



The Problem of Inflated Type I Errors with Simple Group Models

Jeanette Mumford & Thomas E. Nichols
Department of Biostatistics, University of Michigan

Abstract

Group modeling of fMRI data is done according to one of two broad approaches: Ordinary Least Squares (OLS) analysis of contrast images [1], or a mixed models (MM) approach [2,3], where optimal weighting is found based on between- and within-subject variance estimates. OLS users favor OLS's computational simplicity, which consists of 1- or 2- sample t-tests on contrast data, while MM users cite its ideally optimal sensitivity. While literature has focused on sensitivity, little discussion has been made of *specificity*, or false positive rates, of the methods. In this work we focus on specificity of the commonly used OLS method, the 1-sample t-test on contrast images. Using a real dataset with heterogeneous variance across subjects, we compared the specificity of 3 methods: a t-test using $n - 1$ degrees of freedom (DF), a t-test using Satterthwaite effective degrees of freedom (eDF), and a permutation test. The Satterthwaite eDF were found to vary widely when variances were heterogeneous, but the Satterthwaite Chi-square approximation was quite poor. The test with $n - 1$ degrees of freedom were found to be slightly *conservative* in our dataset. The permutation test was valid and the p-values were found to be similar to truth and slightly smaller than p-values computed with $n - 1$ DF.

Introduction

When given a set of first-level contrasts for a voxel across n subjects, (x_1, \dots, x_n) , the most common group model in functional neuroimaging is a one-sample t-test given by

$$T = \bar{x} / (\sqrt{S^2/n}),$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ and

$$S^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1).$$

Inference is carried out by comparing T to a t -distribution with $n - 1$ degrees of freedom (t_{n-1}). This procedure assumes contrast images have homogeneous variance over subjects, when in reality contrast variance can vary between subjects; most severely when the number of fMRI sessions differs between subjects. Although the sample variance, S^2 , is an unbiased estimate for the average of the variances, the one-sample t-test with $n - 1$ degrees of freedom (DF) is only valid and optimally sensitive when the variances are homogeneous across subjects. To improve power, some authors estimate between and within subject variances in order to optimally weight different subjects [2,3,9].

An issue often over-looked is the validity or *specificity* of the one-sample t-test when the variance is heterogeneous over subjects. If the variability of the variances across subjects is large, the effective degrees of freedom (eDF) could be far less than the usual degrees of freedom ($DF = n - 1$).

Our study focuses on the false positive control on a one-sample t when the variance is heterogeneous over subjects. Three methods of calculating p-values were studied:

P_{n-1} one-sample t-test with $n - 1$ DF

P_{eDF} the one-sample t-test using eDF calculated with the Satterthwaite approximation [8] and

P_{perm} a permutation test.

Heterogeneous variance causes eDF to decrease, which causes heavier tails and making large statistic value more common under the null. So we hypothesized that when variances are heterogeneous, the P_{n-1} would be artificially small. Since a one-sample nonparametric permutation test [5,6 randomize in 7] makes no assumption of homogeneous variances, we carried out a nonparametric test to see if any bias in p-values could be corrected. A Monte-Carlo (MC) simulation was used to calculate correct p-values (P_{MC}) to compare to the p-values across the different inference methods.

Methods

Data

- Finger tapping experiment of the right hand [4]
 - 12 normal control subjects
 - Block design
 - Rest plus 3 tasks:
 - pseudorandomly cued to tapping of the index finger, sequentially tap the fingers, or randomly tap the fingers
 - Modeling
 - fMRIB software library (FSL) [7]
 - Two-level “FLAME” modeling
 - “COPE” - 12 contrasts, comparing sequential finger tapping to random finger tapping
- Subject-specific mixed-effects variance estimates of contrasts ($\sigma_1^2, \dots, \sigma_{12}^2$) for each of 226,000 voxels

P-value Computation

- P_{11} : Compare T to t-distribution with $n - 1 = 11$ DF
- P_{eDF} : Compare T to t-distribution with Satterthwaite approximation to eDF

$$eDF = 2E^2(S^2)/Var(S^2) = \sum_i \sigma_i^2 / [\sum_i \sigma_i^4 + (n-1)^{-2} \sum_{i \neq j} \sigma_i^2 \sigma_j^2]$$

- P_{perm} : Compare T to empirical t-distribution based on permutations
 - Create all possible permutations created by changing the signs of the contrasts ($2^{12} = 4096$) and compute a test statistic for each permutation
 - $P_{perm} = \%$ of 4096 permuted test statistics as or larger than T

Simulation Details

- For each realization, generate 10,000 sets of $N(0, \sigma_i^2)$ data for each of the 12 subjects and use to calculate 10,000 test statistics
- $P_{MC} = \%$ of 10,000 test statistics as or larger than T

Results

eDF under Variance Heterogeneity

Satterthwaite eDF were found to be far from 11. Figure 1 shows that P_{11} were too small when compared to P_{eDF} . If P_{eDF} were exact, this would indicate considerable anticonservativeness (invalidity) of OLS P-values (P_{11}).

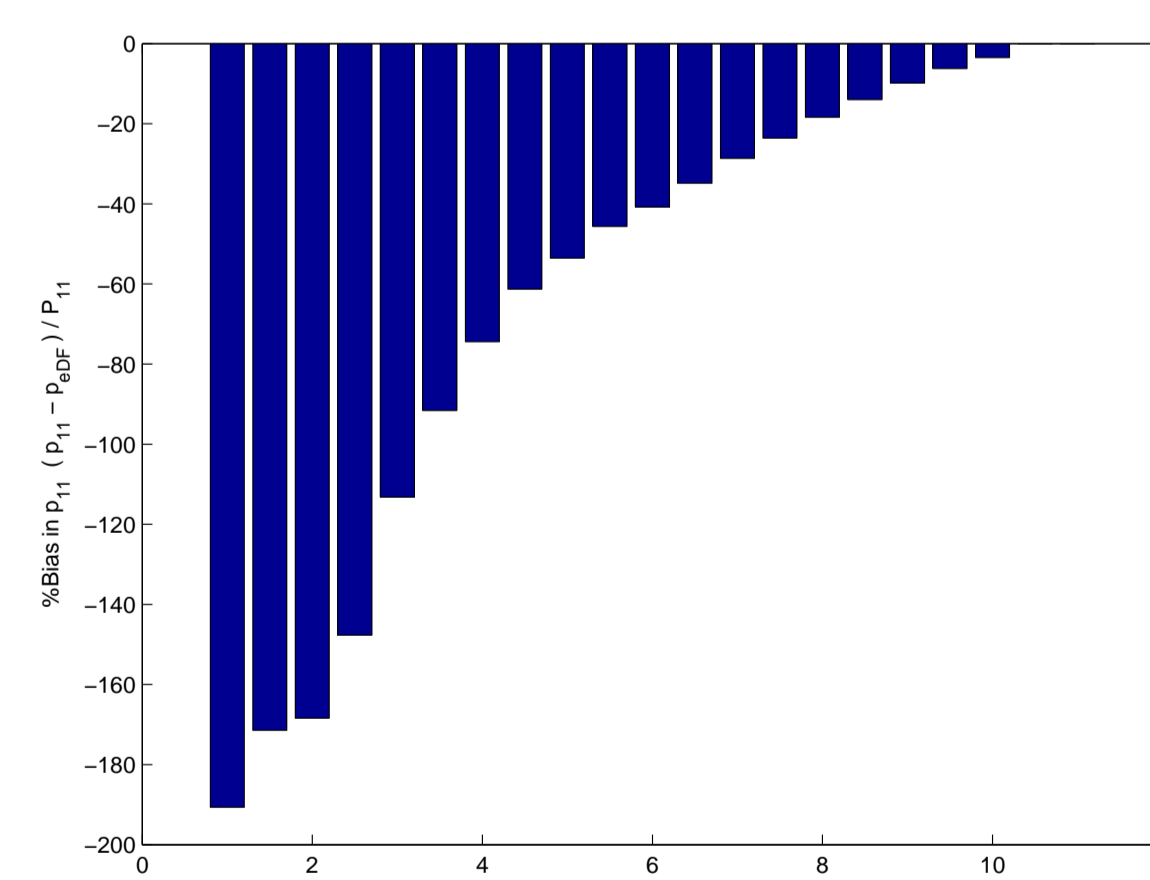


Figure 1: Median bias in P_{11} , compared to P_{eDF} by eDF, for $P_{11} < 0.05$. As expected, bias becomes more negative for lower eDF.

When P_{eDF} were compared to P_{MC} , however, we found that P_{eDF} was very conservative ($P_{MC} < P_{eDF}$) and the P_{11} values were only slightly conservative (Figure 2). Therefore, the test with 11 DF is performing *better* than with eDF.

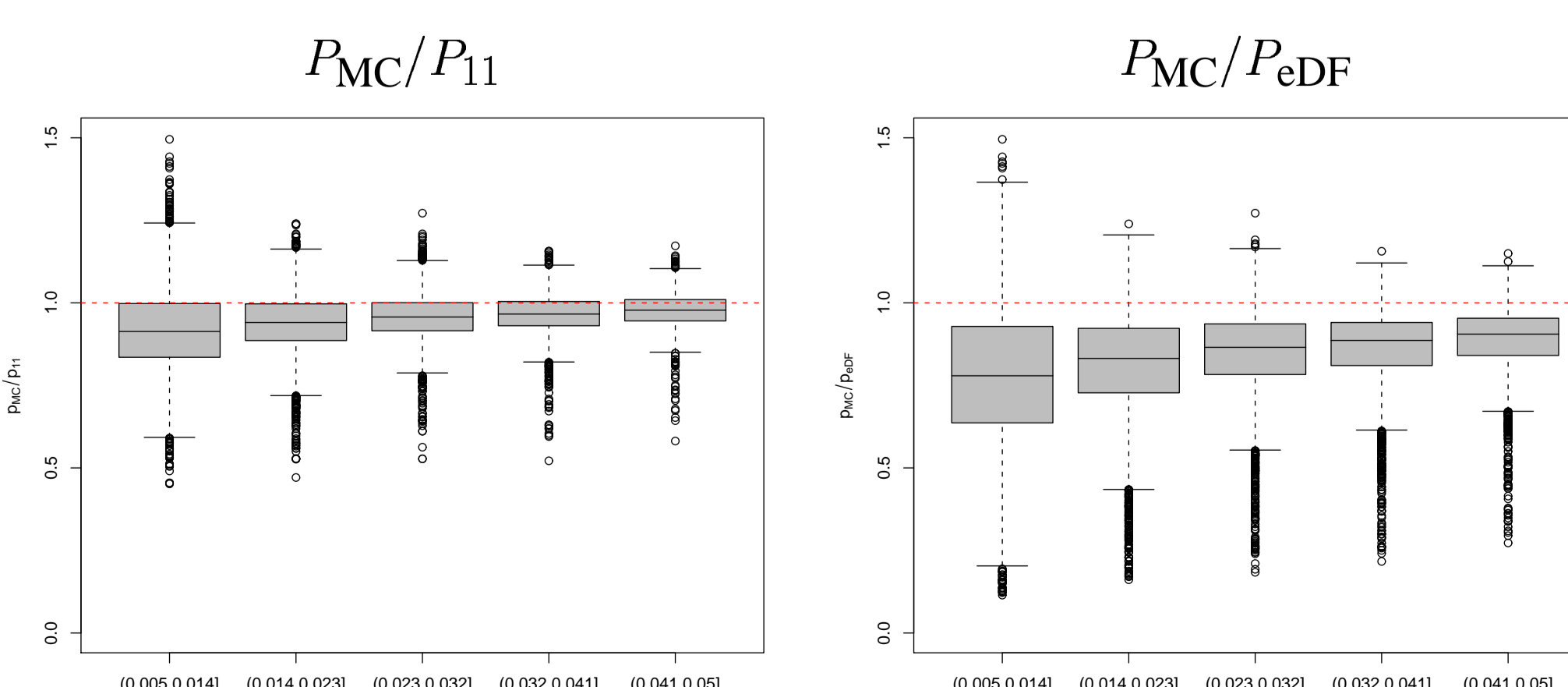


Figure 2: Boxplots of the ratio of P_{MC} to P_{11} and P_{eDF} illustrating that P_{11} is actually slightly conservative (left) and P_{eDF} is extremely conservative (right)

eDF P-value Bias

In order to understand the poor performance of P_{eDF} , we examined the implied distribution of S^2 and T for each method. While the first two moments of the distribution of S^2 based on eDF and the MC distribution match, the tails of the T distribution based on eDF are too wide.

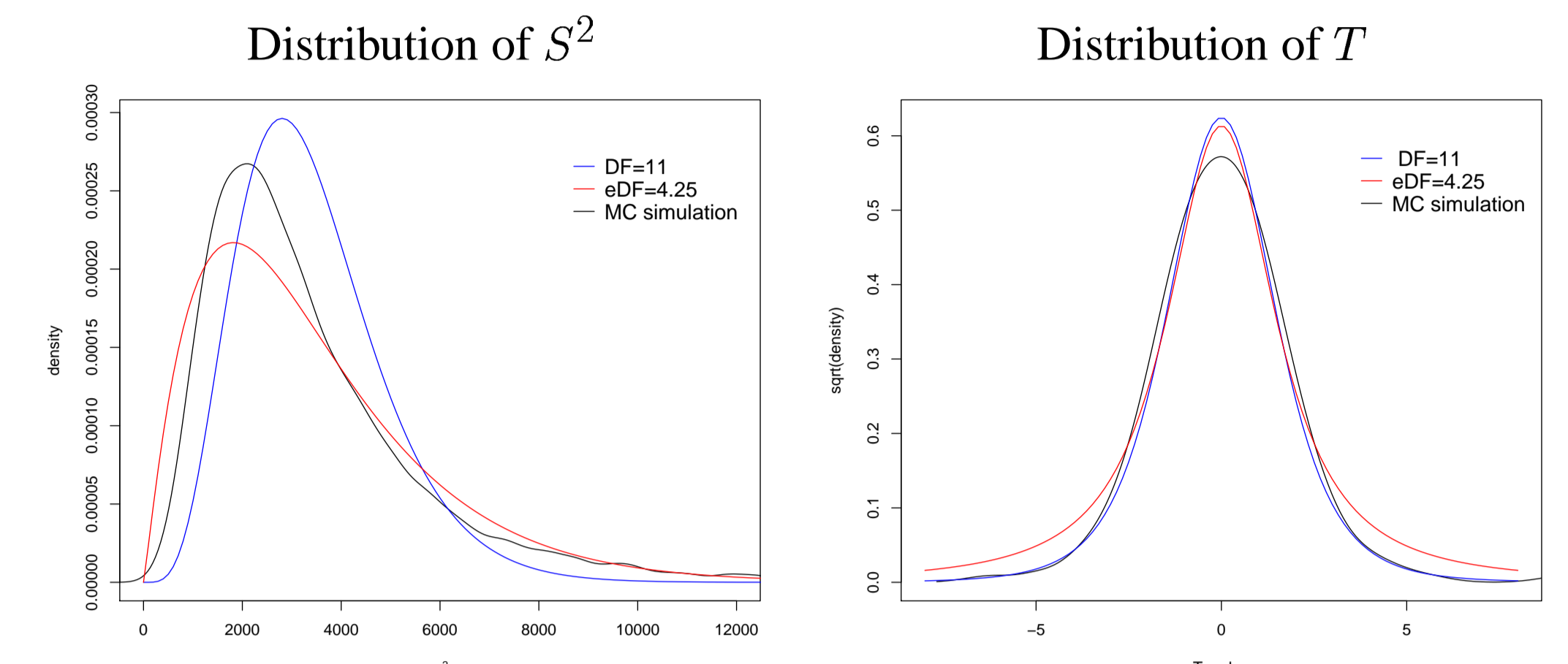


Figure 3: The distribution of S^2 (left) shows that the first two moments of the eDF and MC simulated distributions match, but the tails do not match. The T distributions (right) show that the eDF distributions has much larger tails than that of the MC simulation or $DF=11$, which explains the conservative p-values found using eDF

Permutation P-values

While nonparametric inferences are known to be exact, we confirmed this by comparing P_{perm} to P_{MC} . In Figure 4 left, the ratio P_{MC}/P_{perm} has mean of 1 (median less than one due to skew). To gauge advantage of using a permutation test over OLS, we also compared P_{perm} to P_{11} . In Figure 4 right, the ratio P_{MC}/P_{perm} has mean slightly greater than 1 (though a median of about 1 than one due to skew).

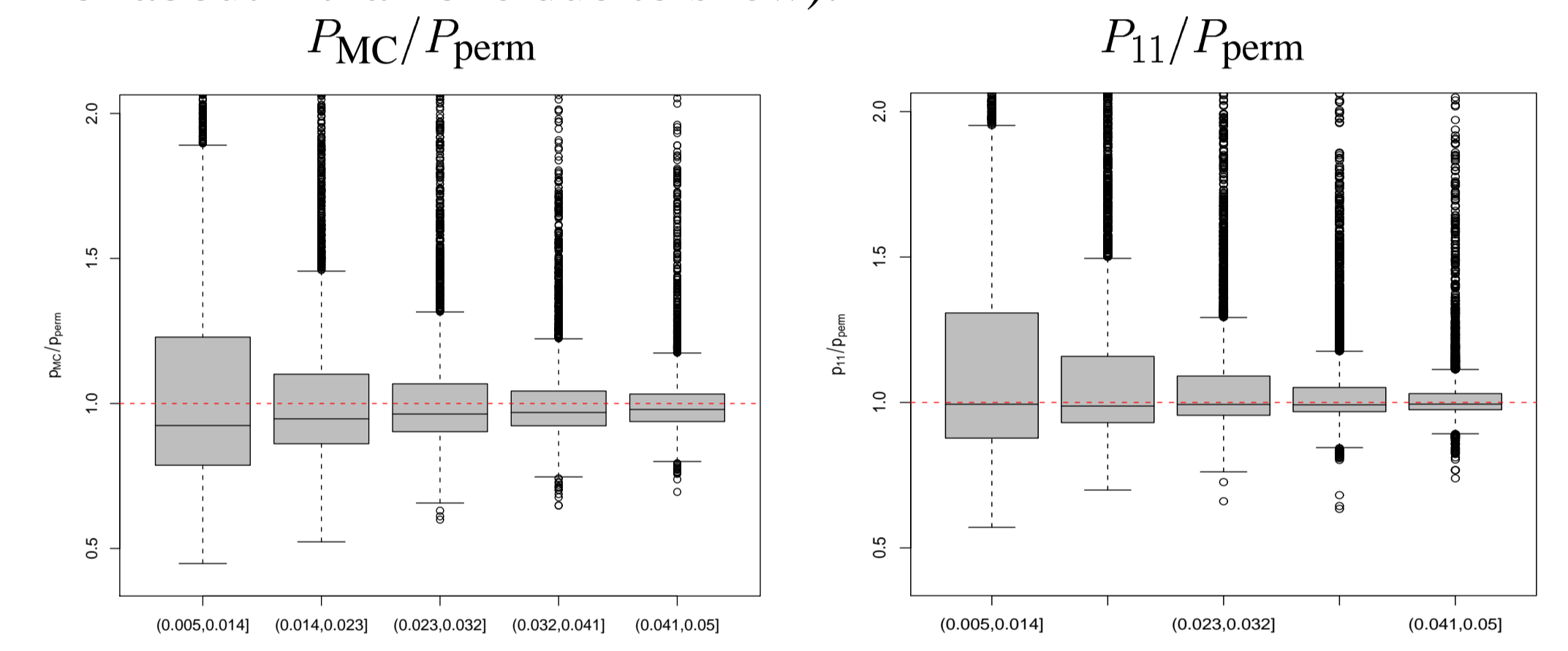


Figure 4: Permutation test applied to the real data, produces p-values that are very close to P_{MC} (left), as predicted by its exactness. There is considerable variability, the right figure shows that the mean permutation P-value is smaller than P_{11} .

Conclusions

While methods have been proposed to account for heterogeneous variance over subjects, the one-sample t-test is still the most common group analysis tool. While the Satterthwaite eDF were found to vary widely depending on the degree of heterogeneous variance, we found the eDF chi-square approximation to be quite poor in this setting. Surprisingly, p-values computed with $n - 1$ DF were not too biased, but were slightly conservative. Since the permutation test makes no assumption about heterogeneous variances, it performed quite well and on average the permutation p-values were a close match to the MC p-values.

The eDF gave results that were very conservative. This is due to the eDF being based on moment-matching, and not necessarily matching the tails of the null distributions.

In summary, if a Mixed Models approach is not taken, the OLS P-values are valid though slightly conservative. A slight improvement in power may be obtained with a nonparametric permutation test.

References

- [1] Holmes & Friston 1998, NI 7:5754. [2] Beckmann *et al.* 2003, NI 20:1052-1063. [3] Woolrich *et al.* 2004, NI 21: 1732-1747. [4] Johansen-Berg *et al.* 2002, PNAS USA99: 14518-14523. [5] Nichols & Holmes 2001, HBM 15:1-25. [6] SnPM2, <http://www.sph.umich.edu/ni-stat/SnPM> [7] FSL, <http://www.fmrib.ox.ac.uk/fsl> [8] Satterthwaite, 1946, Biometrics 110-114. [9] Friston, Stephan, Lund, Morcom, Kiebel, 2005, NI 24:244-252.