Group Analysis

Martin M. Monti
UCLA Psychology

NITP

Aggregating multiple subjects

- When we conduct multi-subject analysis we are trying to understand whether an effect is "significant" across a group of people.
- Whether something is significant depends on the variance we assess it against:

Classical statistical hypothesis testing proceeds by comparing the difference between the expected and hypothesized effect against the "yardstick" of variance.

[Holmes & Friston, 1998]

Variance at the group level

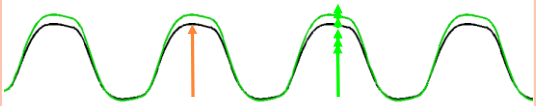
- ***Fixed Effects (FFX)***: is about the intra-subject variability. An effect is compared against the "yardstick" of the precision with which it can be measured (for each subject). The different subjects are considered to be "fixed."
- ***Random Effects (RFX)***: is about the inter-subject variability. An effect is compared against the "yardstick" of how much variability there is across different subjects. The different subjects are considered to be a "random" sample from a greater population.
- ***Mixed Effects (MFX)***: is about intra-subject & inter-subject variability.

Fixed Effects: Intra-subject variability

- Only variation (over sessions) is measurement error
- True Response magnitude is ***fixed***

Adapted from T Nichols

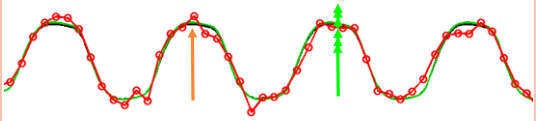
Random Effects: Inter-subject variability



- Source of variation
  - Response magnitude
- Response magnitude is ***random***
  - Each subject/session has random magnitude
  - But note, the population mean is ***fixed***

Adapted from T Nichols

Mixed Effects



- Two sources of variation
  - Measurement error
  - Response magnitude
- Response magnitude is ***random***
  - Each subject/session has random magnitude
  - But note, the population mean is ***fixed***

Adapted from T Nichols

In other words …

FFX Model:

$$y_{ij} = d_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim (0, \sigma_w^2)$$

Subj. effect; Meas. error

In other words …

FFX Model:

$$y_{ij} = d_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim (0, \sigma_w^2)$$

But d_i is a random variable!

$$d_i = d_{pop} + z_i \qquad z_i \sim (0, \sigma_b^2)$$

Population effect; Subj. variability (around dpop)

In other words …

FFX Model:

$$y_{ij} = d_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim (0, \sigma_w^2)$$

But d_i is a random variable!

$$d_i = d_{pop} + z_i \qquad z_i \sim (0, \sigma_b^2)$$

MFX Model:

$$y_{ij} = d_{pop} + z_i + \varepsilon_{ij}$$

Population effect; Subj. variability (around d_{pop}); Meas. error

In other words …

FFX Model:

$$y_{ij} = d_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim (0, \sigma_w^2)$$

But di is a random variable!

$$d_i = d_{pop} + z_i \qquad z_i \sim (0, \sigma_b^2)$$

MFX Model:

$$y_{ij} = d_{pop} + \eta \qquad (\eta = z_i + \varepsilon_{ij})$$

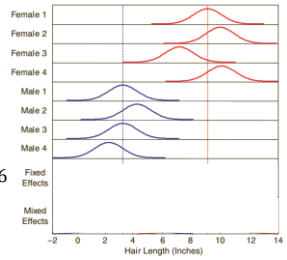
A hairy example

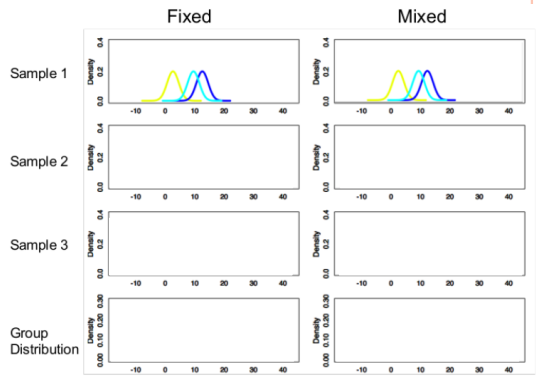
Question: Do M & F hair differ in length?
Experiment: Take 25 hairs from each of 8 Ss (4F, 4M)
[σ^2_w=1, σ^2_b=49]

$$\sigma^2_{FFX}: \frac{\sigma_w^2}{Nn} = \frac{1}{(4 \times 25)} = 0.01$$

$$\sigma^2_{RFX}: \frac{\sigma_b^2}{N} = \frac{49}{4} = 12.25$$

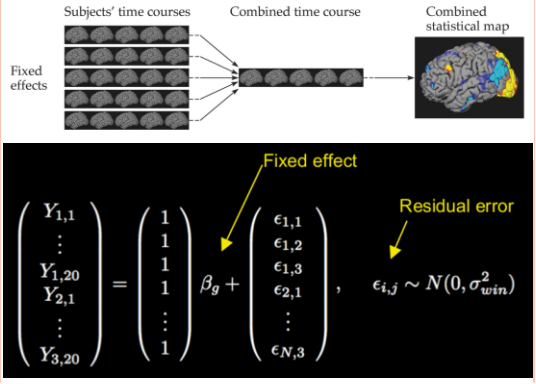
$$\sigma^2_{MFX}: \frac{\sigma_w^2}{Nn} + \frac{\sigma_b^2}{N} = \frac{49}{4} = 12.26$$





By Jeanette Mumford

Group analysis strategies: FFX



Fixed v Random

- Fixed isn't "wrong," just usually isn't of interest
- Fixed Effects Inference
  - "I can see an effect in ***this*** sample"
- Random Effects Inference
  - I can extend my inference to the population: "I expect to see the effect across the population"

15

Group analysis strategies (I): "all-in-one"

- Complete single-level GLM that relates various parameters of interest at the group level to the full set of (time series) data available

$$Y = XX_g\beta_g + \gamma$$

Concatenated timeseries

Single subject design matrix

Second level design matrix (e.g., pat v vol)

Group level parameter

error term

GROUP ANALYSIS STRATEGIES (I): "ALL-IN-ONE"

- Complete single-level GLM that relates various parameters of interest at the group level to the full set of (time series) data available

$$Y = XX_g\beta_g + X\eta_g + \varepsilon$$

Concatenated timeseries

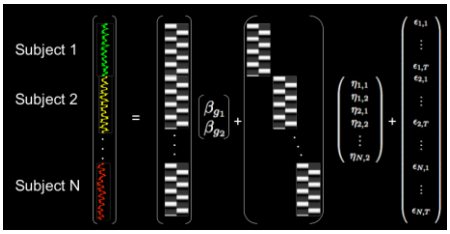
Single subject design matrix

Second level design matrix (e.g., pat v vol)

Group level parameter

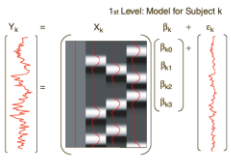
GROUP ANALYSIS STRATEGIES (I): "ALL-IN-ONE"

- Complete single-level GLM that relates various parameters of interest at the group level to the full set of (time series) data available

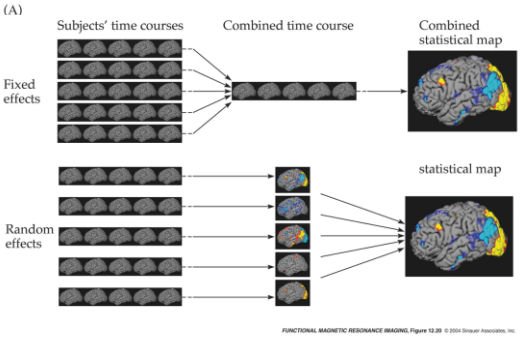


- Computationally intense approach
- What if you acquire 1 more dataset?

GROUP ANALYSIS STRATEGIES (II): THE SUMMARY STATISTIC APPROACH



GROUP ANALYSIS STRATEGIES (II): THE SUMMARY STATISTIC APPROACH



GROUP ANALYSIS STRATEGIES (II): 2ND LEVEL

1. **Perform an OLS [the SPM way]**
   - Assume that 1st level variances ($\sigma_{w_i}^2$) are the same for all subjects (i.e., *homoschedasticity*)*
   - Assume that 1st level design matrices are the same for all subjects (i.e., are *balanced*)*
   - Estimate σ_b^2 from the $(c)\widehat{\beta}_i$ carried forward from the 1st level analyses, use it to assess the average group effect. Essentially, this is a t-test!

   \+ Rapid & simple

   – Are $\sigma_{w_i}^2$ truly the same (distracted subjects, learning, …)?

   – Are 1st level matrices truly the same (forgotten v recalled)?

GROUP ANALYSIS STRATEGIES (II): 2ND LEVEL

2. **Perform a GLS (WLS) [the FSL way]**
   - Carry forward $(c)\widehat{\beta}_i$ ***as well*** as 1st level variance ($\sigma_{w_i}^2$)
   - Estimate σ_b^2 , define (for each subject *j*) the overall variance is: $\hat{\sigma}_{w_j}^2 + \hat{\sigma}_b^2$
   - Perform a GLS where each subject's (2nd level) data is weighted by her overall variance.

Act as weights

$$V_g = \begin{pmatrix} \sigma_{win_1}^2 + \sigma_g^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{win_N}^2 + \sigma_g^2 \end{pmatrix} \rightarrow W_g = \begin{pmatrix} \frac{1}{\sqrt{\sigma_{win_1}^2 + \sigma_g^2}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\sigma_{win_N}^2 + \sigma_g^2}} \end{pmatrix}$$

\+ "Bad" subjects with a large $\sigma_{w_i}^2$ will be down-weighted

\+ Statistically more correct (presumably better for more using designs beyond simple t-test)

– Computationally more intensive (iterative calculation of variance)

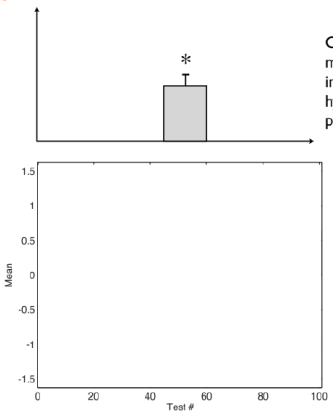
Group analysis strategies (II): The Summary Statistic approach

The debate:

- **Friston** (SPM): Assume homoscedastic 1st level variances and do an OLS.
- **Beckmann 03** (FSL): must use lower level variance in group estimation, else no longer equivalent to the all-in-one approach
- **Friston 05** (SPM): OLS is robust to unequal variances (but can estimate the covariance structure [using ReML] from first level [only significant voxels] and carry that forward).
- **Smith 05** (FSL): Within subject variability can actually be fairly large
- **Mumford 09**: OLS is robust even in the presence of outliers and violations of homoschedasticity, but only for 1 sample t-test. GLS always more optimal strategy.

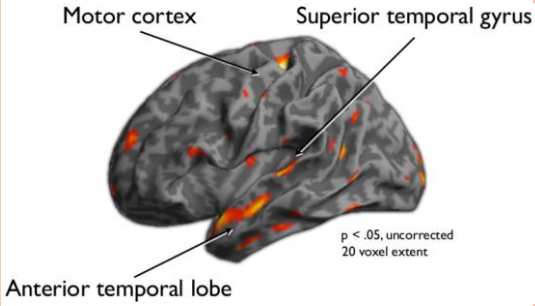
Recap

i. FFX inferences are valid, but only with respect to the sample. May be of interest for single case studies, or small rare populations you can fully sample.
ii. MFX inferences are valid over the population you sample from because you are accounting for sampling variability. This is what you want to do for a typical group study.
iii. The Summary statistic approach is efficient. Run 1st levels independently, then combine the results. If you run 1 more subject, then you only have to re-run the group.



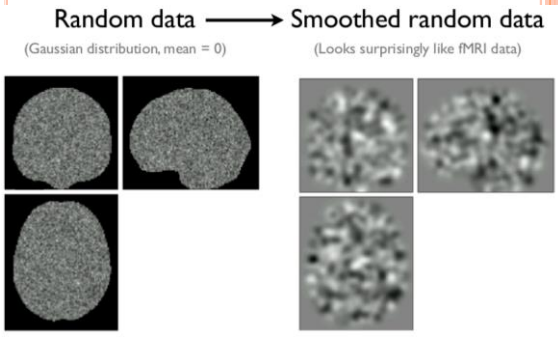
One test significant at $p < .05$ means you have a 5% chance of incorrectly rejecting the null hypothesis, resulting in a false positive

Massive Univariate Approach



Source: Jonathan Peelle

How these data were generated



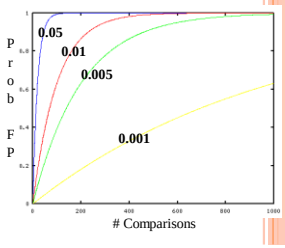
Source: Jonathan Peelle

Multiple comparisons problem

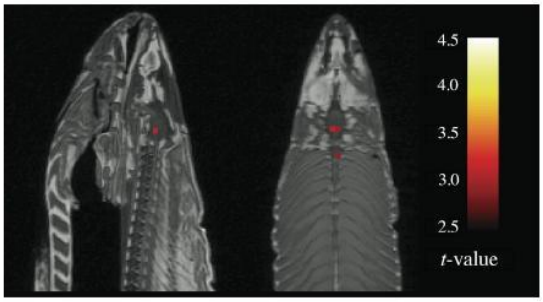
- When you make 1 test, what is the probability that a positive result is, in fact, not true (i.e., false positive)
  → α (say, 5%)
- If we make 2 tests, what is the overall probability (i.e., 'family-wise' probability) of false positives?
  → $1–(1–\alpha)^2$ (at a nominal 5%: 9.75%)
- If we make n tests, what is the overall probability (i.e., 'family-wise' probability) of false positives?
  → $1–(1–\alpha)^n$

Multiple comparisons problem

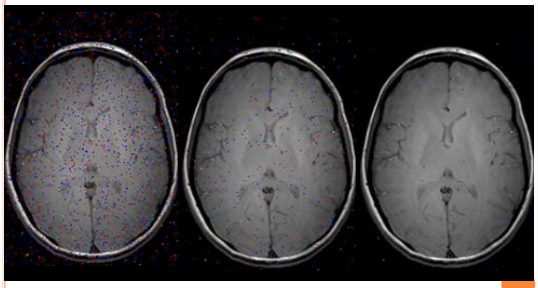
- How many tests do we perform in fMRI analysis?
- Over (say) 100,000 null voxels, how many times will we incorrectly reject H_0?
- ~5,000 voxels (on average!)



Fishy Statistics

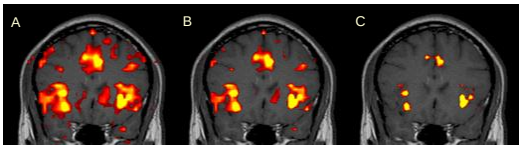


False activations under H0



P < 0.05 (1682 voxels) P < 0.01 (364 voxels) P < 0.001 (32 voxels)

How much correction?



Poor Specificity (risk of false positives)

Good Power

Good Specificity

Poor Power (risk of false negatives)

Correction for multiple comparisons

2 main strategies:

1. ***Family Wise Error (FWE):*** Control for the probability of ***any*** false positives (e.g., Bonferroni, Random Field Theory, Permutation)
2. ***False Discovery Rate (FDR):*** Control proportion of false positives ***among*** rejected tests

FWE (I): Bonferroni

- Main idea: make each individual test more stringent, so overall you end up with your total (i.e., family-wise) 'desired' false positive rate.

$$\alpha_i^{Bonf} = \frac{\alpha_{FW}}{n} \rightarrow \sum_{i=1}^{n} P(T_i > \alpha_i \mid H_0) \leq \alpha_{FW}$$

- For example:
  - Desired familywise false positive rate: $\alpha_{FW} = 0.05$
  - Total number of (independent) tests: 100,000
  - Then the Bonferroni-corrected false positive level *for each individual test* is:

$$\alpha_i^{Bonf} = \frac{\alpha_{FW}}{n} = \frac{0.05}{100{,}000} = 0.0000005$$

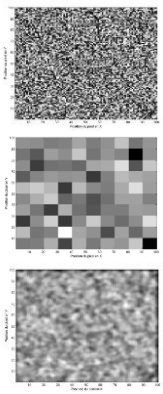
FWE (I): Bonferroni

- Assumes independent tests
- FMRI data spatially correlated (vasculature, spatial smoothing), so the number of independent tests is less than the number of voxels

→ Overly stringent

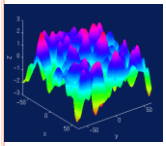
→ Increases Type II error

- Difficult to find what is n in order to calculate the correct α_{bonf}



FWE (II): Random Field Theory

- Allows to find a threshold in a set of data where it's not easy (or even impossible) to find the number of independent variables
- 3 step approach:
  i. Estimate how smooth the data is ("resels")
  ii. Compute how many peaks would be above the threshold by chance ("Euler Characteristic")
  iii. Calculate the threshold that yields desired FWER



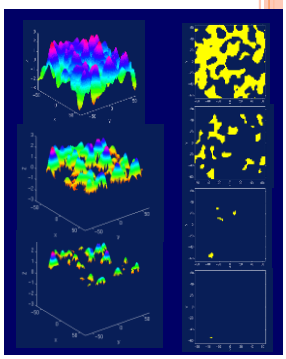
1. Smoothness Parametrization

We can't compute the # of independent voxels, but we can compute the number of resolution elements (i.e. "resels").

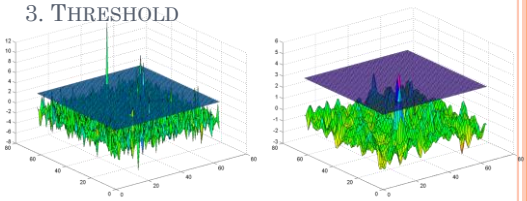
- RESELS – Resolution Elements
  - 1 RESEL = $FWHM_x \times FWHM_y \times FWHM_z$
  - RESEL Count R
    - $R = V \sqrt{|\Lambda|}$ ← *The only data-dependent part of* $E(\chi_u)$
    - Volume of search region in units of smoothness
    - Eg: 10 voxels, 2.5 voxel FWHM smoothness, 4 RESELS
- RESELs not # of independent 'things' in the image

2. Euler Characteristic

- Topological measure [χ]
- Threshold an image at u
- EC = # of blobs - # holes
- At high u:
  EC = # of blobs
  P(blob) = E[EC]
- Under H0, α_{FWE} = E[EC]



3. Threshold



$$\alpha_{FW} = E[\chi] = R(4\log_e 2)(2\pi)^{\frac{-3}{2}} Z_u e^{\frac{-1}{2}Z_u^2}$$

Given the smoothness of my data (R), what threshold (Z) do I need to set so that I have α_{FW} chance (~E[EC]) of having peak above threshold?

False Discovery Rate (FDR)

- FDR controls the expected proportion of false positive values among supra-threshold values (i.e., false claims v false tests):
- p < 0.05 FWE means: There is only a 5% chance any result is a false positive.
- p < 0.05 FDR means: No more than 5% of active voxels are false positives.

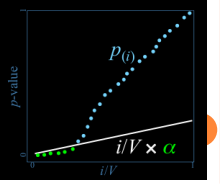
False Discovery Rate (FDR)

Benjamini & Hochberg Procedure

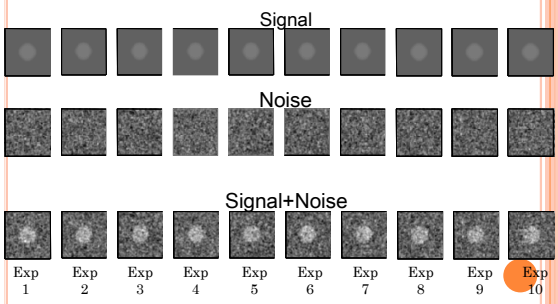
- Select desired limit α on FDR
- Order p-values, $p_{(1)} \le p_{(2)} \le \ldots \le p_{(V)}$
- Let r be largest i such that

$$p_{(i)} \le i/V \times \alpha$$

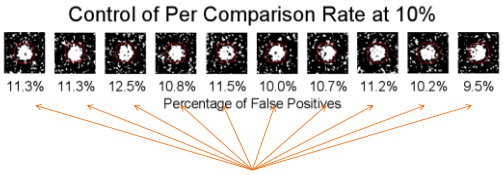
- Reject all hypotheses corresponding to $p_{(1)}, \ldots, p_{(r)}$.



Comparing correction methods



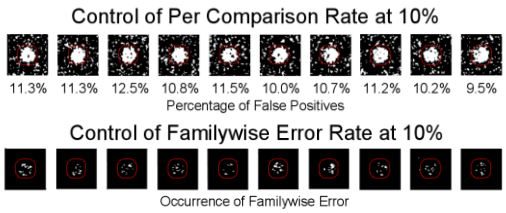
No correction ($\alpha = 0.1$)



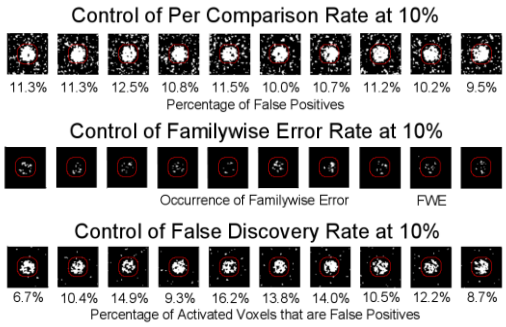
On average, 10% of the 'false' voxels are incorrectly declared "active."

In *each* experiment we have about 10% false alarms

FWE ($\alpha = 0.1$)



FDR ($\alpha = 0.1$)



Resources

- Monti M.M. (2011) Statistical analysis of fMRI time-series: A critical evaluation of the GLM approach. *Frontiers in Human Neuroscience*, 5(28).
- Mumford, J. A., and Nichols, T. (2009). Simple group fMRI modeling and inference. Neuroimage 47, 1469–1475.
- Mumford, J. A., and Poldrack, R. A. (2007). Modeling group fMRI data. Soc. Cogn. Affect. Neurosci. 2, 251–257.
- Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in fMRI. Neuroimage 20, 1052–1063. .
- Poldrack R.A., Mumford J.A., Nichols T.E. (2011) Handbook of Functional MRI Analysis, Cambridge University Press.
- Lazar, N. (2008). The statistical analysis of functional MRI data. Springer.
- Friston K.J., et al Statistical Parametric Mapping: The Analysis of Functional Brain Images, chapter 8.