

# Reproducibility and Power

Thomas Nichols

Department of Statistics & WMG  
University of Warwick

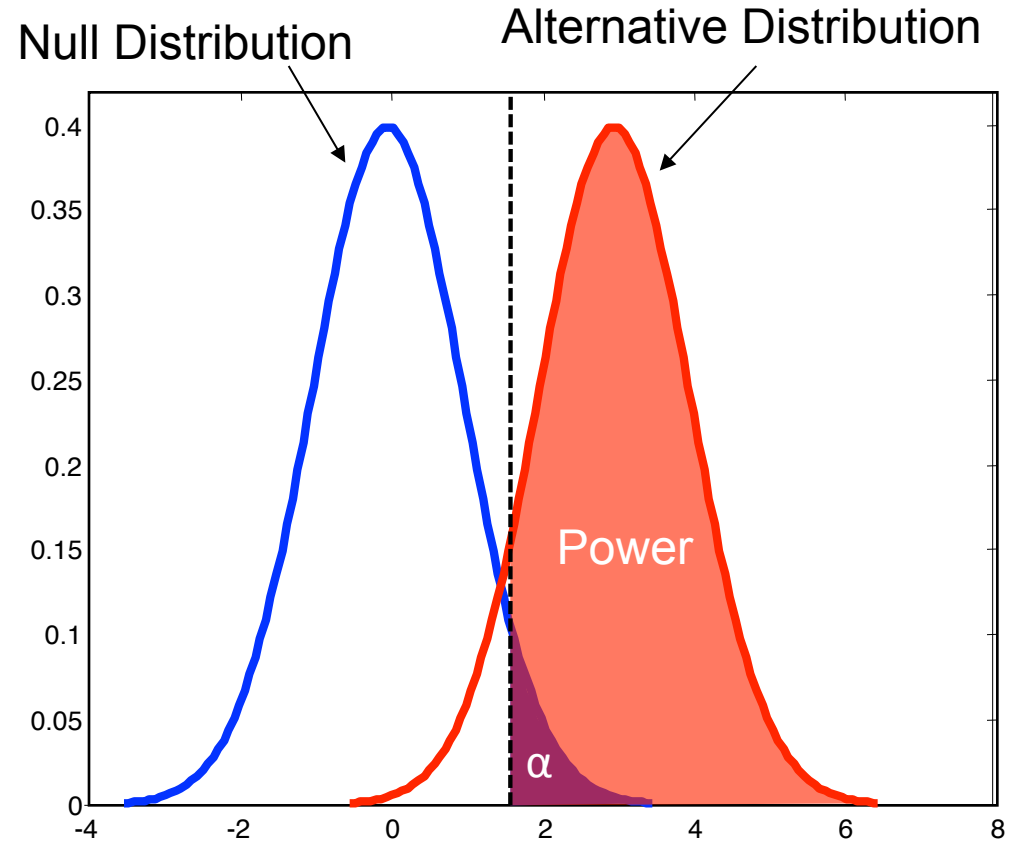
“Reproducible Neuroimaging” Educational Course  
OHBM 2015

# Power & Reproducibility

- Power Review
- Practical Power Methods for Neuroimaging
- Why you should care (Reproducibility)

# Power: 1 Test

- Power:  
Probability of rejecting  $H_0$  when  $H_A$  is true
- Must specify:
  - Sample size  $n$
  - Level  $\alpha$   
(allowed false positive rate)
  - Standard deviation  $\sigma$   
(population variability; not StdErr)
  - Effect magnitude  $\Delta$
- Last two can be replaced with
  - Effect size  $\delta = \Delta/\sigma$



# Power: Statistic vs. Data Units

- 10 subjects
- % BOLD stdev  $\sigma = 0.5$

One-Sample T-test...

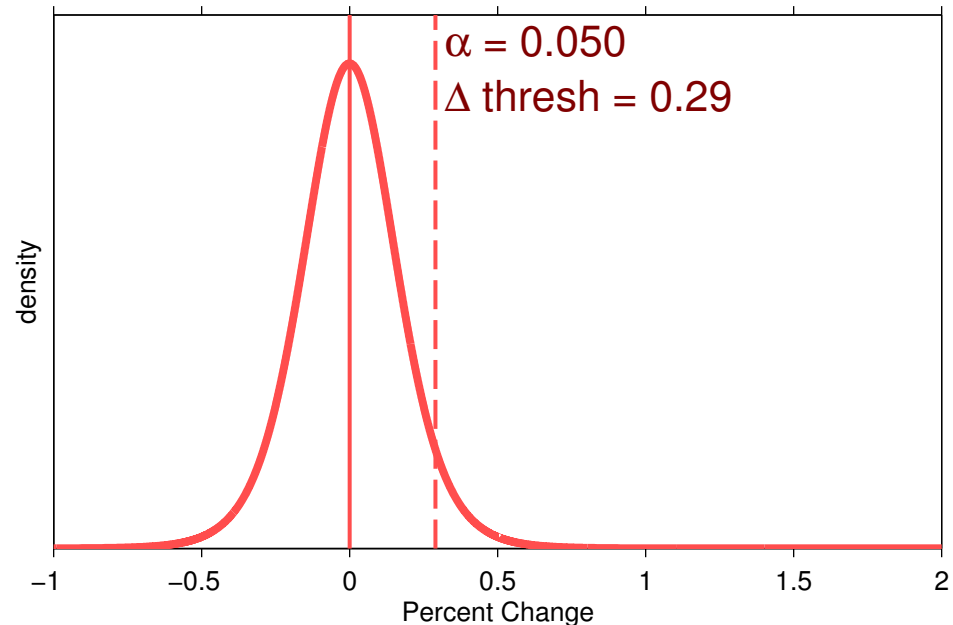
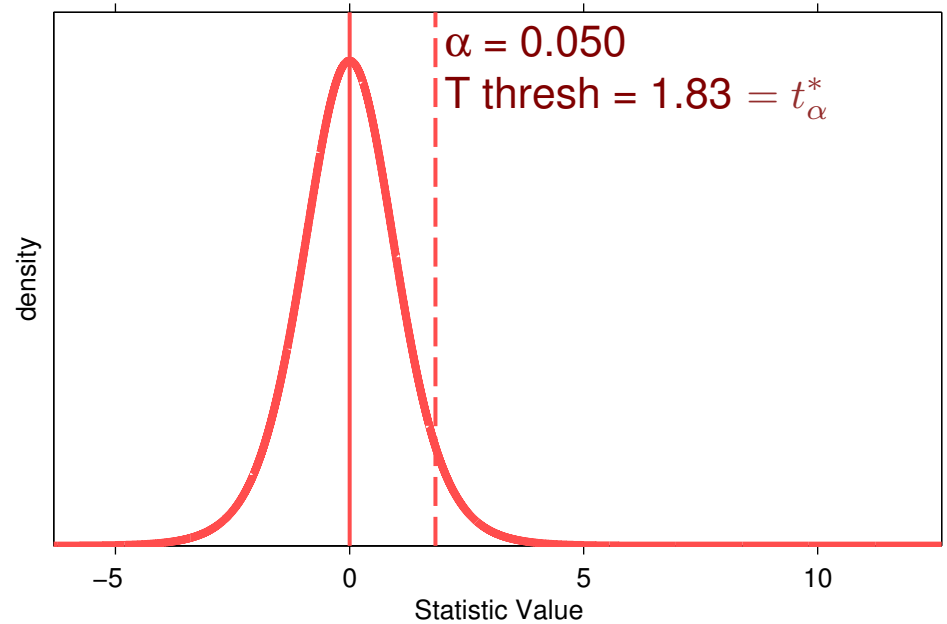
$$t = \frac{\bar{x}}{s/\sqrt{n}}$$

Reject  $H_0$  if...

$$\frac{\bar{x}}{s/\sqrt{n}} \geq t_{\alpha}^*$$

Equivalently, reject  $H_0$  if...

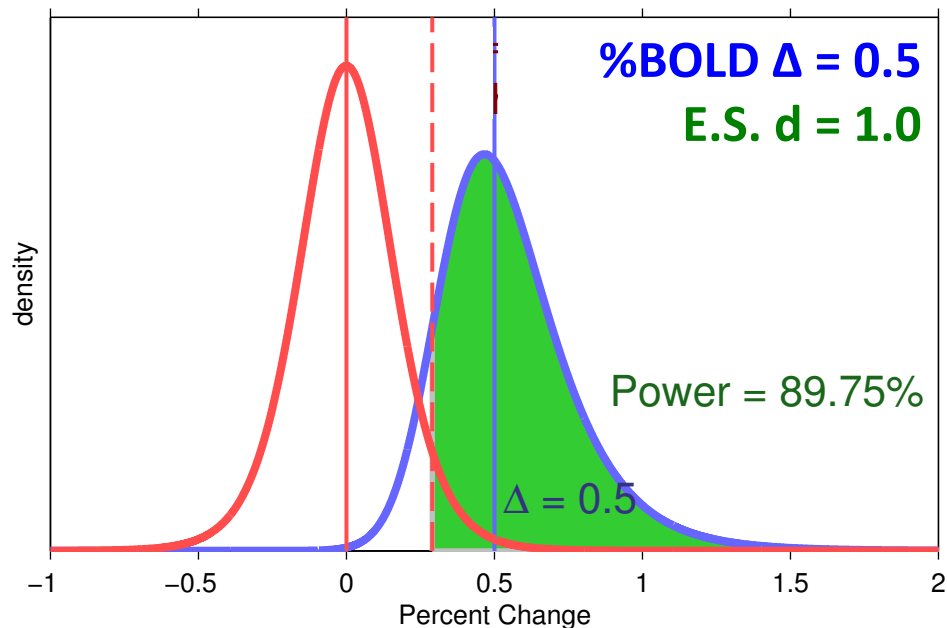
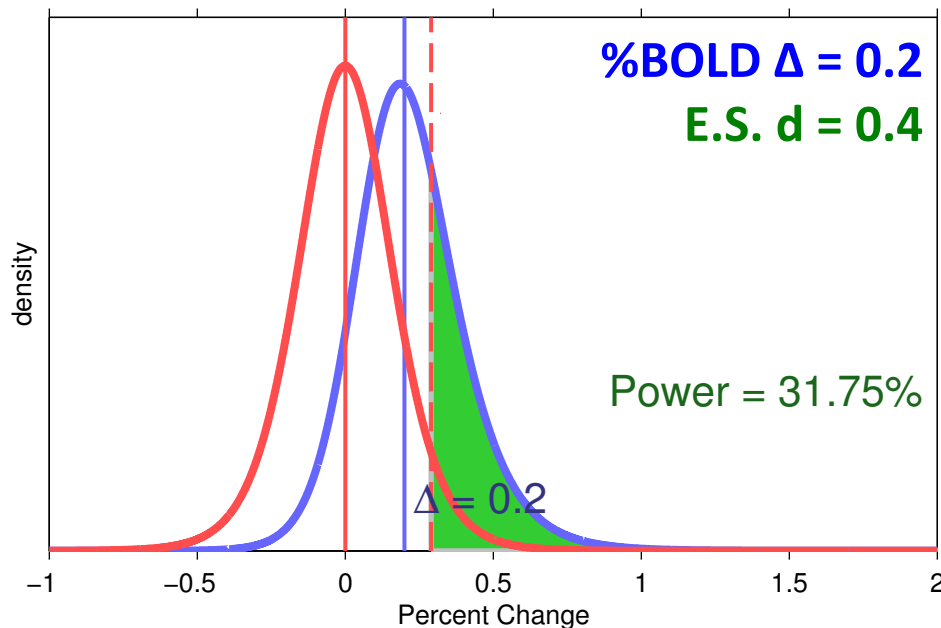
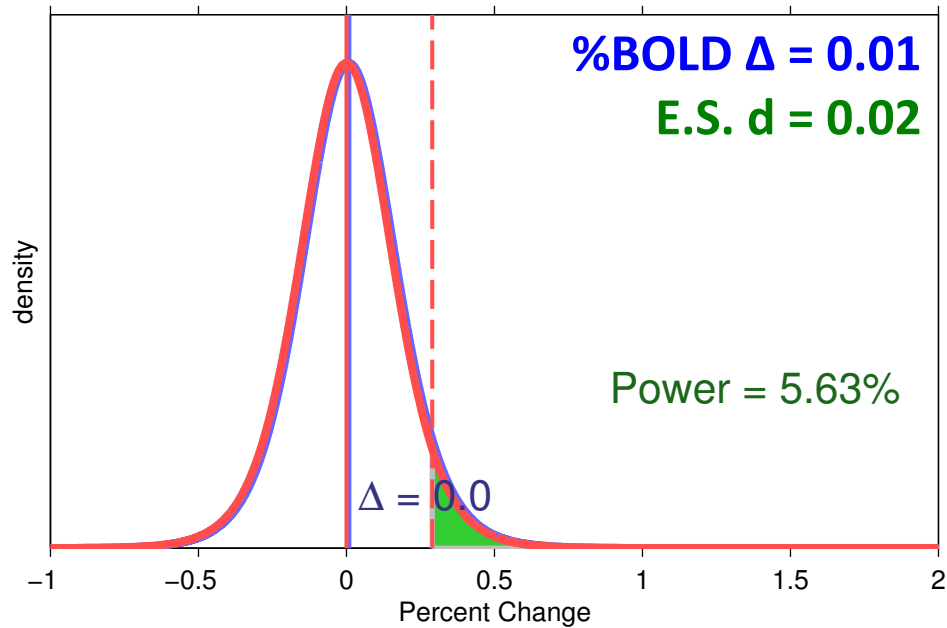
$$\bar{x} \geq t_{\alpha}^* \times s/\sqrt{n}$$



# Power & Effect Magnitude

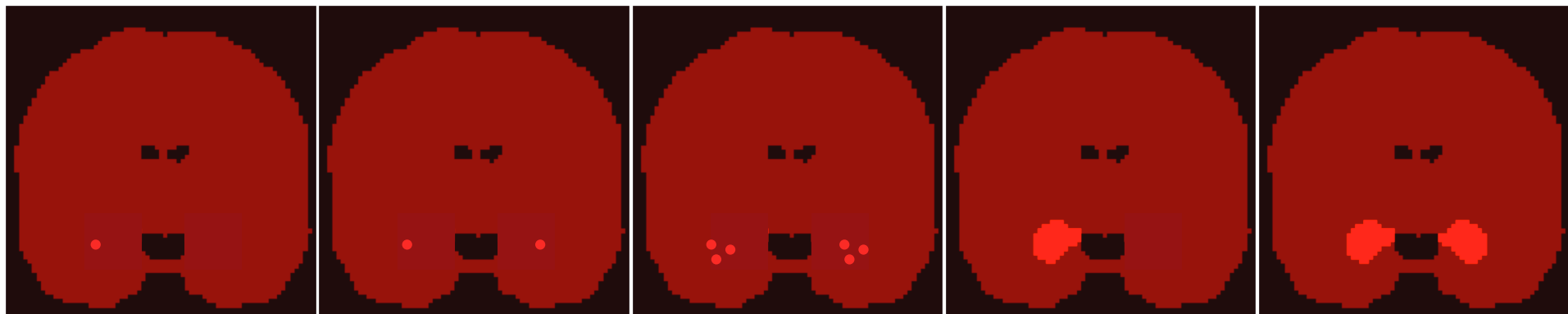
- 10 subjects
- % BOLD stdev  $\sigma = 0.5$
- True %BOLD  
 $\Delta = 0.01, 0.2, 0.5$
- Effect Size (Cohen's  $d$ ) =  $\Delta/\sigma$   
 $d = 0.02, 0.4, 1.0$

*... assuming these are the right numbers!*



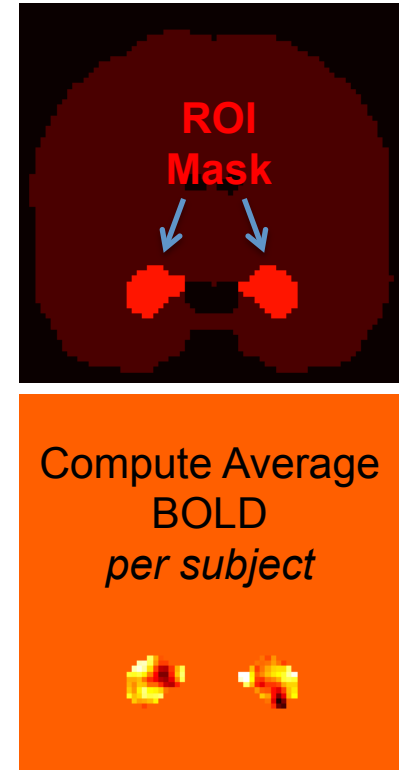
# Power: 100,000 Tests?

- Set  $\alpha$  to reflect multiple testing (easy part)
  - E.g. FWE, for a given search volume & smoothness
    - MNI mask, FWHM [8 8 8]mm, 3301 RESELS  
 $t^* = 4.93$ , then  $\alpha = 0.0000004$
- Alternative:  $\delta_1, \delta_2, \delta_3, \dots, \delta_{99,999}, \delta_{100,000}$  (hard part)
  - Must consider structure of alternative
  - These 10 voxels active at  $\Delta$ , and those other 20...
  - Oh, and don't forget to specify  $\sigma_1, \sigma_2, \sigma_3 \dots$  too!



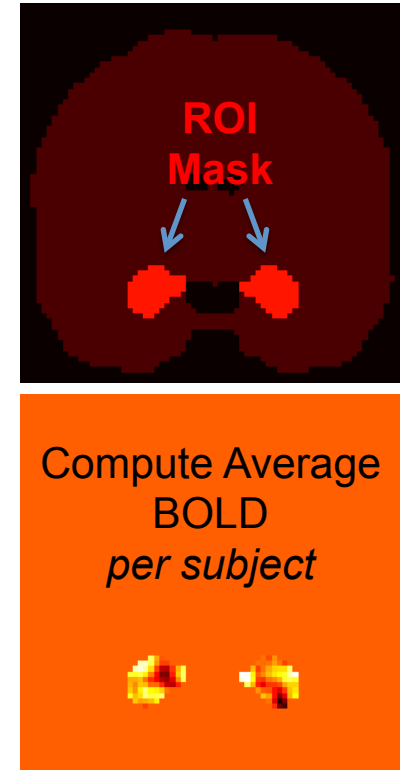
# Practical Power Estimation: Clinical Trial 'Primary Outcome'

- Define 'Primary Outcome'
  - E.g. Average %BOLD in amygdala
- Compute outcome  $\Delta$ ,  $\sigma$ 
  - Previous literature
  - Pilot data
    - E.g. compute %BOLD in amygdala for *each subject*
    - Compute mean, SD
- Set  $\alpha$ 
  - Uncorrected  $\alpha$  if taking clinical trial approach
  - $\alpha$  reflecting multiple testing, otherwise
- Compute power
  - Matlab, or your favorite power app (e.g. G\*power)



# Practical Power Estimation: Clinical Trial 'Primary Outcome'

- Limitations...
- Not flexible
  - $\Delta$  &  $\sigma$  computed only relevant for same design as literature/pilot data
  - Modify design and may not be relevant
    - E.g. shorten run length
    - Re-arrange event spacing
  - Requires pilot image data
- Doesn't account for spatial statistics
  - Yes, can set  $\alpha$  to account for voxel-wise FWE
  - But not cluster-wise





# Practical Power Estimation: Mixed Effects fMRI Modeling

- Power in group fMRI depends on **d** and within- & between-subject variability...

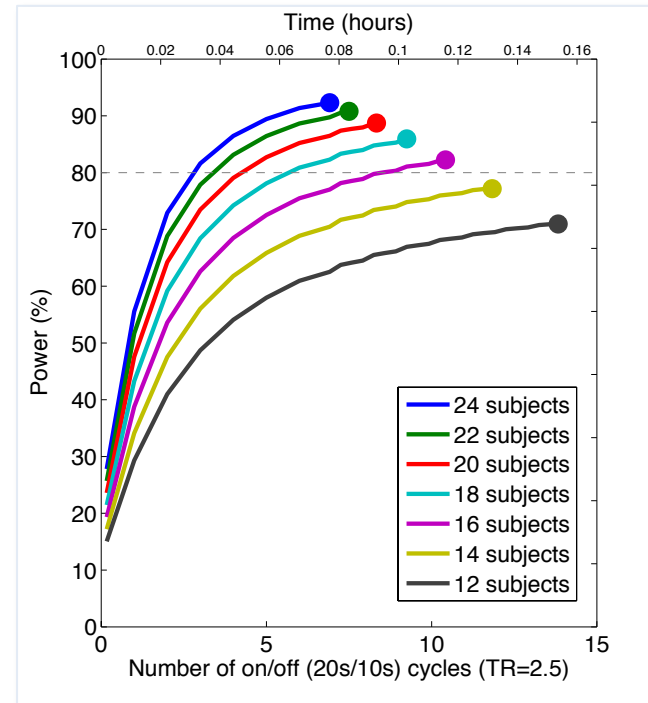
2<sup>nd</sup> Level fMRI Model

$$\hat{\beta}_{cont} = X_g \beta_g + \varepsilon_g$$

$$\text{Cov}(\varepsilon_g) = V_g = \underbrace{\text{diag}\{c(X_k^T V_k^{-1} X_k)^{-1} \sigma_k^2 c^T\}}_{\text{Within subject variability}} + \underbrace{\sigma_B^2 I_N}_{\text{Between subject variability}}$$

# Practical Power Estimation: Mixed Effects fMRI Modeling

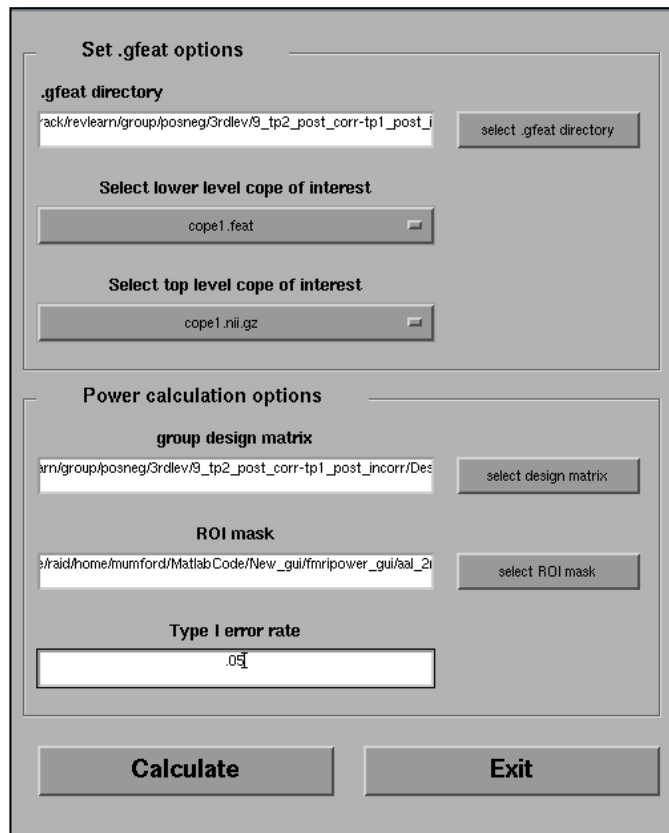
- Requires specifying
  - Intra-subject correlation  $V_k$
  - Intra-subject variance  $\sigma_k^2$
  - Between subject variance  $\sigma_B^2$
  - Not to mention  $X_k, c$  &  $d$
- But, then gives flexibility
  - Can consider different designs
    - E.g. shorter runs more subjects.
    - Optimize event design for optimal power



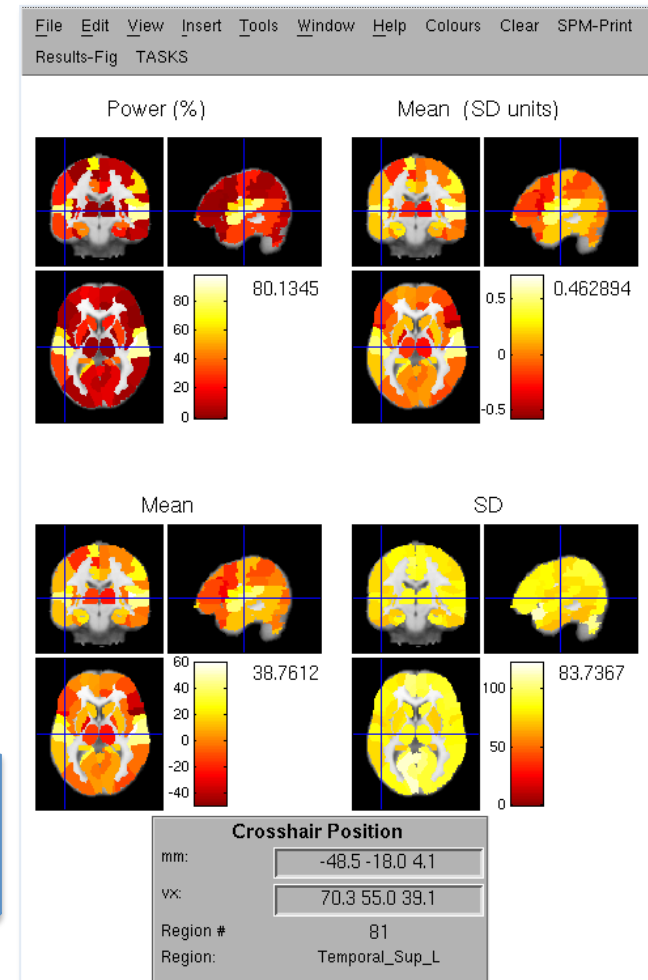
$$\text{Cov}(\epsilon_g) = V_g = \underbrace{\text{diag}\{c(X_k^T V_k^{-1} X_k)^{-1} \sigma_k^2 c^T\}}_{\text{Within subject variability}} + \underbrace{\sigma_B^2 I_N}_{\text{Between subject variability}}$$

# Practical Power Estimation: Mixed Effects fMRI Modeling

- Toolbox to estimate all this from existing data
  - Jeanette Mumford's <http://fmripower.org>

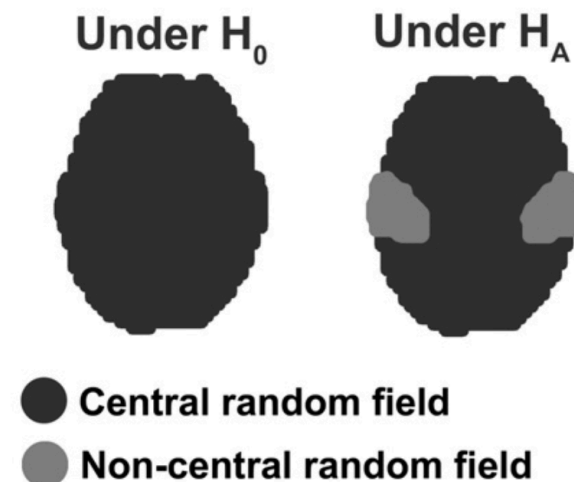


Works with  
FSL & SPM!



# Practical Power Estimation: RFT Cluster-wise Inference

- All previous methods ignored space or worked voxel-wise only
- Cluster-wise inference is popular
  - Gives greater sensitivity
  - Though, pitfalls (Woo, et al. NeuroImage, 2014)
- Power for RFT Cluster Inference
  - Can provide power given a mask of signal
  - Or provide maps of ‘local power’

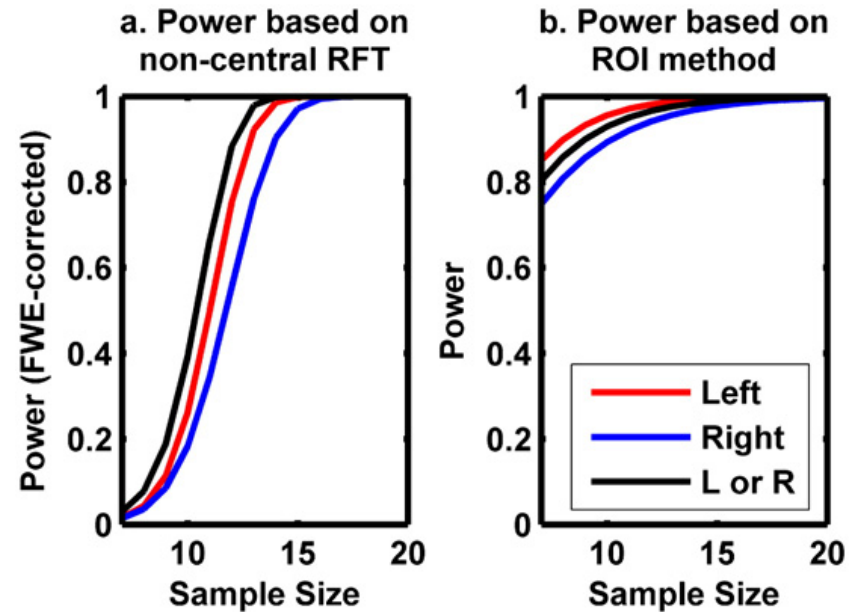


# RFT Cluster-wise Inference

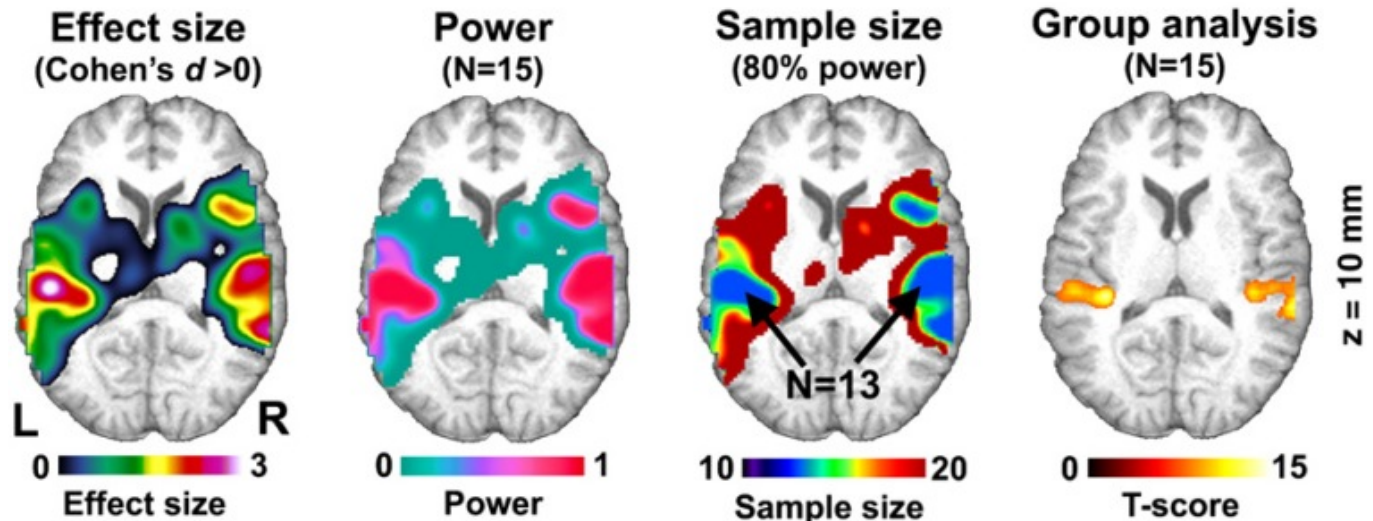
- PowerMap by Satoru Hayasaka

<http://sourceforge.net/projects/powermap>

## FWE vs Primary Outcome Power

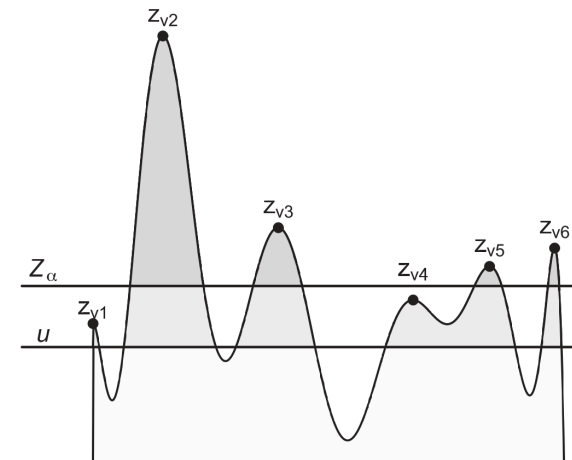


## Mapping Local Power



# Practical Power Estimation: Peak Power

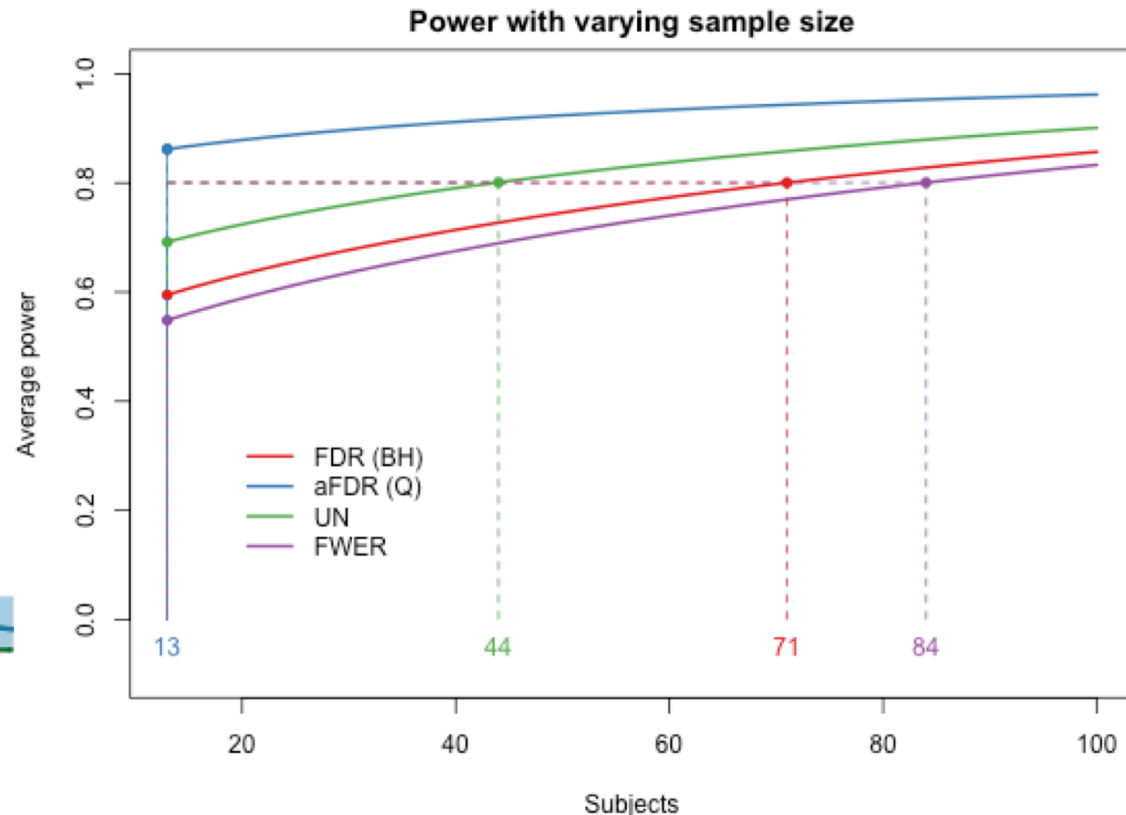
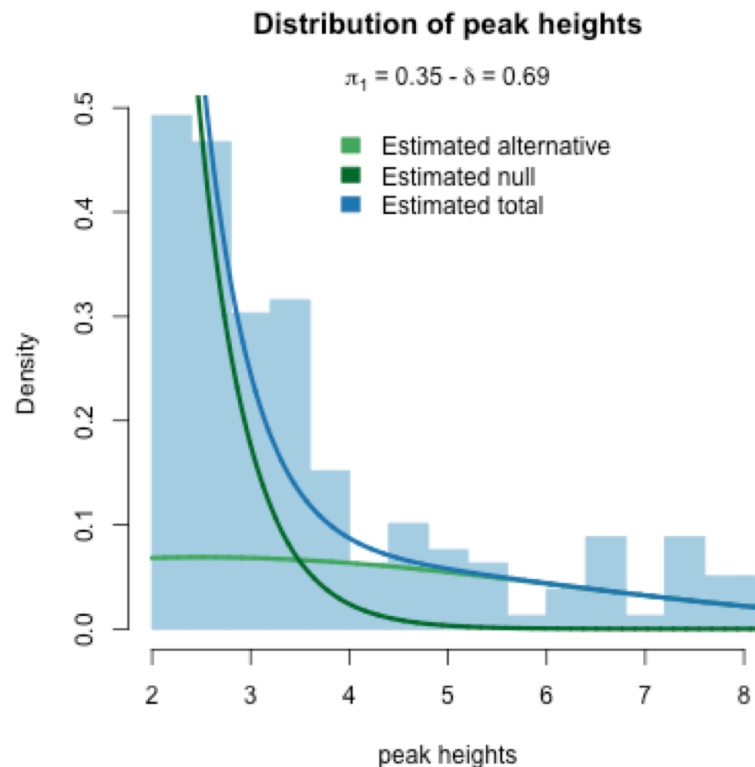
- Mixed-effects & RFT Cluster both complicated
- Peak inference
  - Considers only local maxima above threshold  $u$
  - Make P-values for peak height (SPM8+)
- Peak power calculations simpler to specify
  - Mean peak height
    - Can translate to/from  $d$  if needed
  - Size of activated region
  - Plus, mask size, FWHM smoothness



# Practical Power Estimation: Peak Power

- NeuroPower by Joke Durnez
  - Estimates the 2 key parameters from pilot data's T map
  - Creates power curves

<http://neuropower.shinyapps.io/neuropower>



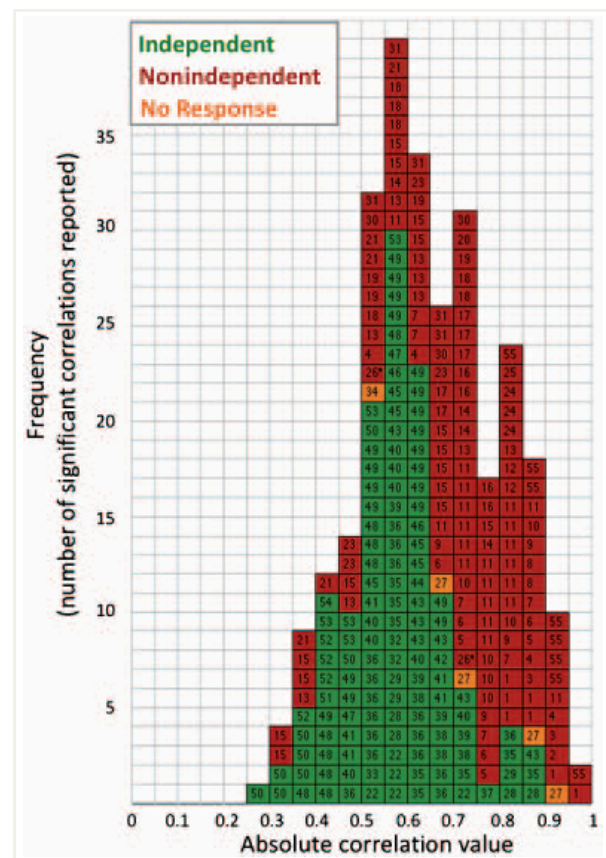
# Power Tools Redux

- Primary 'Clinical Trials' outcome
  - Easy but not flexible
  - Doesn't reflect image-wise analyses actually done
- Mixed effects power
  - Flexible, harder to specify
- RFT Cluster power
  - Only method for widely used cluster inference
- Peak power
  - Easy to specify
- But, again, why do we care?



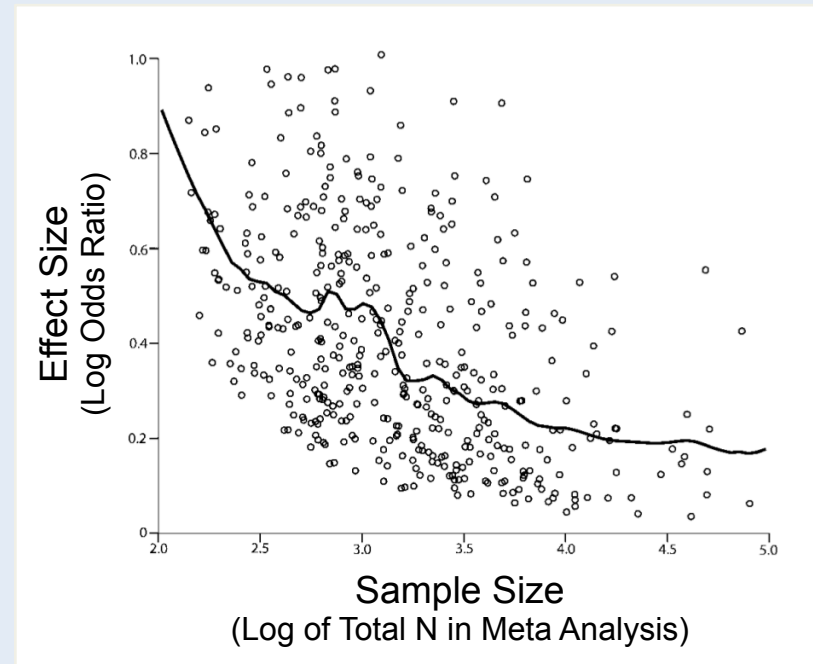
# Power Dangers

- Retrospective Power
  - Power is a probability of a *future* true positive
  - Can't take current data (e.g.  $t=1.3$ ) and say “What was my power for this result?”
- Estimating Effect Sizes
  - Voodoo correlations!
    - Effect size at peak is biased
      - Circularly defined as *the best* effect
  - Must use independent ROIs
    - Independent data, contrasts
    - Anatomical ROI



# Power & Replicability

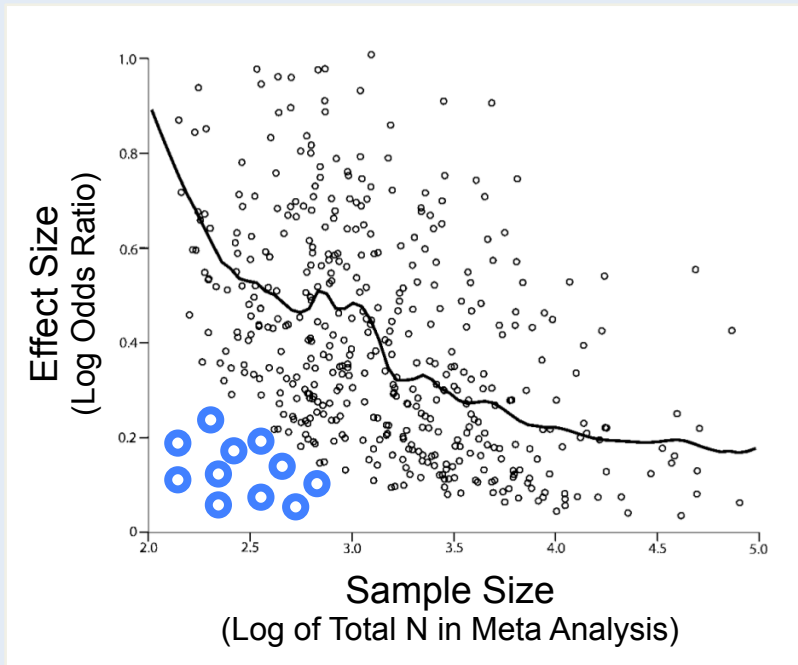
- I got a significant result, who cares about power!?
  - Law of Small Numbers aka “Winner’s Curse”
    - Small studies over-estimate effect size
  - A Low N study...
    - Has low power, likely to fail
    - Significant if...
      - Randomly-high effect, or
      - Randomly-low variance
  - Low power = hard to replicate!
- 256 meta analyses
    - For a binary effect (odds ratio)
    - Drawn from Cochran database
  - Lowest N, biggest effect sizes!



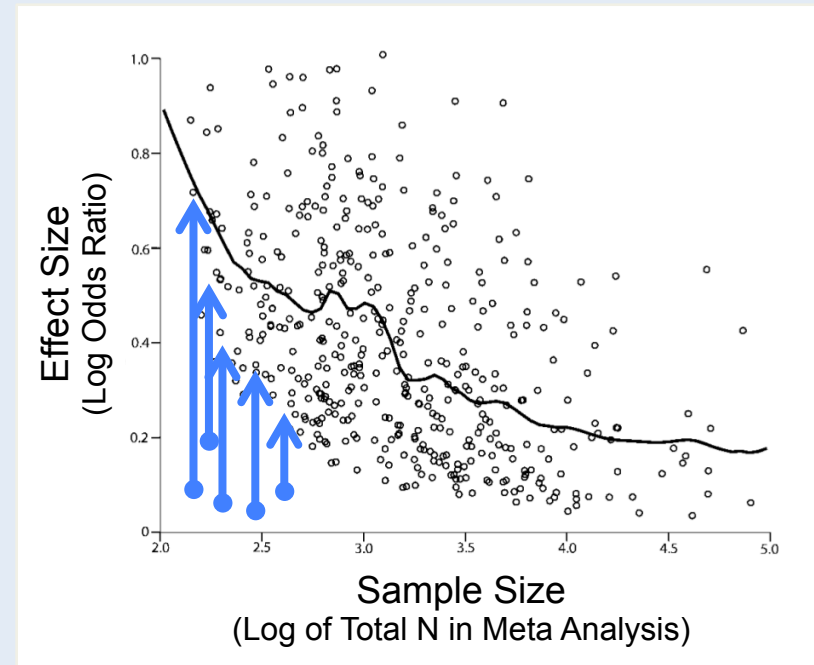
Ioannidis (2008). “Why most discovered true associations are inflated.” *Epidemiology*, 19(5), 640-8.

# Low N studies: The Dark Side

- Suppressed studies & Biased effects
  - $P > 0.05$  not published
  - Biases that afflict small studies more than large studies



File drawer problem  
(Unpublished non-significant studies)

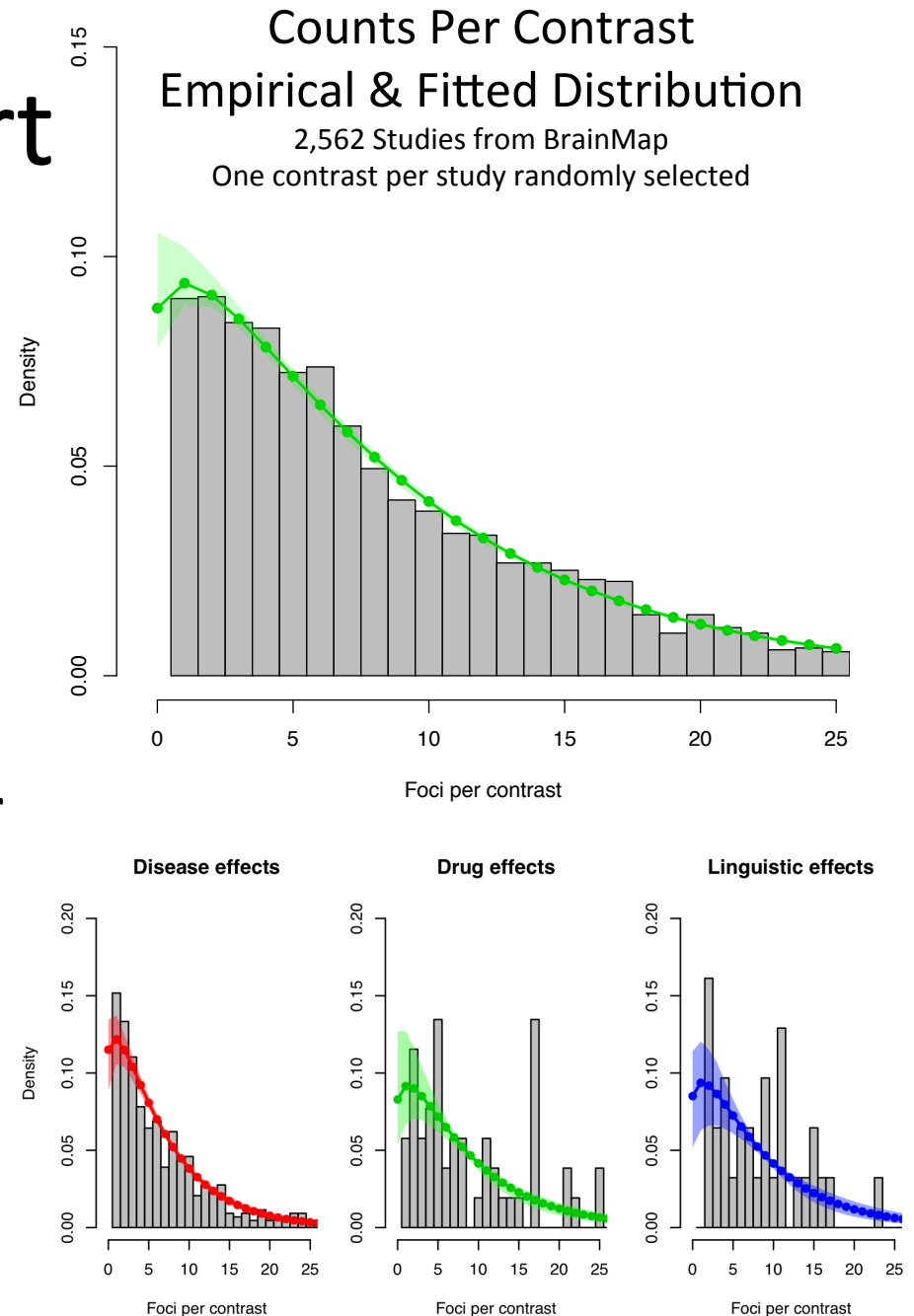


Bias  
(Fishing or Vibration Effects)

# Self-Promotion Alert

- Estimating the size of the “File Drawer” in fMRI
- Use meta-analysis foci counts to infer number of missing (0 count) studies
- About 1 study missing per 10 published
  - 9.02 per 100
  - 95% CI (7.32, 10.72)
  - Varies by subarea

Pantelis Samartidis, et al. Poster 4038-W  
“Estimating the prevalence of ‘file drawer’ studies”



# Vibration Effects

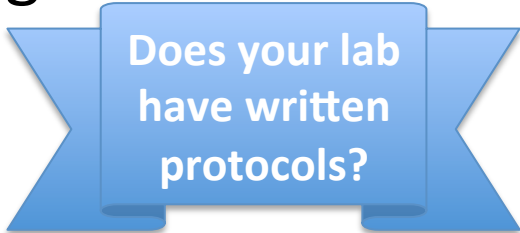
- Sloppy or nonexistent analysis protocols

“Try voxel-wise whole brain, then cluster-wise, then if not getting good results, look for subjects with bad movement, if still nothing, maybe try a global signal regressor; if still nothing do SVC for frontal lobe, if not, then try DLPFC (probably only right side), if still nothing, will look in literature for xyz coordinates near my activation, use spherical SVC... surely that'll work!”

- You stop when you get the result you expect
- These “vibrations” can only lead to inflated false positives

- Afflicts well-intended researchers

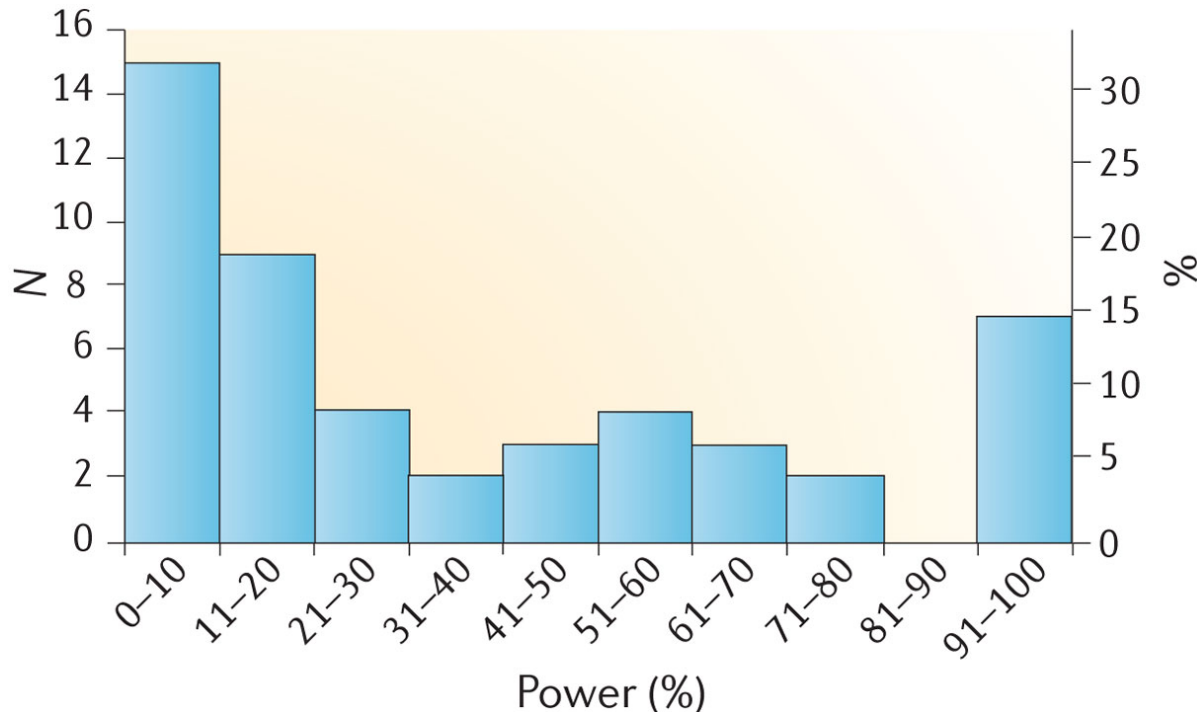
- Multitude of preprocessing/modelling choices
  - Linear vs. non-linear alignment
  - Canonical HRF? Derivatives? FLOBS?



Does your lab  
have written  
protocols?

# Power failure: Button et al.

- Meta-Analysis of (non-imaging) Neuroscience Meta-Analyses
- Recorded median power per meta-analysis
  - Median median power **21%**



50% of all neuroscience studies have **at most a 1-in-5** chance of replicating!

# Button et al's Recommendations

- Do power calculations
- Disclose methods & findings transparently
- Pre-register your study protocol and analysis plan
- Make study materials and data available
  - Check out <http://neurovault.org> !
- Work collaboratively to increase power and replicate findings

# Power Conclusions

- Power = Replicability
  - Best gauge on whether you'll find the effect again
- “Primary outcome” power
  - Good way to appease grant reviewers
  - Doesn't reflect how we usually analyze data
- Whole image-wise power possible
  - Now various tools to choose from