Is Z enough? Impact of Meta-Analysis using only Z/T images in lieu of estimates and standard errors

Camille Maumet[1], TODO pain, and Thomas E. Nichols[1,2]

[1] Warwick Manufacturing Group, The University of Warwick, Coventry, UK.
[2] Statistics Department, The University of Warwick, Coventry, UK.

TODO: find another title

Abstract. The abstract should summarize the contents of the paper using at least 70 and at most 150 words. It will be set in 9-point font size and be inset 1.0 cm from the right and left margins. There will be two blank lines before and after the Abstract. . . .

Keywords: TODO

1 Introduction

2 Methods

2.1 Theory

Given a set of k studies, we denote for each study i: its contrast estimate by Y_i, its contrast variance estimate by V_{Y_i}, its standardized statistical map by Z_i and its sample size by n_i.

Combining contrast estimates and their standard error The gold standard approach to combine contrast estimates and their standard errors is to input them into a GLM, creating effectively the third-level of a hierarchical model (level 1, subject; level 2, study; level 3: meta-analysis). The general formulation is provided in the following equation:

$$Y = X\beta + \epsilon \tag{1}$$

where β is the meta-analytic parameter to be estimated, $Y = [Y_1 \dots Y_k]^t$ is the vector of contrast estimates and $\epsilon \sim \mathcal{N}(0, W)$ is the residual term. Eq. (1) can be solved by weighted least square giving:

$$\hat{\beta} = (X^t W X)^{-1} X^t W Y \tag{2}$$

$$\mathrm{Var}(\hat{\beta}) = (X^t W X)^{-1} \tag{3}$$

In a fixed-effects model (i.e. assuming no between-study variances), we have $W = diag(\sigma_1^2 \dots \sigma_k^2)$ where σ_i^2 denotes the contrast variance for study i. In a random-effects model, we have $W = diag(\sigma_1^2 + \tau^2 \dots \sigma_k^2 + \tau^2)$ where τ^2 denotes the between-studies variance. Approximating σ_i^2 by V_{Y_i} and given $\hat{\tau^2}$ an estimate of τ^2 we obtain the statistics detailed in table 1 for one sample tests.

	Statistic	Disbribution under H_0
GLM FFX	$\frac{1}{\sqrt{\sum_{i=1}^{k} 1/V_{Y_i}}} \sum_{i=1}^{k} \frac{Y_i}{V_{Y_i}}$	$\mathcal{T}_{(\sum_{i=1}^{k} n_i - 1) - 1}$
GLM MFX	$\frac{1}{\sqrt{\sum_{i=1}^{k} 1/(V_{Y_i} + \hat{\tau^2})}} \sum_{i=1}^{k} \frac{Y_i}{V_{Y_i} + \hat{\tau^2}}$	$\mathcal{T}_{k-1}$
GLM RFX	$\frac{1}{\widehat{\sigma}_C^2/\sqrt{k}} \sum_{i=1}^{k} \frac{Y_i}{k}$	$\mathcal{T}_{k-1}$
Fisher's	$-2 \sum_{i=1}^{k} \ln(\Phi(-Z_i)))$	$\chi^2_{(2k)}$
Stouffer's	$\frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$	$\mathcal{N}(0,1)$
Stouffer's MFX	$\frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}\hat{\sigma}}$	$\mathcal{T}_{k-1}$
Optimally weighted-Z	$\frac{\sum_{i=1}^{k} \sqrt{n_i} Z_i}{\sqrt{\sum_{i=1}^{k} n_i}}$	$\mathcal{N}(0,1)$

Table 1. Statistics for one-sample meta-analysis tests and distributions under the null hypothesis.

Combining contrast estimates In the absence of standard error, the contrast estimates Y_i can be combined by assuming that the within-study variance σ_i^2 is roughly constant ($\sigma_i^2 \simeq \sigma^2 \ \forall 1 \leq i \leq k$) or a negligible by comparison to the between-study variance($\sigma_i^2 \ll \tau^2 \ \forall 1 \leq i \leq k$). Then $W = diag(\sigma_C^2 \dots \sigma_C^2)$ where σ_C^2 is the combined within and between-stubject variance such as $\sigma_C^2 \simeq \tau^2$ or $\sigma_C^2 \simeq \tau^2 + \sigma^2$. Eq. (1) can be solved by ordinary least square giving:

$$\hat{\beta} = (X^t X)^{-1} X^t Y \tag{4}$$

$$\mathrm{Var}(\hat{\beta}) = (X^t W X)^{-1} \tag{5}$$

Given $\hat{\sigma_C^2}$ an estimate of σ_C^2 we obtain the statistics detailed in table 1 for one sample tests.

Permutations...

Combining standardised statistics In the presence of standardiseds statistical estimates, Fisher proposed to combine the associated p-values [3]. Stouffer's proposed to combine directly the standardised statistic [4]. In [5] following [2], the author proposed a weighted method that weights each study's Z_i by the square root of its sample size [3,7]. All these statistics, assuming fixed-effects and suited only for one-sample tests only are presented in table 1.

As suggested in [1], to get a kind of MFX with Stouffer's approach, the standardised statistical estimates Z_i can be combined in an OLS analysis. The corresponding estimate, referred as Stouffer's MFX is also provided in 1

2.2 Experiments

Simulations To verify the validity of each estimator under the null hypothesis we estimated the false positive rate at $p < 0.05$ uncorrected. For each meta-analysis, we simulated a contrast estimate a variance estimates such as:

$$Y_i \sim \mathcal{N}(0, \frac{\sigma_i^2}{n_i} + \tau^2) \tag{6}$$

$$V_{Y_i} \sim \frac{\sigma_i^2}{n_i - 1}\chi^2_{(n_i-1)} \tag{7}$$

where $\sigma_i^2 \in [1/2, 1, 2, 4]$ is the within-study variance, $\tau^2 \in [0, 1]$ is the between-study variance (fixed-effects if τ^2 is 0, random-effects otherwise). We simulated different number of studies: $k \in [5, 10, 25, 50]$ and for a given meta-analysis, the number of subjects per studies n was selected such as we would have varying number of subjects in a common range for neuroimaging studies. In each simulated meta-analysis we simulated one study with exactly $20, 25, 10$ and 50 subjects. For the remaining studies the number of subjects were drawn from uniform distributions a quarter from $\mathcal{U}(11, 20)$, a quarter from $\mathcal{U}(26, 50)$ and the remaining from $\mathcal{U}(21, 25)$. A total of 32 parameter sets (4 σ_i^2 x 2 τ^2 x 4 k) was therefore tested, 71 repeats with 5041 samples per repeats were simulated.

Real data We first compared the Z-scores obtained by the three approaches using a Bland-Altman plot. Then, as results are usually presented as a thresholded map, we computed the dice similarity score between thresholded maps obtained with Stouffer's and weighted-Z FFX with FLAME FFX for three (uncorrected) thresholds: p ¡ 0.001, 0.01 and 0.05. Finally, as results are best reported using a multiple comparison correction, we defined ground truth activations as the FLAME FFX analysis FDR-corrected at a threshold of p¡0.05 and plotted Receiver-Operating-Characteristics (ROC) curves of Stouffer's and weighted-Z FFX.

3 Results

3.1 Simulations

Fig. 1 displays the false positive rate obtained for the eight estimators over all set of parameters in the absence and presence of random-effects. From this graph, it is clear that the fixed-effects meta-analytic summary statistics, i.e. Fisher's, Stouffer's and weighted-z estimates are overly liberal in the presence of random-effects. As expected the original Fisher's approach is the most invalid. Surprisingly, FFX GLM is also invalid under fixed-effects, maybe suggesting inaccurate degrees of freedoms (here set to $(\sum_{i=1}^{k} n_i - 1) - 1)$). Stouffer's MFX, GLM RFX and permutations on effects or z-statistics provide valid estimates. The permutation estimates present the largest sampling variance.

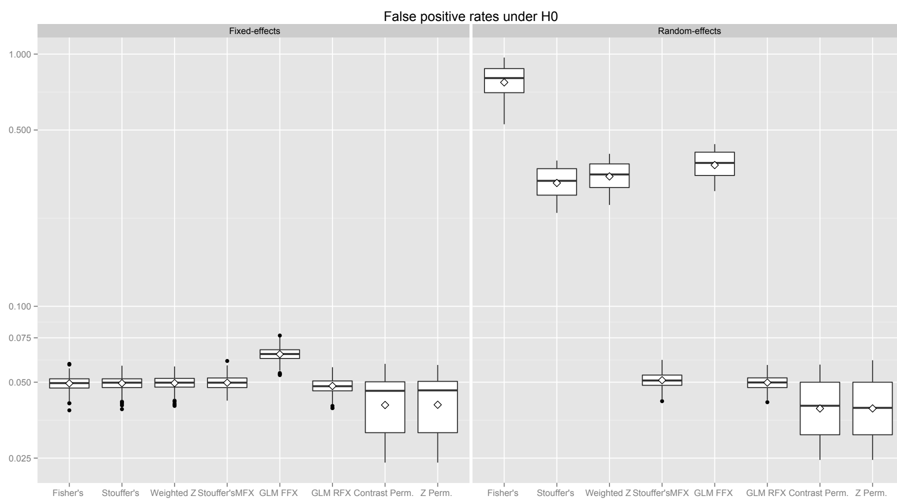

Fig. 1. False positive rates of the meta-analytic estimators under the null hypothesis for $p < 0.05$.

The impact of the number of studies involved in the meta-analysis and of the size of the within-study variance are investigated in fig. 2. The permutation estimates appears conservative (FPR $\simeq 0.03$) when 5 studies are involved. All approaches perform equally as soon as 10 or more studies are included in the meta-analysis.

3.2 Real data

Fig. 3 presents the error made on the z-score estimated by each valid meta-analytic approach by comparison to the z-scores estimated with the Ground truth

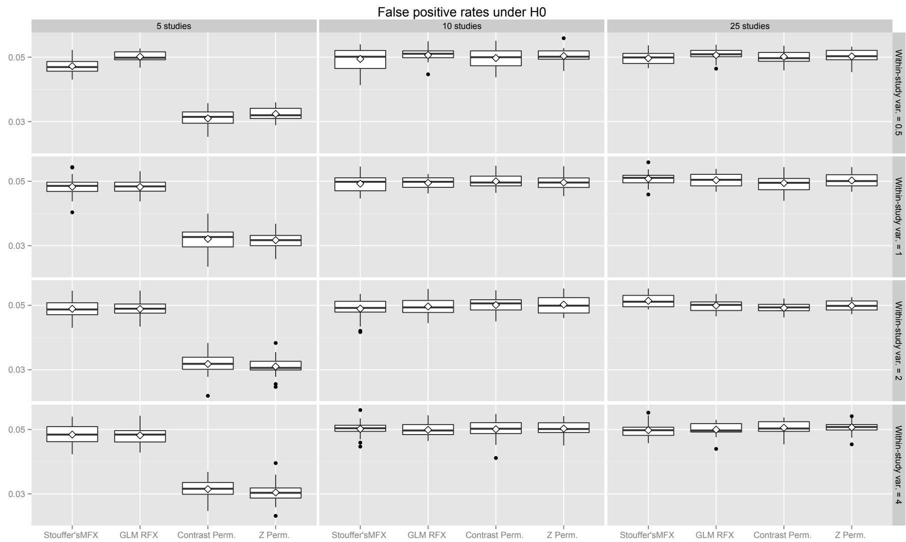

Fig. 2. False positive rates of the valid random-effects meta-analytic estimators under the null hypothesis for $p < 0.05$ as a function of the number of studies and the within-study variance.

MFX GLM approach. GLM RFX, Contrast permutations and z permutations offers valid or more stringent estimates while Stouffer's MFX is more liberal than the gold standard for p-values between 10^{-3} and 10^{-7} and more stringent for p-values smaller than 10^{-7}.

Dice among valids

1. StouffersMFX: 0.9454
2. PermutZ: 0.9450
3. GLMRFX: 0.8994
4. PermutCon: 0.8991

1. WeightedZ: 0.9244
2. Stouffers: 0.9184
3. GLMFFX: 0.8972
4. fishers: 0.8382

AUC between 0 and 0.1 among valids

1. StouffersMFX: 0.8924
2. PermutZ: 0.8919
3. GLMRFX: 0.7809
4. PermutCon: 0.7815

1. WeightedZ: 0.8293
2. Stouffers: 0.8619
3. fishers: 0.6329
4. GLMFFX: 0.6111

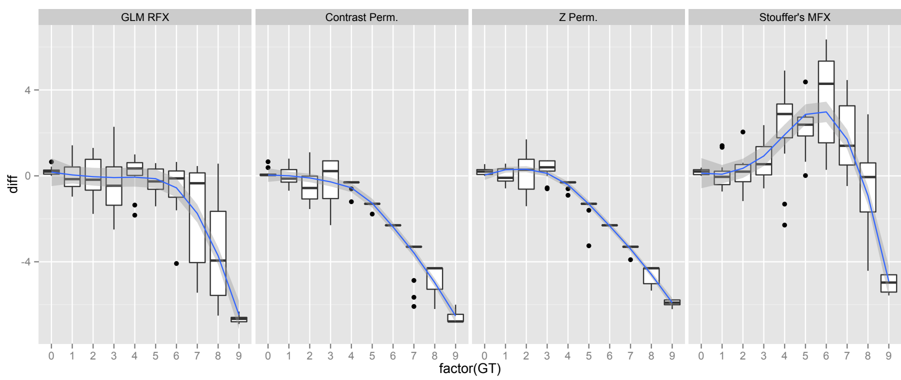

Fig. 3. Difference between the -log10 p-values from each meta-analytic estimates and the -log10 p-values from MFX GLM as a function of -log10 p-values from MFX GLM [TODO change to z-stats instead of -log10(p)]

4 Conclusion

We have found appreciable differences between the Z-score only approaches as compared to a gold-standard approach. Overall the weighted-Z method provided results that were closer to the ground truth than Stouffer's approach. We hypothesize that Stouffer's methods may be attributing greater weights to less-representative subsets of the data. All three procedures are valid, but the gold-standard should be giving the most faithful representation of the population effect. This advocates over the development of tools supporting the sharing E+SE's.

5 Acknowledgements

We gratefully acknowledge the use of this data from the Tracey pain group, FMRIB, Oxford.

References

1. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, 45(3):810–23, 2009.
2. Peter Cummings. On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, 3:171–197, 1958.
3. R.A. Fisher. *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, 1932.
4. S. Stouffer, L. DeVinney, and E. Suchmen. *The American Soldier: Adjustment During Army Life*, volume 1. Princeton University Press, Princeton, NJ, 1949.
5. D V Zaykin. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology*, 24(8):1836–41, 2011.