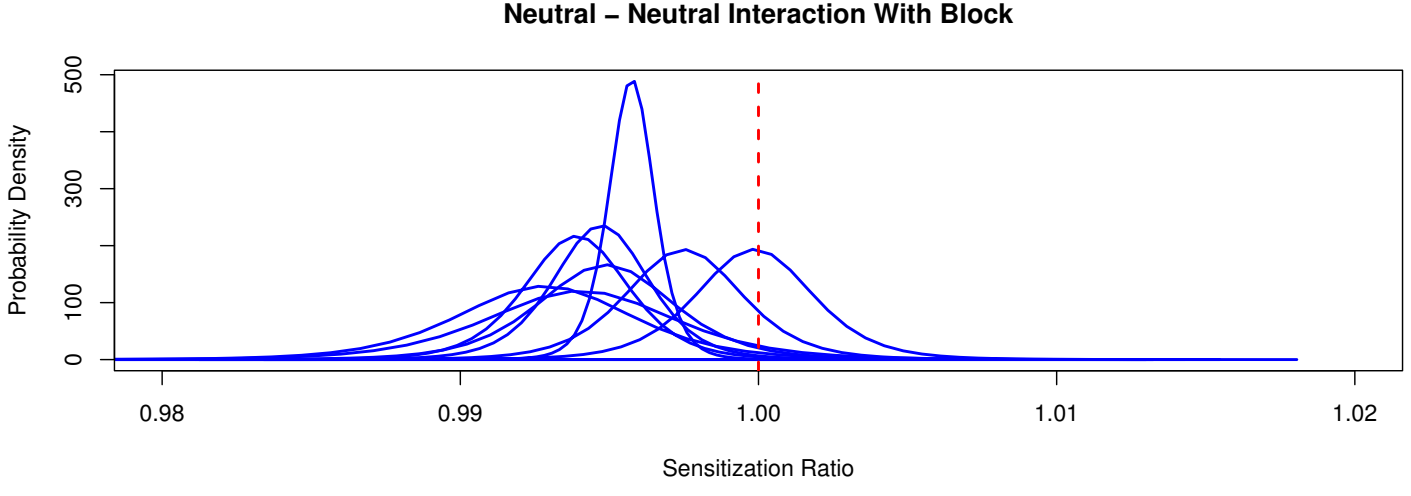


Posterior for LPHIP Sensitization



A/N: Sensitization N/N: Habituation





The Problem of Not Enough Subjects

- Random-effects 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 metholdogy. 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 under two conditions:
  - 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.
  - You *regularize* the estimate of $\hat{\sigma}_b^2$ 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 (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 activation 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.

IV. 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.

Linear Model (GLM) Assumptions

- 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 activation, 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 better.
- 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.

Error Variance Assumptions (Cont'd)

- 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. Standardized contrasts will generally be approximately t-distributed (Central Limit Theorem; if standard errors and degrees of freedom are appropriately estimated).

Gaussian Errors (Cont'd)

- 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).