Finding Activations: Power, Specificity, and Selection Bias

Thomas Nichols, Ph.D.
Department of Statistics &
Warwick Manufacturing Group
University of Warwick

with
Jeanette Mumford, PhD
University of Texas, Austin

OHBM 2011 – Advanced fMRI Course – 26 June, 2011

Outline

- Specificity, Sensitivity & Power
  - How many subjects to get a good result
- Circularity
  - How to guarantee a good (but meaningless) result

Why Power Analysis?

- To answer
  - How many subjects do I need for my study?
  - How many runs per subject should I collect?
- For grants
  - Reviewers
    - Want it (even if they don't believe/understand it)
  - Funders
    - Don't want to waste money on studies likely to fail,
    - Or on studies where fewer subjects would suffice

Test Outcome: One Test

Truth (unobserved) \ Test Decision	Below Thresh.	Above Thresh.
No Signal	☺	☹
Signal Present	☹	☺

Test Outcome: One Test

Test Decision

Truth (unobserved)		Don’t Reject H_0	Reject H_0
	H_0 True	Correct	Type I Error
	H_0 False	Type II Error	Correct

Test Outcome: Long Run

The probability of each kind event occurring

Test Decision

Truth (**un**observed)

	Don't Reject H_0	Reject H_0
H_0 True	Specificity	α
H_0 False	β	Sensitivity Power

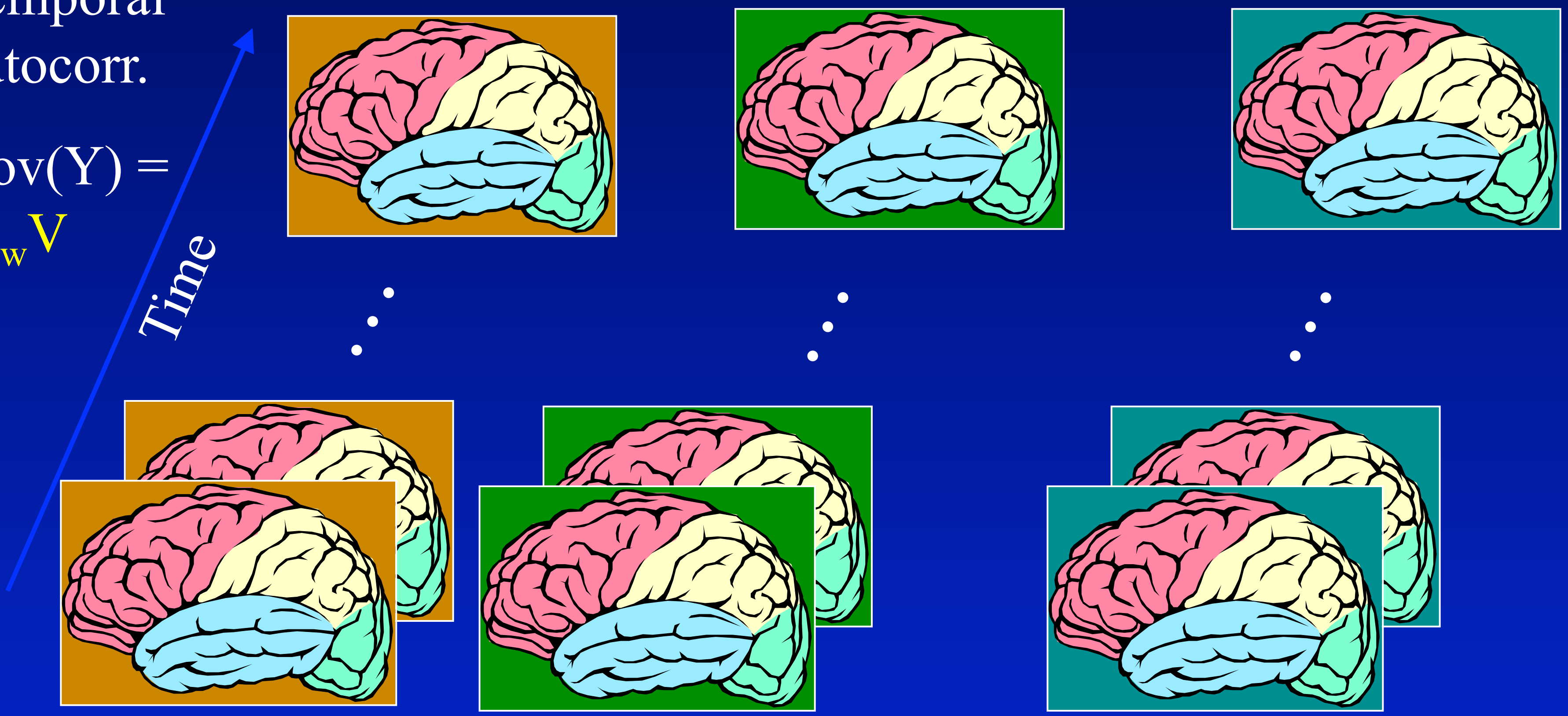
Power Analysis: Necessary information

- N Number of Subjects
  - Adjusted to achieve sufficient power
- α The size of the test you'd like to use
  - Commonly set to 0.05 (5% false positive rate) for ROI
  - Set to corrected level for voxel-wise
- Δ The size of the effect to be detected
  - Based on intuition or similar studies
- σ2 The variance of Δ
  - Has a complicated structure with very little intuition
  - Depends on many things

Why is it so difficult for group fMRI?

Temporal autocorr.

Cov(Y) = $\sigma^2_w V$

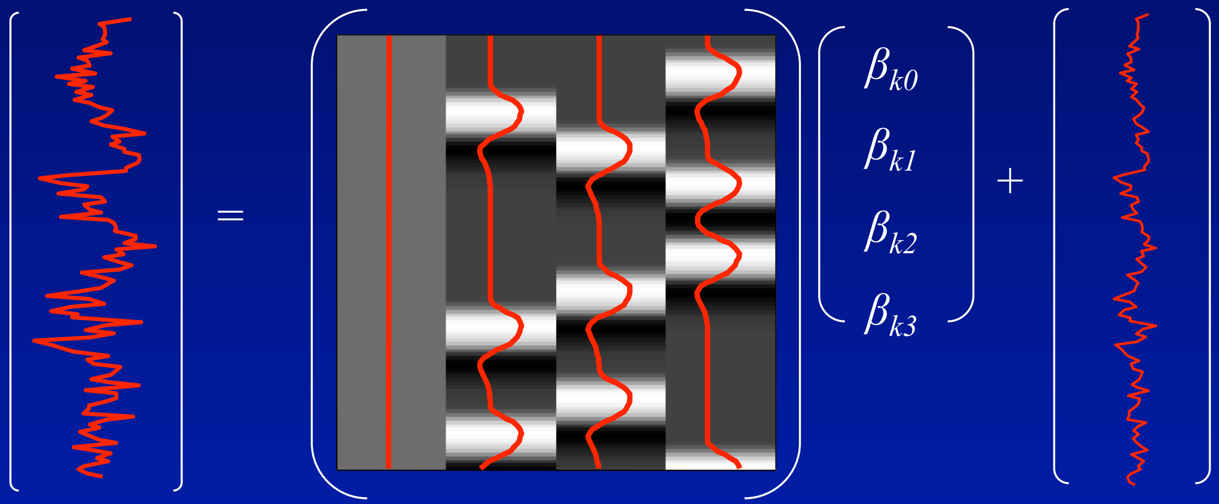


Subject 1 Subject 2 . . . Subject N

Between subject variability, σ^2_B

Level 1: Intra-Subject

$$Y_k \quad = \quad X_k \quad \beta_K \quad + \quad \varepsilon_k$$

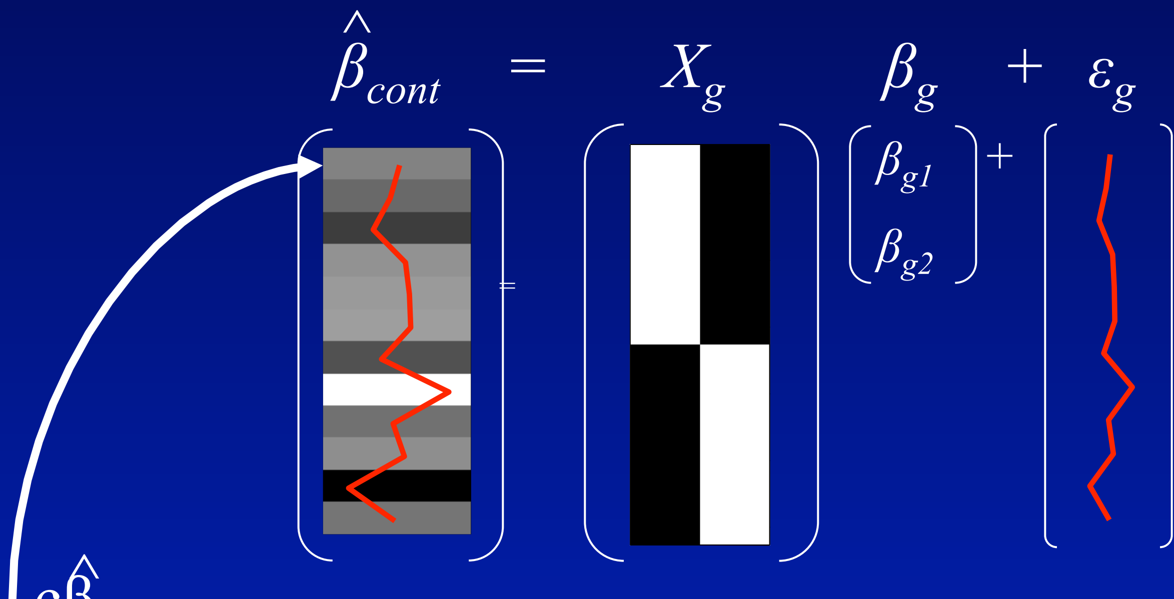


- Y_k : T_k-vector timeseries for subject k
- X_k : $T_k \times p$ design matrix
- β_k : p-vector of parameters
- ε_k : T_k-vector error term, $Cov(\varepsilon_k) = \sigma^2_k V_k$

Level 1: Autocorrelation

- What's your V_k ($N \times N$ matrix)?
  - Who has intuition on magnitude of autocorrelation?
  - True V_k is very complicated
    - SPM uses per-subject, global, cheap & cheerful AR (1) approximation ($\rho \approx 0.2$)
    - FSL uses per-subject, local, tapered & spatially regularized arbitrary Autocorrelation function
- AR(1) + White Noise
  - Estimated from residuals
  - Specified by σ_{WN}, σ_{AR} & ρ

Level 2: Between



- $c\hat{\beta}_k$
- X_g : $N \times p_g$ design matrix
- β_g : p_g-vector of parameters
- ε_g : N-vector error term
  - $\mathrm{Cov}(\varepsilon_g)=V_g=\underbrace{\mathrm{diag}\{c(X^T_k V_k^{-1} X_k)^{-1}\sigma_k^2 c^T\}}_{\text{Within subject variability}} + \underbrace{\sigma_B^2 I_N}_{\text{Between subject variability}}$

Estimating Parameters

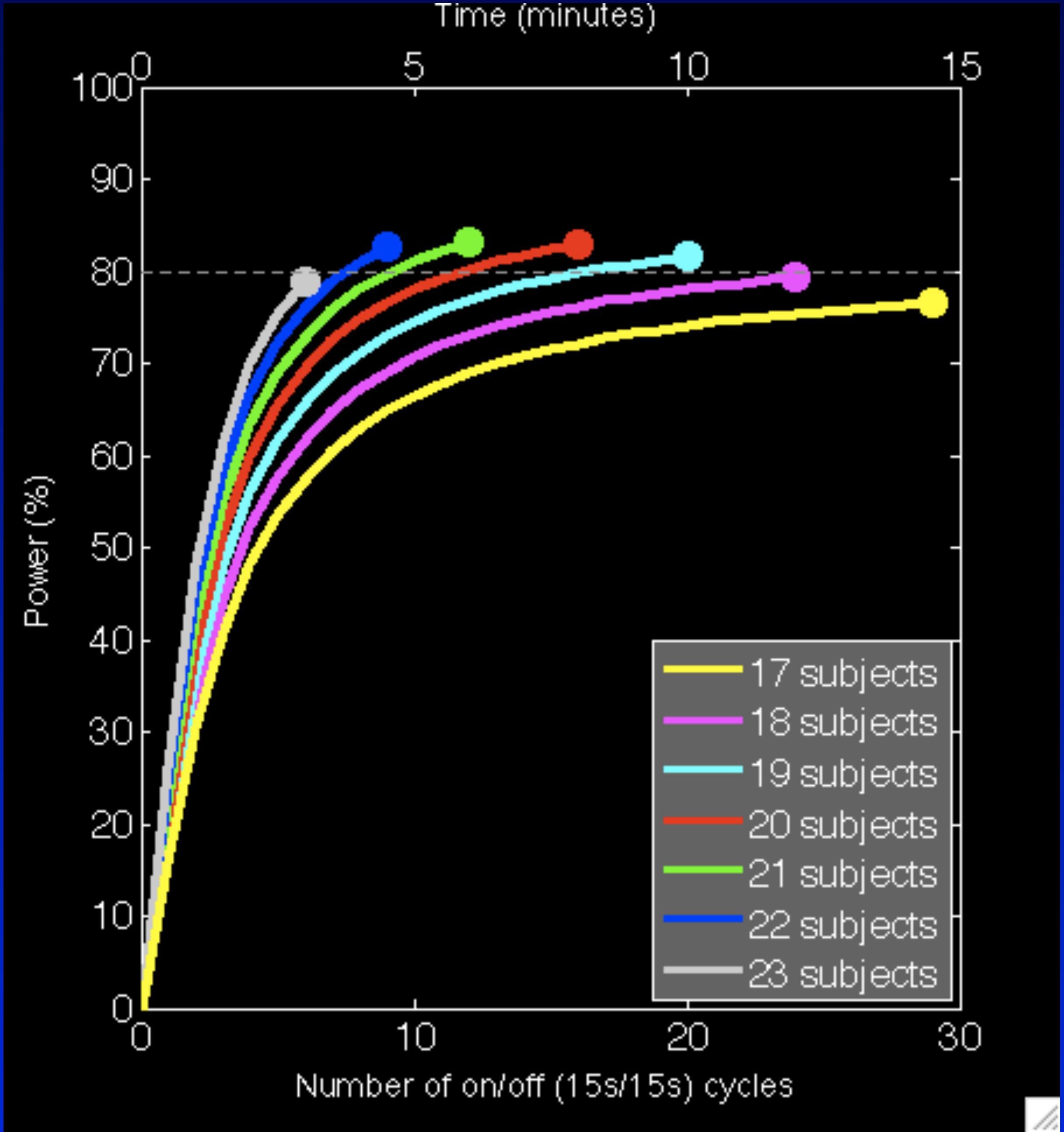
- How to you estimate parameters for a future study?
  - Look at other people's study results for similar studies
    - Usually not enough data reported
  - Look at your own similar studies
- Average parameter estimates over ROIs of interest
- With all this in place, can then estimate

$$\text{Power} = P(\, T(c\beta) > t_\alpha \mid H_A)$$

Model

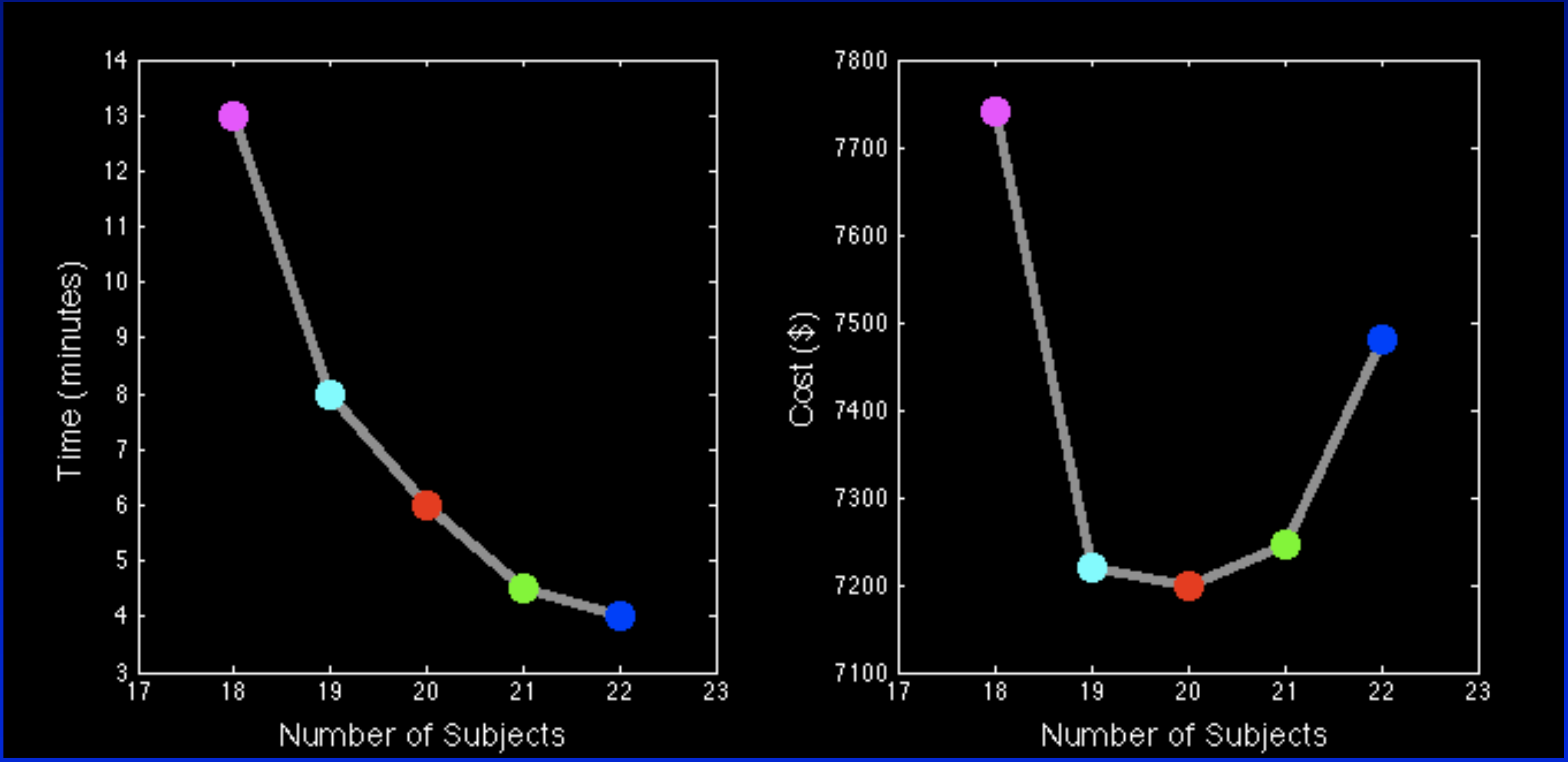
- Block design 15s on 15s off
- TR=3s
- Hrf: Gamma, sd=3
- Parameters estimated from Block study
  - FIAC single subject data
  - Read 3 little pigs
    - Same/different speaker, same/different sentence
    - Looked at blocks with same sentence same speaker
- $\delta = 0.69\%$ $\sigma_g = 0.433\%$
  $\rho = 0.73$, $\sigma_{AR} = 0.980\ \%$, $\sigma_{WN} = 1.313\%$
  $\alpha = 0.005$

Power as a function of run length and sample size



More importantly….cost!

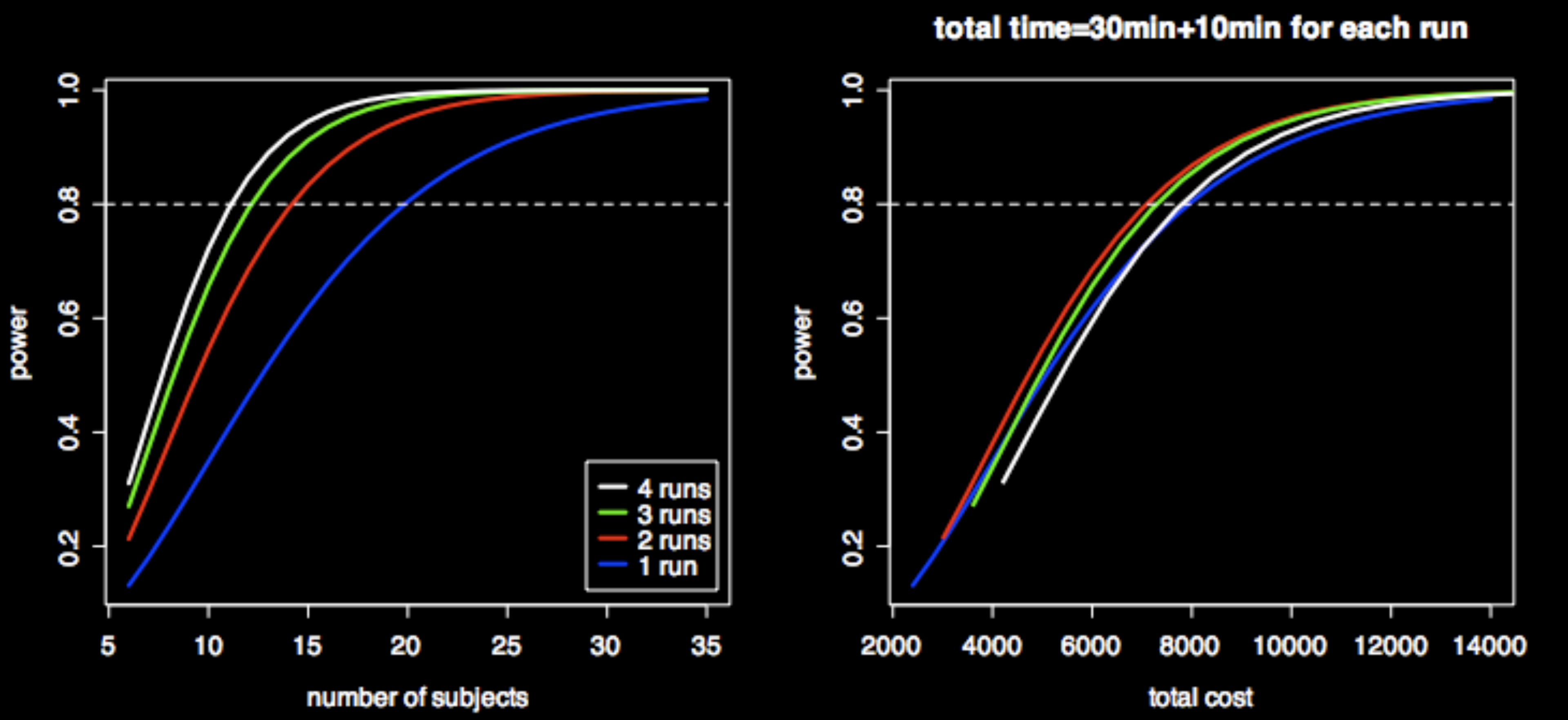
- Cost to achieve 80% power
- Cost=$300 per subject+$10 per each extra minute



How Many Runs?

- Can also expand to a 3 level model and study impact of adding runs
- Example
  - ER study
  - Study used 3 runs per subject
  - Estimate between run variability
  - Assume within subject variability is the same across subjects
  - Assume study design is same across subjects

How many runs?

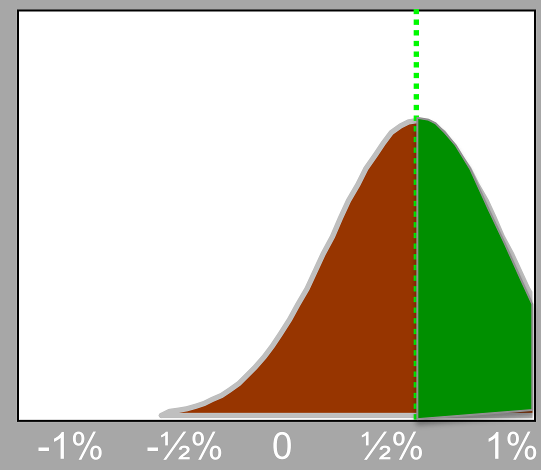


fMRI Power Calculator

- Fmripower by Jeanette Mumford
  - Beta version at fmripower.org
  - ROI based power analysis
  - Works with FSL – SPM version coming soon!
  - Runs in Matlab
  - Current version only allows user to specify different #'s of subjects
    - Assumes # of runs for future study will be the same
    - Assumes between subject variability is same across subjects
    - Doesn't control for multiple comparisons

Post hoc power?

- Power
  - A prediction about a future study
- Irrelevant for completed study
  - Null was true, or it wasn't
  - Study negative:
    - Post hoc power < 50%
  - Study positive:
    - Post hoc power > 50%
- See
  - "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," Hoenig et al, American Statistician, 55(1), 1-6, 2001



Distribution of X-bar
mean percent BOLD change
over subjects

fMRI Power
The Easy Approach

- One sample setting
- Set power for a priori ROI
  - Find paper with ROI non-voodoo results
  - Find effect size
    - $\delta = \bar{x} / \sigma$
    - Or, reverse engineer δ from t
      $t = \bar{x} / (\sigma/\sqrt{n}) \Rightarrow \delta = t / \sqrt{n}$
  - Specify alpha, n $\rightarrow$ Get power!
  - DONE!

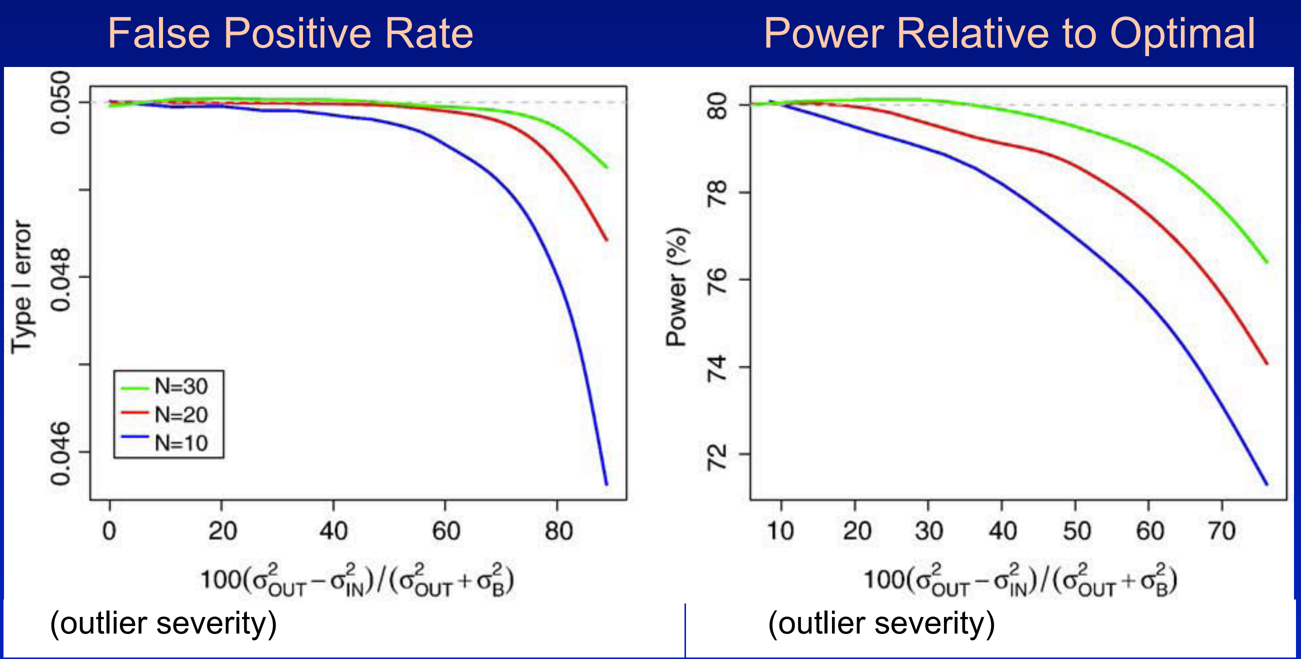
Problems with the Easy Approach

- Only works for 1-sample t-test
- Assumes same experimental design
  - Same experimental efficiency
  - Same run-length
- Assumes same intrasubject noise level
  - Same Tesla, etc

- These limitations motivated Mumford & Nichols (2009) work

Dumb Group Modelling: Not bad, actually

- One-sample t on contrasts vs. Full MFX modelling
  - "Holmes & Friston" almost impossible to break



Mumford & Nichols. Simple group fMRI modeling and inference. *Neuroimage*, 47(4):1469--1475, 2009.

- 2-sample & correlation might give trouble
  - Dramatic imbalance or heteroscedasticity

Power Conclusion

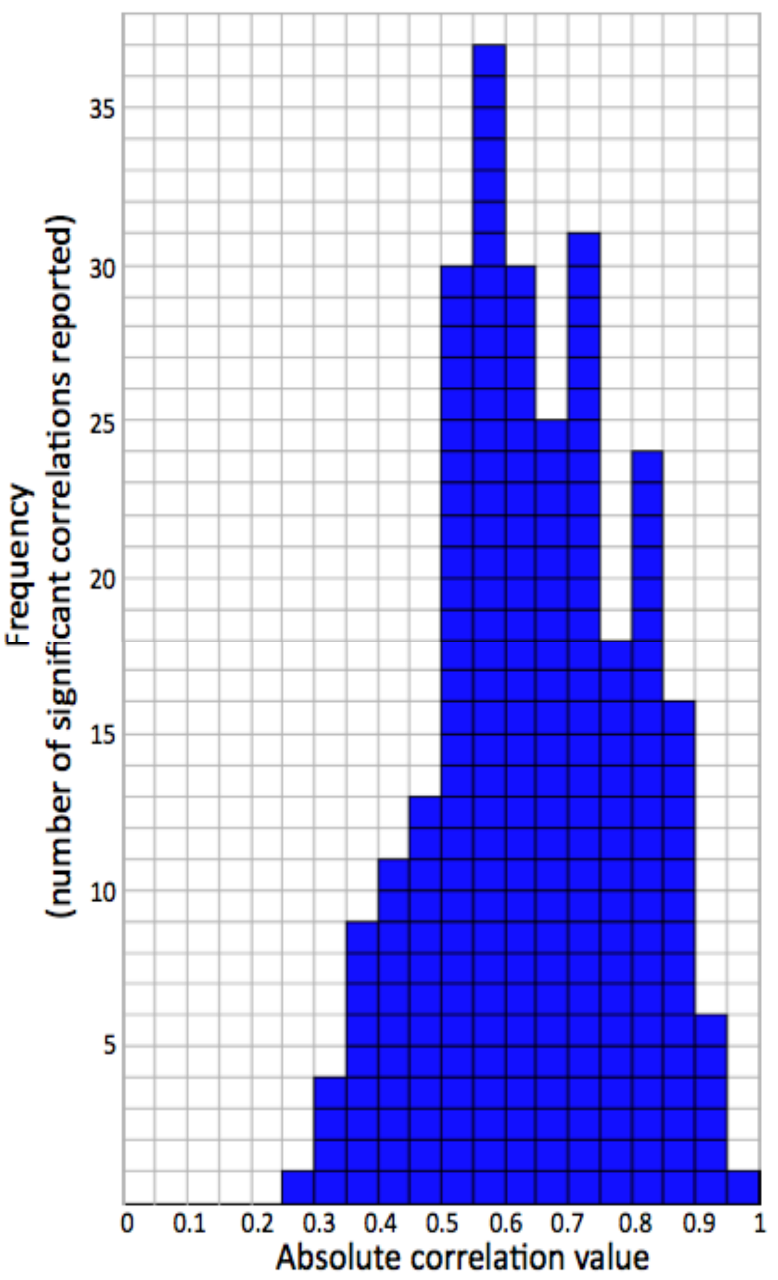
- Power Calculations
  - Useful fiction to keep statisticians employed, grant panels happy
    - You have to assume knowledge of the outcome of the experiment!
- Still, some utility
  - Are very subtle effects detectable at all
  - Relative comparisons of design efficiency

Outline

- Specificity, Sensitivity & Power
  - How many subjects to get a good result
- Circularity
  - How to guarantee a good (but meaningless) result

Voodoo Correlations

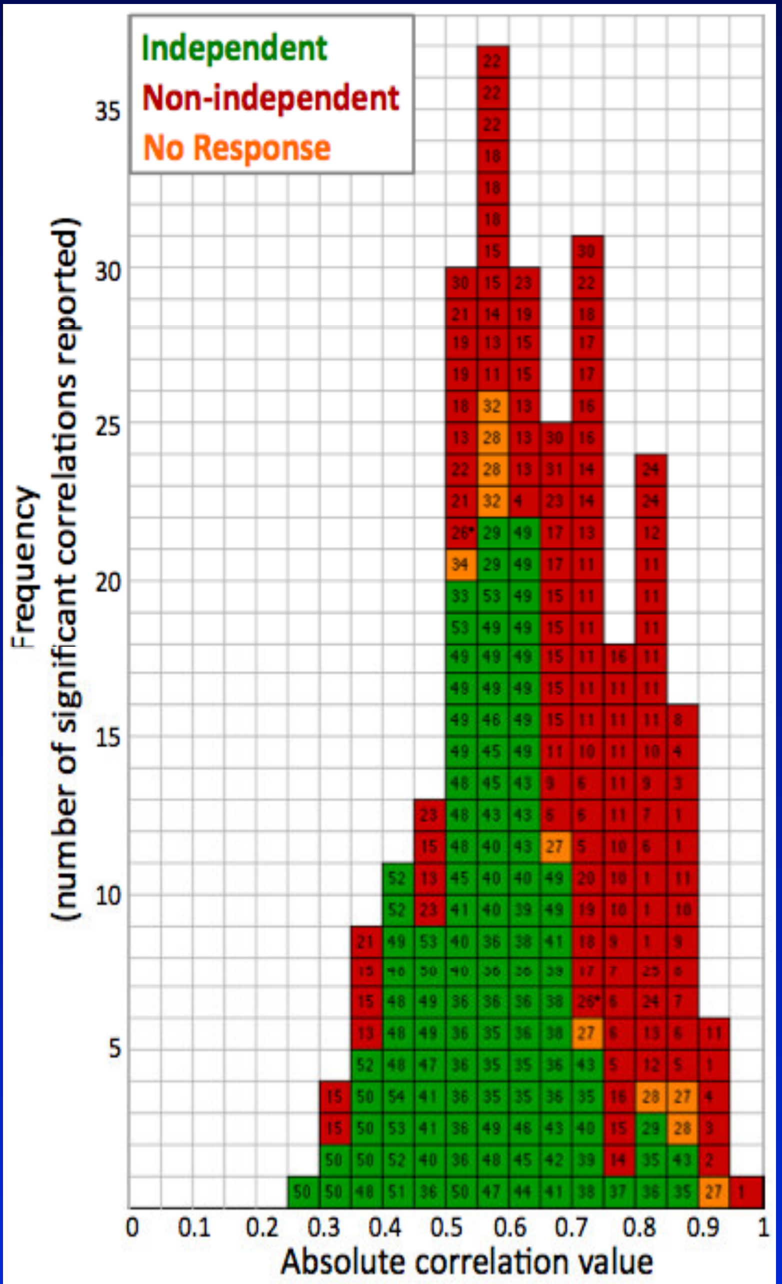
- Meta-analysis
  - Correlations from 54 social-neurosicence pubs
- **All** correlations



Vul et al, (2009). "Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition", *Persp. Psych. Sci*, 4, 274-290, 2009.

Voodoo Correlations

- Meta-analysis
  - Correlations from 54 social-neurosicence pubs
- Non-independent results
  - Peak correlations
- Independent results
  - Correlations from a priori masks



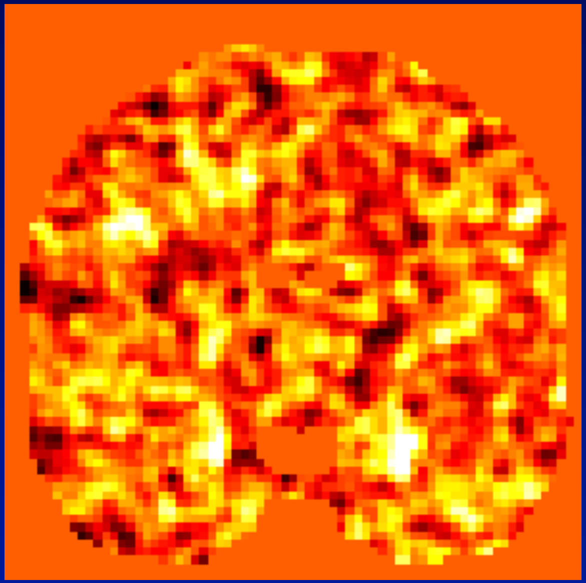
Vul et al, (2009). “Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition”, *Persp. Psych. Sci*, 4, 274-290, 2009.

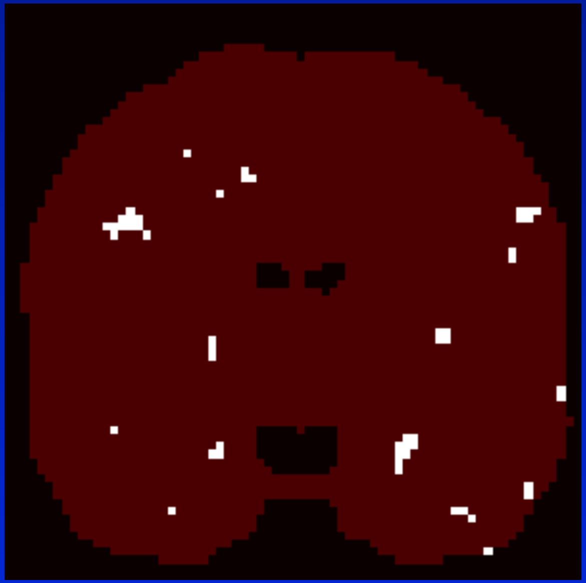
Correlation Bias Explained

- Apples & Oranges
  - Single selected correlation (e.g. GSR & personality score)
    not comparable with
    Search over 100,000 correlations (e.g. best of GSR & 100,000 voxels)
- Sample Correlations ≠ True Correlations
  - Estimated r's bounce around true ρ
  - Peak r's bounce *far* from true ρ especially for small N
- Multiple testing needed
  - Focus on ***inference***, controlling false positives
  - Not ***estimation*** of effect magnitude
- See
  - Kriegeskorte et al. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5), 535-540

Brain Mapping *Inference*

(on where any signal is)



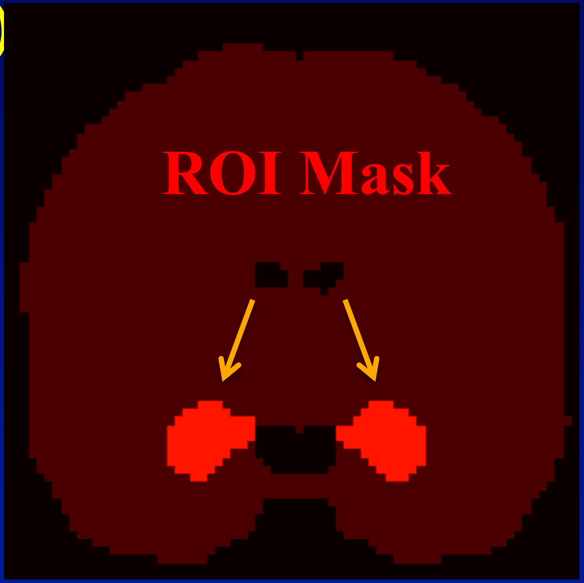


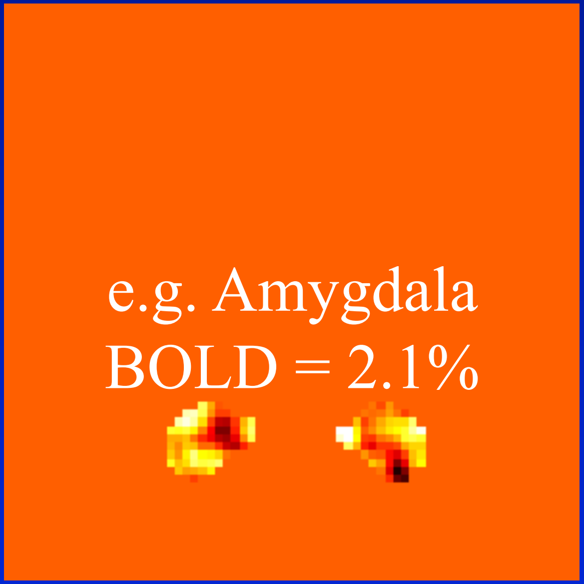
- Perform t-test at 100,000 voxels
- Threshold, mark significant
  - FWE 0.05
    - 95% confident all true positives
  - FDR 0.05
    - 95% true positives on average
- Estimation of effect magnitude?
  - *None*!
  - Only 'estimation' of set of signal voxels

Brain Mapping *Estimation*

(on signal in a given location)

- Define ROI Mask
  - Average voxel-wise %BOLD within mask
- Inference on location?
  - *None!*
  - Location assumed!
  - Only inference is
    - H_0: zero BOLD in ROI
    - H_A: non-zero BOLD in ROI





Estimation Bias from Circularity

- Conditional inference
  - Only measure X in voxels with $Z \geq u$

- Bias – Conditional on a detection
  - $E(X - \mu \mid Z \geq u) = \phi(u^*)/[1\text{-}\Phi(u^*)] \;\; \sigma/\sqrt{N}$

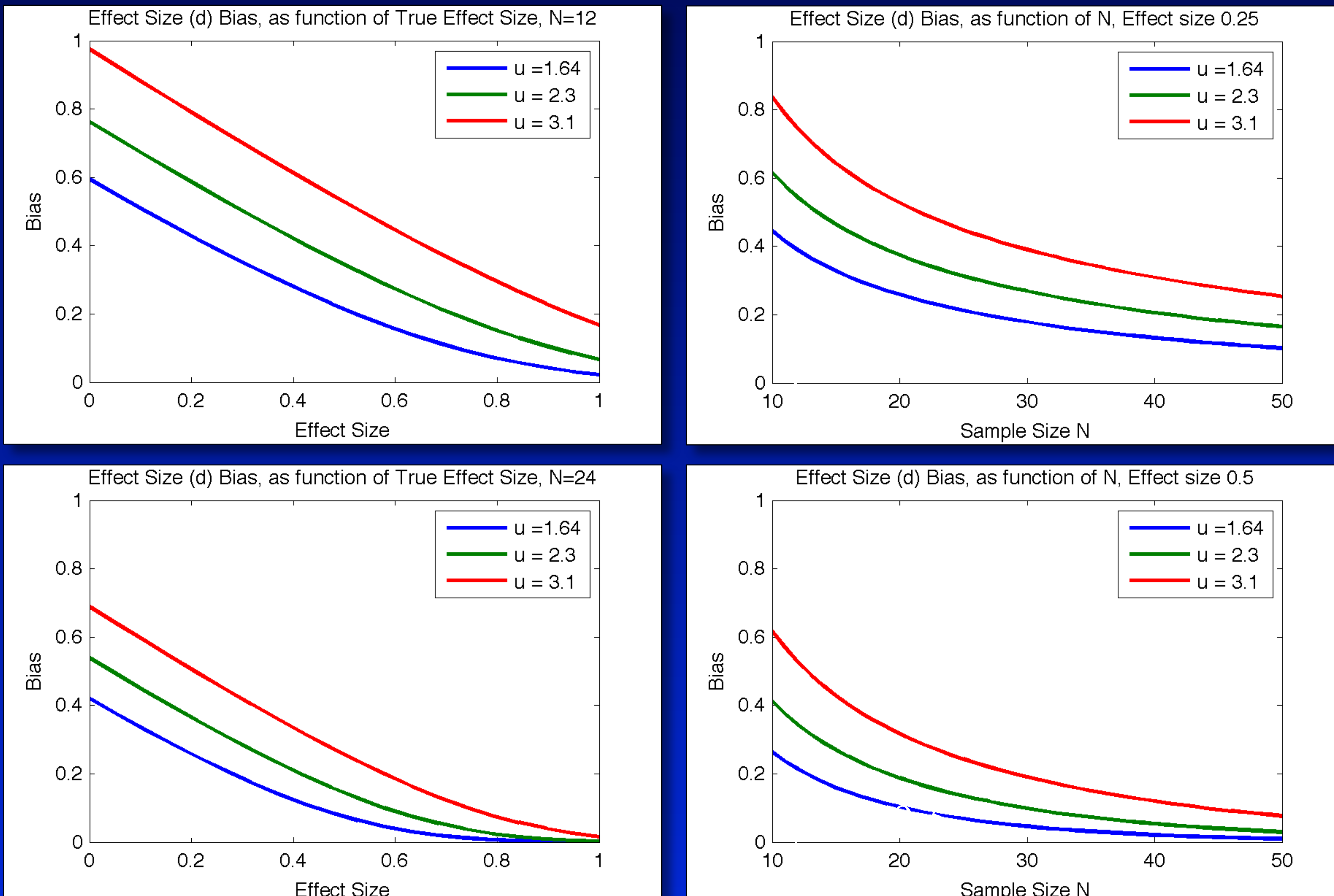
$$u^* = u - \mu \,/\, (\sigma/\sqrt{N})$$

  - Assume no null voxels in ROI

ϕ – CDF of Standard Normal
Φ – PDF of Standard Normal

- Biased by term that depends on
  - Standard Error $\sigma/\sqrt{N}$
  - Shifted threshold u^*
    - Shifted by non-centrality parameter (NCP) $\mu \,/\, (\sigma/\sqrt{N})$

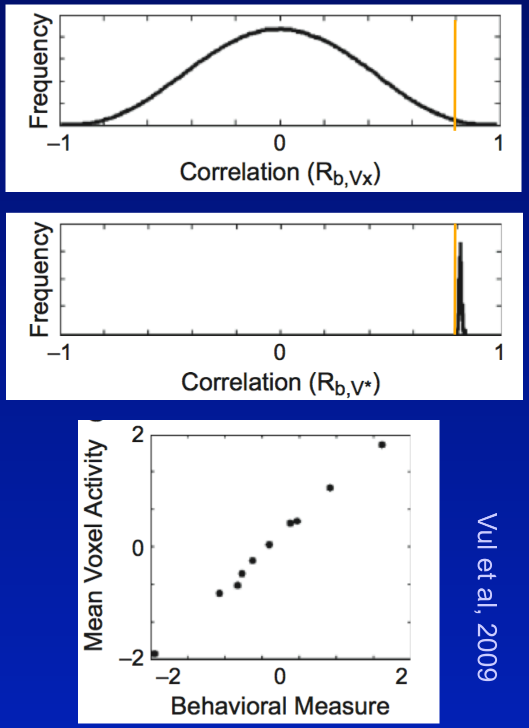
Bias Reduces with Signal & N



Practical Advice (1)

- Emphasize if peak estimate given
  - Especially if % change or *r*
- Consider *not* showing a plot
- Bias-free estimates
  - Compute mean % BOLD for all voxels in a mask
  - Which mask?
    - Define when planning study!
    - Anatomical mask
      - WFUPickAtlas, Harvard-Oxford
    - Functional mask…

N = 10 subjects
1000 *null* correlations



Amazing (null, peak) Result!

Vul et al, 2009

Practical Advice (2) Functional Masks

- From different data
  - Literature or independent subjects
  - From same subjects, unrelated data
- Independent contrast from same acquisition
  - E.g. create mask with (A+B)/2
  - For A−B, compute mean % BOLD within mask
  - Though see Kriegeskorte et al. 2009 supp. mat.
    - Just because contrasts [½ ½] & [-1 1] orthogonal doesn’t mean the COPEs orthogonal
    - Imbalanced design & autocorrelation can prevent perfect independence of COPEs

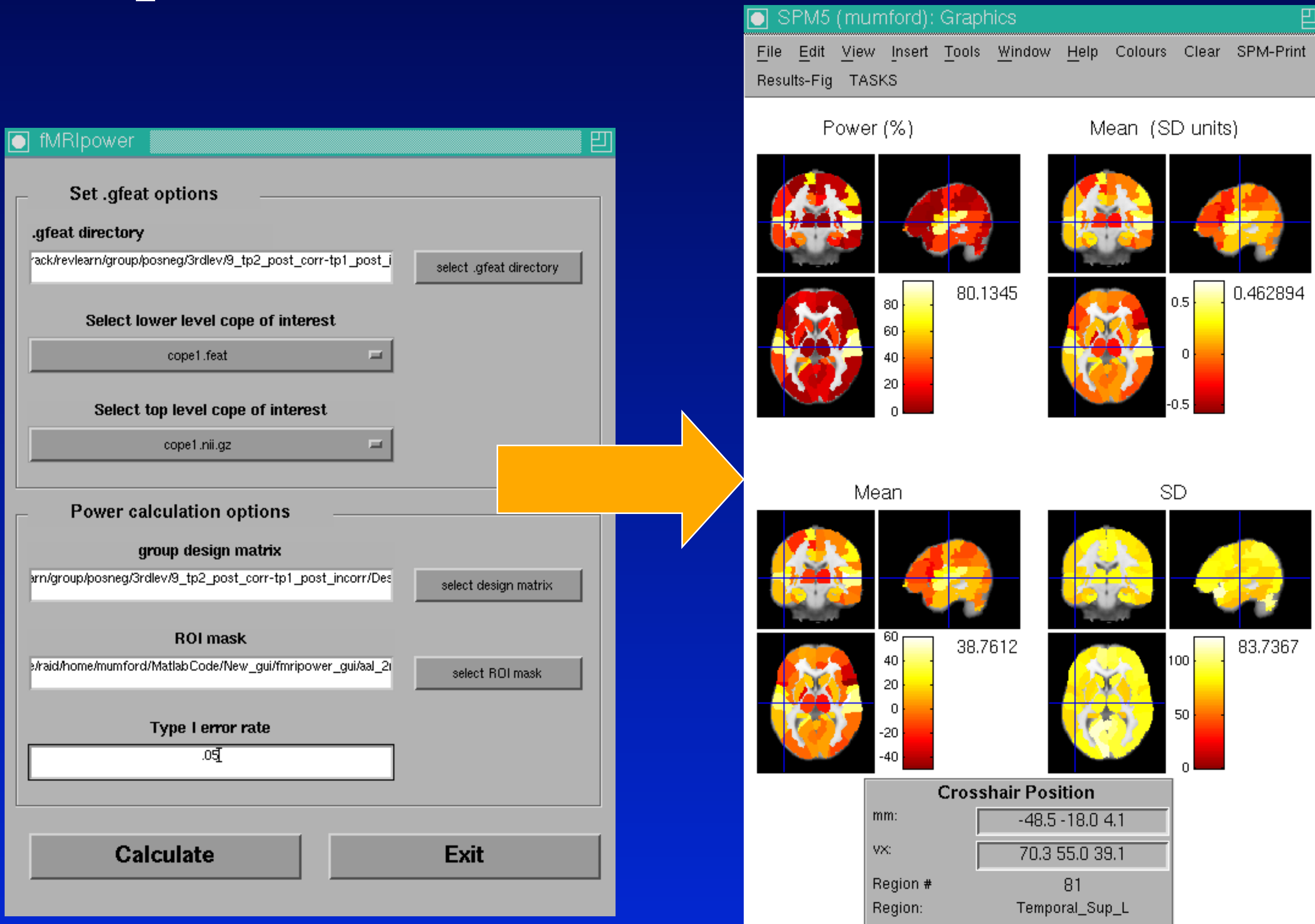
Conclusions: Circularity

- Circularity
  - Selection bias, but severity of bias variable
- Inference
  - Search over space ➔ Localize
- Estimation
  - Assume location ➔ Measure signal
- Make crystal clear what you report

Power Publications

- Desmond & Glover (2002)
  - Desmond & Glover, 2002. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *J. Neurosci. Methods* 118, 115–128.
  - fMRI power method
    - Assumes simple box-car model
    - Ignores temporal autocorrelation
- Mumford & Nichols (2008)
  - Mumford & Nichols, 2008. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, 39(1), 261-268.
  - Allows arbitrary experimental design
  - Accounts for autocorrelation

fmripower


fMRIpower
Set .gfeat options
.gfeat directory
rack/revlearn/group/posneg/3rdlev/9_tp2_post_corr-tp1_post_i
select .gfeat directory
Select lower level cope of interest
cope1.feat
Select top level cope of interest
cope1.nii.gz
Power calculation options
group design matrix
arn/group/posneg/3rdlev/9_tp2_post_corr-tp1_post_incorr/Des
select design matrix
ROI mask
e/raid/home/mumford/MatlabCode/New_gui/fmripower_gui/aal_2
select ROI mask
Type I error rate
.05
Calculate
Exit
SPM5 (mumford): Graphics
File Edit View Insert Tools Window Help Colours Clear SPM-Print
Results-Fig TASKS
Power (%)
Mean (SD units)
80.1345
0.462894
Mean
SD
38.7612
83.7367
Crosshair Position
mm: -48.5 -18.0 4.1
vx: 70.3 55.0 39.1
Region # 81
Region: Temporal_Sup_L

Estimation Bias: What if ROI misses signal?

- Conditional expectation now a mixture
  - $E(X - \mu \mid Z \geq u) =$

$$f_{Ho} \quad \phi(u)/[1-\Phi(u)] \quad \sigma/\sqrt{N} \quad +$$

$$-f_{Ho}\, \mu \quad + (1-f_{Ho})\, \phi(u^*)/[1-\Phi(u^*)] \quad \sigma/\sqrt{N}$$

  where

  f_{Ho} is fraction of ROI that is null
- So now have different directions of bias
  - "Winner's Curse" biases up
  - False positive voxels biases up
  - Diluting true positives biases down

Bias: Effect of Null voxels in ROI