Permutation inference for the General Linear Model

Anderson M. Winkler

FMRIB Analysis Group

03.Sep.2015
in jalapeno: \texttt{~winkler/bin/palm}
PALM — Permutation Analysis of Linear Models — is a tool that allows inference using permutation methods, offering a number of features not available in other analysis software. These features currently include:

- Ability to work with volumetric and surface-based formats, including face-wise data, as well as with non-imaging data;
- A range of various regression and permutation strategies;
- Statistics that are robust to heteroscedasticity;
- Shuffling of sets of observations, to allow, for instance, the analysis of certain designs with repeated measurements, with no missing data;
- Shuffling of observations with complex, tree-like covariance structure (such as for the Human Connectome Project);
- Permutation with sign-flipping;
- Non-Parametric Combination (NPC) for joint inference over multiple modalities, or multiple contrasts, or both together, with various combining functions available;
- Classic multivariate statistics (MANOVA, MANCOVA, CCA) for joint inference over multiple modalities, assessed through robust permutation methods, and also parametrically when approximations exist;
- Correction over multiple contrasts, multiple modalities, for images with or without the same size or geometry, including non-imaging data, controlling the FWER or FDR;
- Fast draft mode, in which a minimum number of permutations is performed for each test;
- Tail approximation, in which small permutation p-values are further refined using a continuous asymptotic approximation, even with a small number of permutations.

PALM requires Matlab or Octave. It can be executed from inside either environment, or directly from the shell. It can also be called from scripts.

PALM is experimental software. As novel features are included, tested, verified, and validated, eventually they will be implemented and made available in randomise or in other tools. PALM is for users who are familiar with statistics and willing to use experimental analysis tools.
ON SMALL DIFFERENCES OF SENSATION.

READ OCTOBER 17, 1884.

By C. S. Peirce and J. Jastrow.

The physiological psychologists assume that two nerve excitations alike in quality will only produce distinguishable sensations provided they differ in intensity by an amount greater than a fixed ratio. The least perceptible difference of the excitations divided by half their sum is what they call the Unterschiedsschwelle. Fechner* gives an experiment to prove the fact assumed, namely: He finds that two very dim lights placed nearly in line with the edge of an opaque body show but one shadow of the edge. It will be found, however, that this phenomenon is not a clearly marked one, unless the lights are nearly in range. If the experiment is performed with lateral shifting of one of the lights, and with a knowledge of the effects of a telescope upon the appearance of terrestrial objects at night, it will be found very far from conclusive.

The conception of the psychologists is certainly a difficult one to seize. According to their own doctrine, in which the observed facts seem fully to bear them out, the intensity of the sensation increases continuously with the excitation, so that the least increase of the latter must produce a corresponding increase of the former. And, indeed, the hypothesis that a continuous increase of the excitation would be accompanied by successive discrete increments of the sensation, contradicts as it would be, would not be sufficient to account for a constant Unterschiedsschwelle.
**Why parametric tests were useful?**

**Practical difficulty**

Often it is impossible to repeat the same experiment multiple times.

**Example:** Consider comparing the effect of the sex of the subjects on school scores. The sex cannot be simply changed and the test repeated. The same for, e.g., presence of a disorder. In some cases, costs associated with data collection are high, or the samples are unique.
Why parametric tests were useful?

Instead:
Sample subjects from a population, randomly and independently, and assume a certain distribution for the data. Compute a statistic. All else being equal, an extreme value suggest a difference in the parameter that is allowed to vary (usually the mean).
Why parametric tests were useful?

Computations involved are quite simple:

Calculate the statistic just once, then refer to a table to find what the critical value is for a given significance level.

TABLE A
Normal Distribution

<table>
<thead>
<tr>
<th>z</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>-1.6449</td>
<td>-1.5982</td>
<td>-1.5548</td>
<td>-1.5141</td>
<td>-1.4758</td>
<td>-1.4392</td>
<td>-1.4051</td>
<td>-1.3722</td>
<td>-1.3408</td>
<td>-1.3106</td>
</tr>
<tr>
<td>0.20</td>
<td>-1.2816</td>
<td>-1.2366</td>
<td>-1.1940</td>
<td>-1.1520</td>
<td>-1.1107</td>
<td>-1.0703</td>
<td>-1.0307</td>
<td>-0.9919</td>
<td>-0.9531</td>
<td>-0.9149</td>
</tr>
<tr>
<td>0.30</td>
<td>-1.0364</td>
<td>-0.9932</td>
<td>-0.9511</td>
<td>-0.9105</td>
<td>-0.8708</td>
<td>-0.8317</td>
<td>-0.7933</td>
<td>-0.7554</td>
<td>-0.7183</td>
<td>-0.6813</td>
</tr>
<tr>
<td>0.40</td>
<td>-0.8166</td>
<td>-0.7839</td>
<td>-0.7522</td>
<td>-0.7210</td>
<td>-0.6906</td>
<td>-0.6605</td>
<td>-0.6305</td>
<td>-0.6006</td>
<td>-0.5710</td>
<td>-0.5413</td>
</tr>
<tr>
<td>0.50</td>
<td>-0.6145</td>
<td>-0.5828</td>
<td>-0.5513</td>
<td>-0.5207</td>
<td>-0.4908</td>
<td>-0.4615</td>
<td>-0.4327</td>
<td>-0.4042</td>
<td>-0.3760</td>
<td>-0.3482</td>
</tr>
<tr>
<td>0.60</td>
<td>-0.4234</td>
<td>-0.3910</td>
<td>-0.3595</td>
<td>-0.3287</td>
<td>-0.2983</td>
<td>-0.2684</td>
<td>-0.2390</td>
<td>-0.2098</td>
<td>-0.1807</td>
<td>-0.1518</td>
</tr>
<tr>
<td>0.70</td>
<td>-0.2733</td>
<td>-0.2420</td>
<td>-0.2112</td>
<td>-0.1807</td>
<td>-0.1503</td>
<td>-0.1202</td>
<td>-0.0903</td>
<td>-0.0606</td>
<td>-0.0310</td>
<td>-0.0015</td>
</tr>
<tr>
<td>0.80</td>
<td>-0.0887</td>
<td>-0.0678</td>
<td>-0.0472</td>
<td>-0.0268</td>
<td>-0.0065</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>-0.0127</td>
<td>-0.0110</td>
<td>-0.0095</td>
<td>-0.0080</td>
<td>-0.0065</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table is arranged like an ordinary logarithm table, with values of the variate, z, in the first column and row, and in the body of the table the corresponding values of the sum of the two "tails" beyond the limits + and −z, i.e. 2(1 − Φz) (see page 57).

This table is taken by consent from Statistical Methods for Research Workers, by Professor R. A. Fisher, published at 16a. by Oliver & Boyd, Edinburgh, and attention is drawn to the larger collection in Statistical Tables, by Professor R. A. Fisher and F. Yates, published at 13a. 6d. by Oliver & Boyd, Edinburgh.
- Random sampling is rare in biomedical research.
- Population model is often violated, along with the assumptions that come with it.
Why parametric tests were useful?

Also good for imaging in the first years:

Thousands of voxels need speed, and parametric is much faster.

FIG. 3. The changes in regional CBF induced by left-hand vibration are shown in an averaged, subtracted image. This image was created by averaging 10 individual subtraction images (including that shown in Fig. 1C) after stereotactic anatomical standardization. Image orientation is the same as that of Fig. 2. The color scale is linear, representing CBF change in ml 100 g⁻¹ min⁻¹. The noise suppression provided by averaging allows the relatively weak (2–4 ml) activations in second somatic sensory cortex (S2), supplementary motor area (SMA), and cerebellum to be seen clearly above the background noise. Even the robust (13 ml) activation of primary somatic sensory cortex (S1) is better defined than in the nonaveraged image. See Fig. 1C for comparison.

Fox et al. (1988)
However...

- Assumptions can be invalid, and **will** be invalid in many voxels.
- Different imaging modalities have different properties. A single set of assumptions **will** be invalid for some of them.
Permutation tests are superior

- No underlying theoretical distribution.
- No stringent assumptions (normality, independence, homogeneous variances).
- Non-random samples are permitted (but need exchangeability if there is no randomisation).
- Wide variety of statistics.
- Good for small datasets.
- All information needed is in the dataset itself, not on an idealised population.
- Resilient to outliers (and robust statistics can be used, even without a limiting distribution).
Permutation tests are superior

Only assumption:

Data are exchangeable.

i.e., the joint distribution remains unchanged if data are swapped.
Principles
General Linear Model

\[ Y = M\psi + \epsilon \]

Permute the data \((Y)\) or the rows of the design matrix \((M)\), re-estimate the parameters \((\psi)\) and residuals \((\epsilon)\), compute a new statistic. Repeat this many times, count how often a permuted statistic is greater or equal than the non-permuted, and divide by the number of tests. That produces the **p-value**.
Two-sample $t$ test

Winkler

Permutation for the GLM
Two-sample $t$ test

\[ Y = M\psi + \epsilon \]

\[
\begin{bmatrix}
0.9670 \\
0.5472 \\
0.9727 \\
0.7148 \\
0.6977 \\
0.2161
\end{bmatrix} =
\begin{bmatrix}
1 & 0 \\
1 & 0 \\
1 & 0 \\
0 & 1 \\
0 & 1 \\
0 & 1
\end{bmatrix} \times
\begin{bmatrix}
\psi_1 \\
\psi_2
\end{bmatrix} +
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\epsilon_6
\end{bmatrix}
\]

\[ \psi_i = ? \quad \epsilon_i = ? \]

Estimation: \( \hat{\psi} = (M'\mathbf{M})^{-1}M'Y, \hat{\epsilon} = Y - M\hat{\psi}. \)
Two-sample $t$ test

$$Y = M\psi + \epsilon$$

\[
\begin{pmatrix}
0.9670 \\
0.5472 \\
0.9727 \\
0.7148 \\
0.6977 \\
0.2161
\end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} +0.8290 \\ +0.5429 \end{pmatrix} + \begin{pmatrix} +0.1380 \\ -0.2817 \\ +0.1437 \\ +0.1719 \\ +0.1549 \\ -0.3268 \end{pmatrix}
\]

$$t = +1.3258$$

How likely is a value at least as large as this if there is no effect?
Two-sample $t$ test

$t = +1.33$  $t = -0.91$  $t = +0.25$  $t = -0.93$  $t = +0.42$  $t = +2.24$  $t = +0.41$  $t = -0.45$  $t = -2.40$  $t = -0.47$

$t = +0.47$  $t = +2.40$  $t = +0.45$  $t = -0.41$  $t = -2.24$  $t = -0.42$  $t = +0.93$  $t = -0.25$  $t = +0.91$  $t = -1.33$
Two-sample $t$ test

$t = +1.33$  $t = -0.91$  $t = +0.25$  $t = -0.93$  $t = +0.42$  $t = +2.24$  $t = +0.41$  $t = -0.45$  $t = -2.40$  $t = -0.47$

$t = +0.47$  $t = +2.40$  $t = +0.45$  $t = -0.41$  $t = -2.24$  $t = -0.42$  $t = +0.93$  $t = -0.25$  $t = +0.91$  $t = -1.33$
Two-sample t test

There were 3 cases of a statistic at least as large as the one observed. We have run 20 permutations. Thus:

\[ p = \frac{3}{20} = 0.15 \]
One-sample $t$ test
One-sample *t* test

\[ Y = M\psi + \epsilon \]

\[
\begin{bmatrix}
+0.0734 \\
+0.9999 \\
+0.9684 \\
-0.8775
\end{bmatrix}
= \begin{bmatrix}
1 \\
1 \\
1 \\
1
\end{bmatrix}
\times \begin{bmatrix}
\psi_1 \\
\psi_2 \\
\psi_3 \\
\psi_4
\end{bmatrix}
+ \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4
\end{bmatrix}
\]

\[ \psi_i = ? \quad \epsilon_i = ? \]

Estimation: \( \hat{\psi} = (M'M)^{-1}M'Y, \hat{\epsilon} = Y - M\hat{\psi} \).
One-sample *t* test

\[ Y = M \psi + \epsilon \]

\[
\begin{bmatrix}
+0.0734 \\
+0.9999 \\
+0.9684 \\
-0.8775
\end{bmatrix}
= \begin{bmatrix}
1 \\
1 \\
1
\end{bmatrix}
\times \begin{bmatrix}
+0.2911
\end{bmatrix}
+ \begin{bmatrix}
-0.2176 \\
+0.7089 \\
+0.6773 \\
-1.1686
\end{bmatrix}
\]

\[ t = +0.6544 \]

How likely is a value at least as large as this if there is no effect?
One-sample \( t \) test

\[
\begin{array}{cccccccc}
+1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
+1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
+1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
+1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
+1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
+1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
+1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
+1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
\end{array}
\]

\[
t = +0.65 \quad t = +0.56 \quad t = -0.45 \quad t = -0.54 \quad t = -0.42 \quad t = -0.50 \quad t = -2.70 \quad t = -3.31
\]
One-sample $t$ test

- $t = +0.65$
- $t = +0.56$
- $t = -0.45$
- $t = -0.54$
- $t = -0.42$
- $t = -0.50$
- $t = -2.70$
- $t = -3.31$

- $t = +3.31$
- $t = +2.70$
- $t = +0.50$
- $t = +0.42$
- $t = +0.54$
- $t = +0.45$
- $t = -0.56$
- $t = -0.65$
One-sample \( t \) test

There were 3 cases of a statistic at least as large as the one observed. We have run 16 sign flippings. Thus:

\[
p = \frac{3}{16} = 0.1875
\]
Single test – Parametric

Test statistic
-4 -3 -2 -1 0 1 2 3 4
Relative frequency
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4

Permutation for the GLM
Single test – Permutation

![Graph showing test statistic and relative frequency.](image-url)
Many tests (maximum) – Parametric
Many tests (maximum) – Permutation
Exactness

- If all permutations are done → exact.
- If many permutations are done → exact.

In either sense there are no distributional assumptions. However, in the second sense it means that $P(p \leq \alpha) = \alpha$. 
A bit more on the assumptions

- Exchangeable errors (EE) $\rightarrow$ Permutations.
- Independent and symmetric errors (ISE) $\rightarrow$ Sign-flippings.
- Both (EE + ISE) $\rightarrow$ Permutations + Sign-flippings.
How many permutations are possible?

- Depends on the design: repeated rows reduce the number of possible permutations.
- Regardless, the number grows very quickly, and can be astronomical even with small samples.
  ex.: 24 subjects, single continuous EV, $J = 24! \approx 6 \cdot 10^{23}$. 
How many permutations should we run?

- The $p$-value itself is an estimate, and more permutations narrow confidence interval around it, giving more credence to the results.

- More permutations do not increase power, although too few may not be powerful: smallest $p = 1/J$. 
Permutation for the GLM
**Problem:** Nuisance variables. What do we do?

\[ Y = M\psi + \epsilon \]
Model partitioning

\[ Y = X\beta + Z\gamma + \epsilon \]
Model partitioning: Before

\[ Y = M \psi + \epsilon \]

\[
\begin{bmatrix}
    y_1 \\
    y_2 \\
    y_3 \\
    y_4 \\
    y_5 \\
    y_6 \\
    y_7 \\
    y_8
\end{bmatrix} =
\begin{bmatrix}
    +0.53 & 0 & 1 & 0 \\
    +1.83 & 0 & 1 & 0 \\
    -2.25 & 0 & 1 & 0 \\
    +0.86 & 0 & 1 & 0 \\
    0 & +0.72 & 0 & 1 \\
    0 & -1.06 & 0 & 1 \\
    0 & +1.71 & 0 & 1 \\
    0 & -1.20 & 0 & 1
\end{bmatrix} \times
\begin{bmatrix}
    \psi_1 \\
    \psi_2 \\
    \psi_3 \\
    \psi_4
\end{bmatrix} +
\begin{bmatrix}
    \epsilon_1 \\
    \epsilon_2 \\
    \epsilon_3 \\
    \epsilon_4 \\
    \epsilon_5 \\
    \epsilon_6 \\
    \epsilon_7 \\
    \epsilon_8
\end{bmatrix}
\]
Model partitioning: After

\[ Y = X\beta + Z\gamma + \epsilon \]

\[
\begin{bmatrix}
    y_1 \\
    y_2 \\
    y_3 \\
    y_4 \\
    y_5 \\
    y_6 \\
    y_7 \\
    y_8
\end{bmatrix}
= 
\begin{bmatrix}
    +0.11 \\
    +0.62 \\
    -0.98 \\
    +0.24 \\
    -0.41 \\
    +0.66 \\
    -1.01 \\
    +0.75
\end{bmatrix}
\times 
\begin{bmatrix}
    \beta_1
\end{bmatrix}
+ 
\begin{bmatrix}
    +0.38 & 1 & 0 \\
    +1.29 & 1 & 0 \\
    -1.59 & 1 & 0 \\
    +0.60 & 1 & 0 \\
    +0.51 & 0 & 1 \\
    -0.75 & 0 & 1 \\
    +1.21 & 0 & 1 \\
    -0.85 & 0 & 1
\end{bmatrix}
\times 
\begin{bmatrix}
    \gamma_1 \\
    \gamma_2 \\
    \gamma_3
\end{bmatrix}
+ 
\begin{bmatrix}
    \epsilon_1 \\
    \epsilon_2 \\
    \epsilon_3 \\
    \epsilon_4 \\
    \epsilon_5 \\
    \epsilon_6 \\
    \epsilon_7 \\
    \epsilon_8
\end{bmatrix}
\]
Model partitioning

Features:
- $\hat{\beta} = C'\hat{\psi}$
- $\text{Cov}(\hat{\beta}) = C'C\text{Cov}(\hat{\psi})C$
- $\text{span}(X) \cup \text{span}(Z) = \text{span}(M)$
- $X'Z = 0$

In other words:
The two models are equivalent.
Permutation matrix (P)

<table>
<thead>
<tr>
<th>Permutation matrix</th>
<th>Original data</th>
<th>Permuted data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 0 0 0 0 0 0</td>
<td>-4.01</td>
<td>-1.98</td>
</tr>
<tr>
<td>0 0 0 0 0 0 1 0 0</td>
<td>-2.97</td>
<td>+2.03</td>
</tr>
<tr>
<td>0 0 0 0 1 0 0 0 0</td>
<td>-1.98</td>
<td>0.00</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 0 1</td>
<td>-1.01</td>
<td>+4.02</td>
</tr>
<tr>
<td>0 1 0 0 0 0 0 0 0</td>
<td>0.00</td>
<td>-2.97</td>
</tr>
<tr>
<td>0 0 0 0 0 1 0 0 0</td>
<td>+0.96</td>
<td>+0.96</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 1 0</td>
<td>+2.03</td>
<td>+3.00</td>
</tr>
<tr>
<td>1 0 0 0 0 0 0 0 0</td>
<td>+3.00</td>
<td>-4.01</td>
</tr>
<tr>
<td>0 0 0 1 0 0 0 0 0</td>
<td>+4.02</td>
<td>-1.01</td>
</tr>
</tbody>
</table>
Sign-flipping matrix ($S$)

\[
\begin{array}{cccccccc}
  -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & +1 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
\end{array}
\times
\begin{array}{cccc}
  -4.01 \\
  -2.97 \\
  -1.98 \\
  -1.01 \\
  0.00 \\
  +0.96 \\
  +2.03 \\
  +3.00 \\
  +4.02 \\
\end{array}
= \begin{array}{cccc}
  +4.01 \\
  +2.97 \\
  +1.98 \\
  +1.01 \\
  0.00 \\
  +0.96 \\
  +2.03 \\
  +3.00 \\
  -4.02 \\
\end{array}
\]
Permutation and sign-flipping matrix (B = PS)

Permutation and sign-flipping matrix

Original data

Permuted & sign-flipped

\[
\begin{array}{cccccccccc}
0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\times \begin{array}{cccccccccc}
-4.01 & -2.97 & -1.98 & 0.00 & -1.98 & 0.00 & -2.03 & 0.00 & -2.03 & 0.00 \\
0.00 & +0.96 & +2.03 & +3.00 & +3.00 & +3.00 & +3.00 & +3.00 & +3.00 & +3.00 \\
+1 & +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 & +1 \\
\end{array}
\begin{array}{cccccccccc}
-1.98 & -2.03 & -1.98 & 0.00 & -1.98 & 0.00 & -2.03 & 0.00 & -2.03 & 0.00 \\
+2.97 & +0.96 & +2.03 & +3.00 & +3.00 & +3.00 & +3.00 & +3.00 & +3.00 & +3.00 \\
-1.01 & -1.01 & -4.02 & -4.02 & -4.02 & -4.02 & -4.02 & -4.02 & -4.02 & -4.02 \\
-1.01 & -1.01 & -4.02 & -4.02 & -4.02 & -4.02 & -4.02 & -4.02 & -4.02 & -4.02 \\
\end{array}
\]
H is the hat matrix, such that \( \hat{Y} = HY \), \( R \) is the residual-forming matrix, such that \( \hat{\epsilon} = RY \). \( I = H + R \). The subscript indicates whether these consider the full model (\( M \)) or just the nuisance (\( Z \)). \( P \) is a permutation or sign-flipping matrix.
Freedman–Lane method

1. Regress the data against the full model and compute the unpermuted statistic.

2. Regress the data against the nuisance-only model and compute the residuals.

3. Shuffle the residuals, then add back the estimated nuisance effects.

4. Regress this data against the full model and compute the permuted statistic.

5. Repeat the Steps 2–4 many times.

6. Count how many times the permuted statistic was found to be equal to or larger than the unpermuted, and divide by the number of permutations. The result is the p-value.

There are computational simplifications that can be used.
The \( F \)-statistic

Consider the statistic:

\[
F = \frac{\hat{\psi}' C (C'(M'M)^{-1}C)^{-1} C'\hat{\psi}}{\text{rank}(C)} \left/ \frac{\hat{\epsilon}'\hat{\epsilon}}{N - \text{rank}(M)} \right.
\]

\[
= \frac{\hat{\beta}' (X'X) \hat{\beta}}{\text{rank}(C)} \left/ \frac{\hat{\epsilon}'\hat{\epsilon}}{N - \text{rank}(X) - \text{rank}(Z)} \right.
\]

When \( \text{rank}(C) = 1 \), \( \hat{\beta} \) is a scalar and the \( t \) statistic can be expressed as a function of \( F \) as \( t = \text{sign}(\hat{\beta}) \sqrt{F} \).
The statistic must be pivotal.

- But if there are heterogeneous variances and groups of different sizes, the $F$-statistic is no longer pivotal.
- And with it, the $t$ statistic, the correlation coefficient $r$, and the $R^2$, are no longer pivotal either.
The $G$-statistic

Consider the statistic:

$$G = \frac{\hat{\psi}' C (C'(M'WM)^{-1}C)^{-1} C'\hat{\psi}}{\Lambda \cdot s}$$

where $W$ is a diagonal matrix that has elements:

$$W_{nn} = \frac{\sum_{n' \in g_n} R_{n'n'}}{\hat{\epsilon}'_{gn} \hat{\epsilon}_{gn}}$$

$$\Lambda = 1 + \frac{2(s - 1)}{s(s + 2)} \sum_g \frac{1}{\sum_{n \in g} R_{nn}} \left(1 - \frac{\sum_{n \in g} W_{nn}}{\text{trace}(W)}\right)^2$$

and where $s = \text{rank}(C)$. 
**Variance groups (vgs)**

\( W \) can be seen as a weighting matrix, the square root of which normalises the model such that the errors have then unit variance and can be ignored.
W can be seen as a weighting matrix, the square root of which normalises the model such that the errors have then unit variance and can be ignored.
The $G$-statistic

$G$ generalises a number of well known statistical tests:

<table>
<thead>
<tr>
<th>Condition</th>
<th>$\text{rank } (C) = 1$</th>
<th>$\text{rank } (C) &gt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homoscedastic errors, unrestricted exchangeability</td>
<td>Square of Student’s $t$</td>
<td>$F$-ratio</td>
</tr>
<tr>
<td>Homoscedastic within $\nu G$, restricted exchangeability</td>
<td>Square of Aspin–Welch $\nu$</td>
<td>Welch’s $\nu^2$</td>
</tr>
</tbody>
</table>
Heterogeneous variances: the $G$-statistic

- When rank $(C) = 1$, the $t$-equivalent to the $G$-statistic is
  \[ \nu = \hat{\psi}' C (C'(M'WM)^{-1}C)^{-\frac{1}{2}} \], which is the same $\nu$-statistic for the Behrens-Fisher problem.

- In the absence of variance groups, $G$ and $\nu$ are equivalent to $F$ and $t$ respectively.

- Although not typically necessary, approximate parametric $p$-values for the $G$-statistic can be computed from an $F$-distribution with degrees of freedom $\nu_1 = s$ and $\nu_2 = \frac{2(s-1)}{3(\Lambda-1)}$. 
Examples
**Example 1: Mean effect**

**Scenario**

Consider a multi-subject fMRI study to investigate the BOLD response associated with a novel experimental task. After the first-level analysis (within subject), maps of contrasts of parameter estimates for each subject are used in a second level analysis. The regressor for the effect of interest (the mean effect) is simply a column of ones; nuisance variables, such as handedness, can be included in the model.
## Example 1: Mean effect

<table>
<thead>
<tr>
<th>Coded data (Y)</th>
<th>EB</th>
<th>VG</th>
<th>Model (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( m_1 )</td>
</tr>
<tr>
<td>Subject 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 11</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 12</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Contrast 1 (\( C'_1 \))  \( +1 \)  0
Contrast 2 (\( C'_2 \))  \( -1 \)  0
Example 1: Mean effect

Solution

Permutations of the data or of the design matrix do not change the model with respect to the regressor of interest. However, by treating the errors as symmetric, instead of permutation, the signs of the ones in the design matrix, or of each datapoint, can be flipped randomly to create the empirical distribution from which inference can be performed. The procedure can be performed as in either the Freedman–Lane or Smith methods.
Example 2: Multiple regression

Scenario

Consider the analysis of a study that compares patients and controls with respect to brain cortical thickness, and that recruiting process ensured that all selected subjects are exchangeable. Elder subjects may, however, have thinner cortices, regardless of the diagnosis. To control for the confounding effect of age, it is included in the design as a nuisance regressor. Sex is also included.
## Example 2: Multiple regression

<table>
<thead>
<tr>
<th>Coded data (Y)</th>
<th>EB</th>
<th>VG</th>
<th>Model (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( m_1 )</td>
</tr>
<tr>
<td>Subject 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subject 7</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Subject 8</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Subject 9</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Subject 10</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Subject 11</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Subject 12</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Contrast 1 (( C'_1 ))</td>
<td>+1</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>Contrast 2 (( C'_2 ))</td>
<td>-1</td>
<td>+1</td>
<td>0</td>
</tr>
</tbody>
</table>
Example 2: Multiple regression

Solution

The permutation strategy follows the Freedman–Lane or Smith methods, with the residuals of the reduced model being permuted under unrestricted exchangeability.
Example 4: Unequal group variances

Scenario

Consider a study using FMRI to compare whether the BOLD response associated with a certain cognitive task would differ among subjects with autistic spectrum disorder (ASD) and control subjects, while taking into account differences in age and sex. In this hypothetical example, the cognitive task is known to produce more erratic signal changes in the patient group than in controls. Therefore, variances cannot be assumed to be homogeneous with respect to the group assignment of subjects.
Example 4: Unequal group variances

<table>
<thead>
<tr>
<th>Coded data (Y)</th>
<th>EB</th>
<th>VG</th>
<th>Model (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>m₁ m₂ m₃ m₄</td>
</tr>
<tr>
<td>Subject 1</td>
<td>1</td>
<td>1</td>
<td>1 0 a₁ s₁</td>
</tr>
<tr>
<td>Subject 2</td>
<td>1</td>
<td>1</td>
<td>1 0 a₂ s₂</td>
</tr>
<tr>
<td>Subject 3</td>
<td>1</td>
<td>1</td>
<td>1 0 a₃ s₃</td>
</tr>
<tr>
<td>Subject 4</td>
<td>1</td>
<td>1</td>
<td>1 0 a₄ s₄</td>
</tr>
<tr>
<td>Subject 5</td>
<td>1</td>
<td>1</td>
<td>1 0 a₅ s₅</td>
</tr>
<tr>
<td>Subject 6</td>
<td>1</td>
<td>1</td>
<td>1 0 a₆ s₆</td>
</tr>
<tr>
<td>Subject 7</td>
<td>2</td>
<td>2</td>
<td>0 1 a₇ s₇</td>
</tr>
<tr>
<td>Subject 8</td>
<td>2</td>
<td>2</td>
<td>0 1 a₈ s₈</td>
</tr>
<tr>
<td>Subject 9</td>
<td>2</td>
<td>2</td>
<td>0 1 a₉ s₉</td>
</tr>
<tr>
<td>Subject 10</td>
<td>2</td>
<td>2</td>
<td>0 1 a₁₀ s₁₀</td>
</tr>
<tr>
<td>Subject 11</td>
<td>2</td>
<td>2</td>
<td>0 1 a₁₁ s₁₁</td>
</tr>
<tr>
<td>Subject 12</td>
<td>2</td>
<td>2</td>
<td>0 1 a₁₂ s₁₂</td>
</tr>
</tbody>
</table>

Contrast 1 (C₁'): +1 -1 0 0
Contrast 2 (C₂'): -1 +1 0 0
Example 4: Unequal group variances

Solution

This is an example of the classical Behrens–Fisher problem. To accommodate heteroscedasticity, two permutation blocks are defined according to the group of subjects. Under the assumption of independent and symmetric errors, the problem is solved by means of random sign-flipping, using the well known Welch’s $v$ statistic, a particular case of the $G$ statistic.
That’s all folks.

Twitter: @AndersonWinkler | Blog: brainder.org