Group Analysis

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

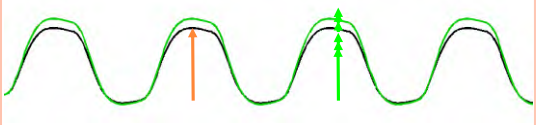
- ***Fixed Effects (FFX)***: is about the intra-subject variability. An effect is compared against the "yardstick" of the precision with which it can be measured (for each subject). The different subjects are considered to be "fixed."
- ***Random Effects (RFX)***: is about the inter-subject variability. 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 di 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 di is a random variable!

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

MFX Model:

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

Population effect; Subj. variability (around dpop); Meas. error

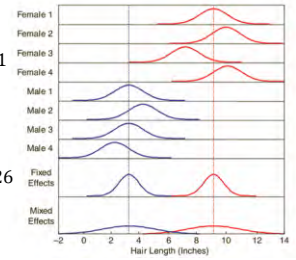
A hairy example

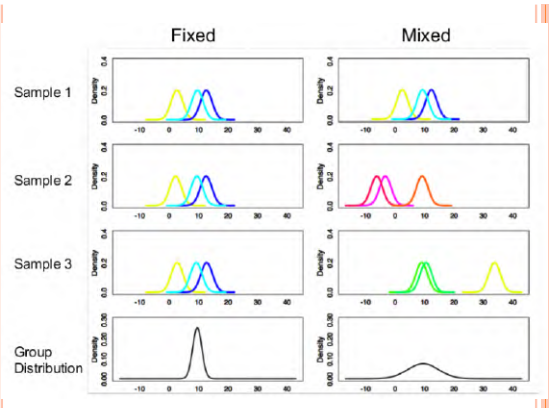
Question: Do M & F hair differ in length?
Experiment: Take 25 hairs from each of 8 Ss (4F, 4M)
[σ^2_w=1, σ^2_b=49]

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

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

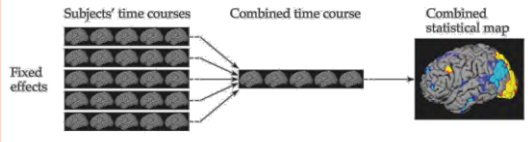
$$\sigma^2_{MFX}: \frac{\sigma_w^2}{Nn} + \frac{\sigma_b^2}{N} = \frac{49}{4} = 12.26$$

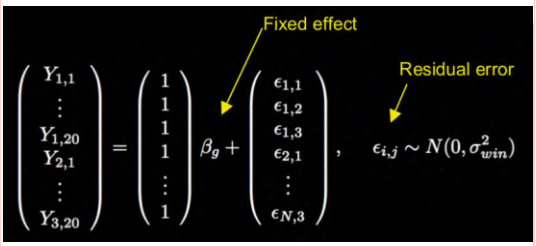




By Jeanette Mumford

Group analysis strategies: FFX





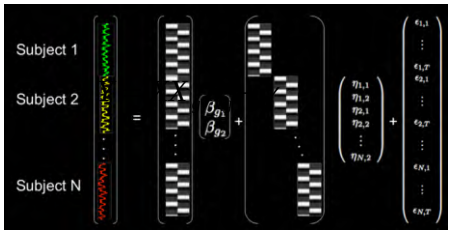
Fixed v Random

- Fixed isn't "wrong," just usually isn't of interest
- Fixed Effects Inference
  - "I can see this effect in this cohort"
- Random Effects Inference
  - "I can extend my inference to the population"

14

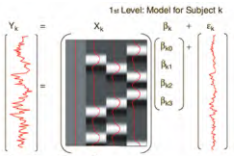
Group analysis strategies (I): "all-in-one"

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

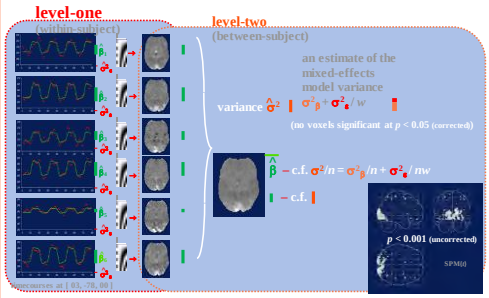


- Computationally intense approach
- What if you acquire 1 more dataset?

Group analysis strategies (II): The Summary Statistic approach



Holmes & Friston: Random Effects



Adapted from T Nichols

Group analysis strategies (II): The Summary Statistic approach

$$\begin{cases} Y = X\beta + \varepsilon & [1^{st}\, lvl\ FFX] \end{cases}$$

$$\rightarrow Y = XX_g\beta_g + X\varepsilon_g + \varepsilon$$

Group analysis strategies (II): The Summary Statistic approach

$$\begin{cases} Y = X\beta + \varepsilon & [1^{st}\, lvl\ FFX] \\ \beta = X_g\beta_g + \varepsilon_g & [2^{st}\, lvl\ RFX] \end{cases}$$

$$\begin{cases} Y = X\beta + \varepsilon & [1^{st}\, lvl\ FFX] \\ \hat{\beta} = X_g\beta_g + \eta & [2^{st}\, lvl\ MFX] \end{cases}$$

$$\eta = \frac{\sigma_w^2}{n} + \sigma_b^2$$

Group analysis strategies (II): The Summary Statistic approach

SPM (I): Assume homoscedastic 1st level variances and do an OLS

To maintain equivalence with all-in-one model assume:
1. first level variances must be equal (σ_w^2)
2. First level X matrices must be the same (i.e., "balanced" for all subjects)

Group analysis strategies (II): The Summary Statistic approach

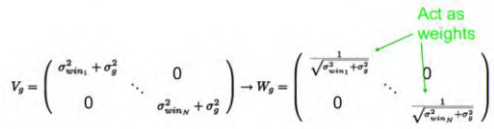
SPM (I): Assume homoscedastic 1st level variances and do an OLS.
FSL: Carry forward c/β estimates ***and*** covariance matrix. (i.e., do a GLS)

$$RFX: \beta = X_g \beta_g + \varepsilon_g \qquad \varepsilon_g \sim (0, \sigma_b^2 I)$$

Group analysis strategies (II): The Summary Statistic approach

SPM (I): Assume homoscedastic 1st level variances and do an OLS.
FSL: Carry forward c/β estimates ***and*** covariance matrix. (i.e., do a GLS)

$$RFX: \beta = X_g \beta_g + \varepsilon_g \qquad \varepsilon_g \sim (0, \sigma_b^2 V_g)$$



Group analysis strategies (II): The Summary Statistic approach

SPM (I): Assume homoscedastic 1st level variances and do an OLS.
FSL: Carry forward c/β estimates ***and*** covariance matrix. (i.e., do a GLS)
SPM (II): Estimate the covariance structure (ReML) from first level (only significant voxels) and carry that forward.

Group analysis strategies (II): The Summary Statistic approach

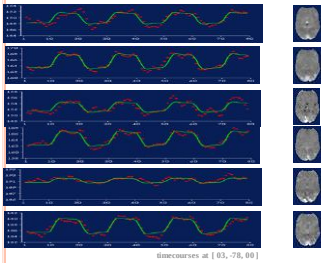
SPM (I): Assume homoscedastic 1st level variances and do an OLS.
FSL: Carry forward c/β estimates ***and*** covariance matrix. (i.e., do a GLS)
SPM (II): Estimate the covariance structure (ReML) from first level (only significant voxels) and carry that forward.
Data:

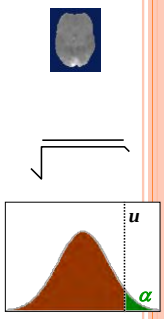
Beckmann 03 (FSL): must weight variances
Friston 05 (SPM): OLS is robust to unequal variances
Mumford 09: OLS is robust, but only for 1 sample t-test, GLS always more optimal strategy.

Recap

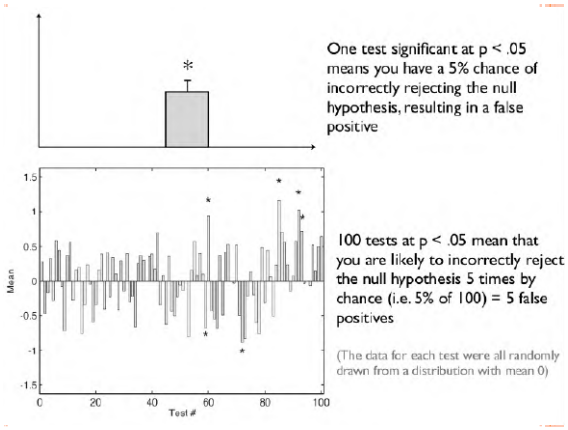
i. FFX inferences are valid, but only with respect to the sample. May be of interest for single case studies, or small rare populations you can fully sample.
ii. MFX inferences are valid over the population you sample from because you are accounting for sampling variability. This is what you want to 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.

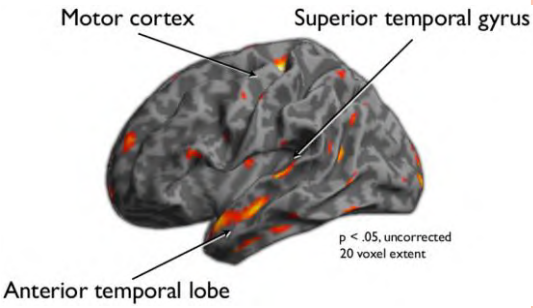
Massive Univariate Approach





$\alpha = p(t > u \mid H_0)$



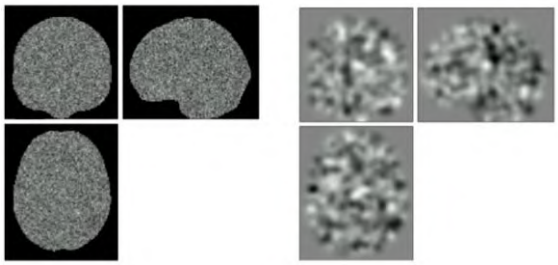


Source: Jonathan Peelle

How these data were generated

Random data ⟶ Smoothed random data

(Gaussian distribution, mean = 0) (Looks surprisingly like fMRI data)



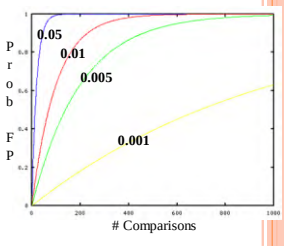
Source: Jonathan Peelle

Multiple comparisons problem

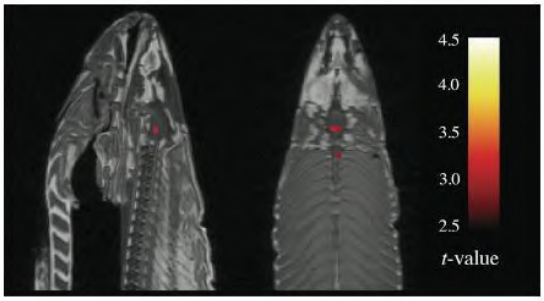
- When you make 1 test, what is the probability that a positive result is, in fact, not true (i.e., false positive)
  → α (say, 5%)
- If we make 2 tests, what is the overall probability (i.e., 'family-wise' probability) of false positives?
  → $1–(1–\alpha)^2$ (at a nominal 5%: 9.75%)
- If we make *n*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

Multiple comparisons problem

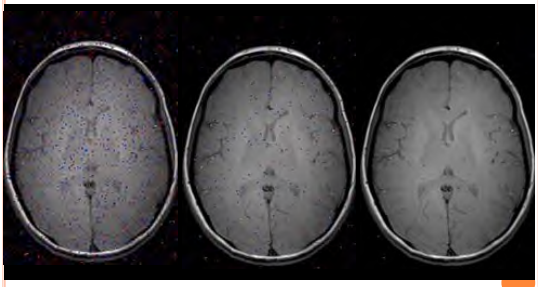
- How many tests do we perform in fMRI analysis?
- Over (say) 100,000 null voxels, how many times will we incorrectly reject H_0?
- ~5,000 voxels (on average!)



Fishy Statistics

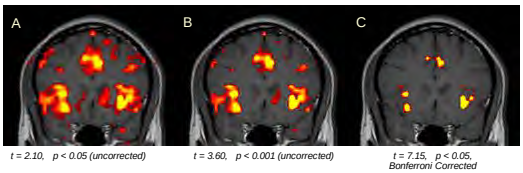


False activations under H0



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

How much correction?



t = 2.10, p < 0.05 (uncorrected) t = 3.60, p < 0.001 (uncorrected) t = 7.15, p < 0.05, Bonferroni Corrected

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:

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

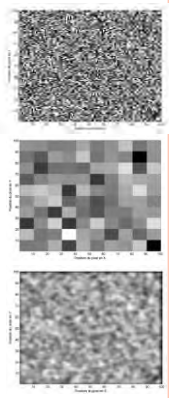
FWE (I): Bonferroni

- 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

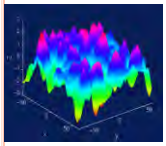
→ Increases Type II error

- 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



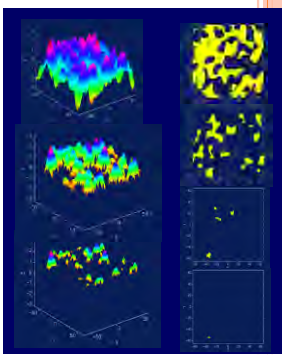
1. Smoothness Parametrization

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

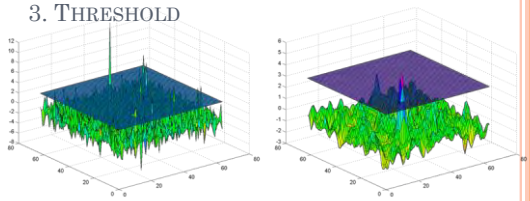
- RESELS – Resolution Elements
  - 1 RESEL = $FWHM_x \times FWHM_y \times FWHM_z$
  - RESEL Count R
    - $R = V \sqrt{|\Lambda|}$ ← *The only data-dependent part of* $E(\chi_u)$
    - Volume of search region in units of smoothness
    - Eg: 10 voxels, 2.5 voxel FWHM smoothness, 4 RESELS
- RESELs not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

2. Euler Characteristic

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



3. Threshold



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

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.

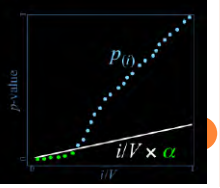
False Discovery Rate (FDR)

Benjamini & Hochberg Procedure

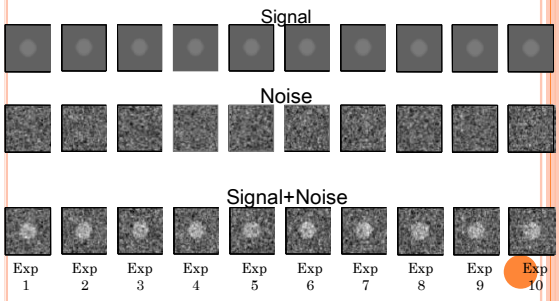
- Select desired limit α on FDR
- Order p-values, $p_{(1)} \le p_{(2)} \le \ldots \le p_{(v)}$
- Let r be largest i such that

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

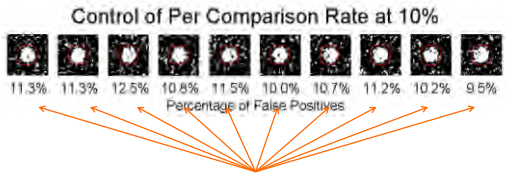
- Reject all hypotheses corresponding to $p_{(1)}, \ldots, p_{(r)}$.



Comparing correction methods



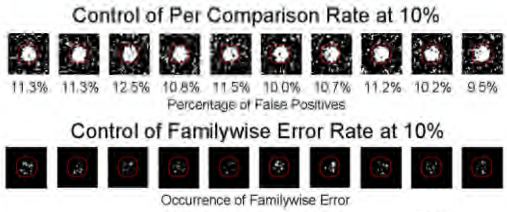
No correction (α = 0.1)



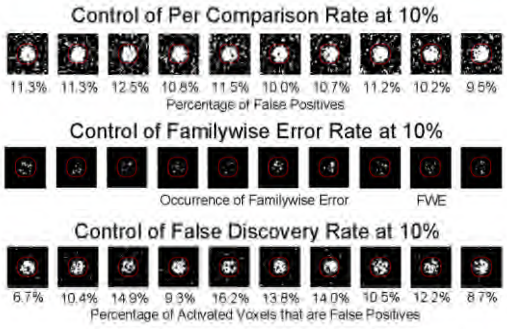
On average, 10% of the 'false' voxels are incorrectly declared "active."

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

FWE (α = 0.1)



FDR (α = 0.1)



Real Data: FWE (RFT) v FDR