Chapter Six

A Non-Parametric Approach

In this chapter, a non-parametric approach to assessing functional mapping experiments is presented. A multiple comparisons randomisation test is developed for simple activation studies, which is shown to maintain strong control over familywise Type I error. A step-down procedure with strong control is introduced, and computationally feasible algorithms presented. The methods are illustrated on a real PET data set, with a pseudo *t*-statistic image formed with a smoothed variance estimate. For the given data set the approach is found to outperform many of the parametric methods, particularly with the pseudo *t*-statistic. This, together with the flexibility and guaranteed validity of a non-parametric method, makes the approach very attractive, despite the computational burden imposed. The practicalities of the method are discussed, including extensions to other experimental paradigms, other test statistics, and permutation tests.

A paper, with similar content to this chapter, has been accepted for publication in the *Journal of Cerebral Blood Flow and Metabolism* (Holmes *et al.*, 1995).

6.1. Introduction and Motivation

Current methods for assessing the significance of statistic images are either parametric, or rely on simulation. Currently available approaches were discussed extensively in chapter 3, where they were found to be lacking in some areas.

Shortcomings of parametric methods

Parametric approaches are based on the assumption of specific forms of probability distribution for the voxel values in the statistic images. Hypotheses are specified in terms of the parameters of the assumed distributions. Usually, scan data are taken to be normally distributed, giving known distributional forms for certain statistics. These statistic images are then assessed using (approximate) results for continuous random fields, under the assumption that the statistic image is a good lattice representation of a continuous random field with matching marginal distribution and variance-covariance matrix of partial derivatives. This use of continuous random fields was seen to be particularly inappropriate for *t* and *F* statistics whose denominator has low degrees of freedom (§3.3.6.5.). Thus, parametric methods restrict the form of the voxel statistic to those for which distributional results are available, and rely on a multitude of assumptions and approximations, the validity of which is often in doubt, but seldom checked.

Simulation approaches require the simulation of null statistic images whose properties match those of true null statistic images, a match which is often dubious as discussed in §3.5.1.

Non-parametric methods

Having encountered problems with classical parametric methods when analysing EEG data, Blair *et al*. (1994) applied non-parametric methods. Originally expounded by Fisher (1935), Pitman (1937a, 1937b), and later Edgington (1964, 1969a, 1980), these methods are receiving renewed interest as modern computing power makes the computations involved feasible. See Edgington (1969a) for a thorough and readable exposition of randomisation tests.

Parametric methods make formal assumptions about the underlying probability model, up to the level of a set of unknown parameters. Scientific hypotheses formulated in terms of these parameters are then tested on the basis of the assumptions. In contrast, non-parametric methods test simple hypotheses about the mechanism generating the data, using minimal assumptions.

Non-parametric approach for functional mapping experiments

In the remainder of this chapter the theory for randomisation and permutation tests for functional mapping experiments is developed. For simplicity, we shall concentrate on the simple multiple subject activation experiment, with statistic image to be assessed using a single threshold. As we shall see, the approach is by no means limited to this scenario.

6.2. Theory

The rationale behind randomisation tests and permutation tests is intuitive and easily understood. In a simple activation experiment, the scans are labelled as "baseline" or "active" according to the condition under which the scan was acquired, and a statistic image is formed on the basis of these labels. If there is really no activation effect then the labelling of the scans as "baseline" and "active" are artificial, and any labelling of the scans would lead to an equally plausible statistic image. Under the null hypothesis that the labelling is arbitrary, the observed statistic image is randomly chosen from the set of those formed with all possible labellings. If each possible statistic image is summarised by a single statistic, then the probability of observing a statistic more extreme than a given value, is simply the proportion of the possible statistic images with summary statistic exceeding that value. Hence, *p*-values can be computed and tests derived. In this section we formalise this heuristic argument, concentrating on randomisation tests, where the probabilistic justification for the method comes from the initial random assignment of conditions to scan times.

Experiment

Consider the following simple multi-subject activation experiment with N subjects, each scanned repeatedly under two conditions denoted by A and B, with M repetitions of each condition. The conditions are presented alternately to each individual. Half the subjects are randomly chosen to receive condition A first, then B, followed by (M-1) further AB pairs (AB order, conditions presented ABAB…). The other half of the subjects receive condition B first (BA order, conditions presented BABA…). The randomisation of subjects to condition presentation order in this way prevents linear time effects confounding any condition effect in the statistic image.

6.2.1. Statistic images

We shall consider a proportional scaling approach to the normalisation for changes in gCBF (gA), constructing paired *t*-statistic images as described in §2.3.1., generalising to "pseudo" *t*-statistics calculated using smoothed variance images. We adopt this approach because of its simplicity, robustness, and because it illustrates some of the problems with statistic images with low degrees of freedom. Note that any method for producing (statistic) images, whose extreme values indicate activation, can be used. In particular, more general modelling of the effect of global changes via ANCOVA is possible, at an increased computational burden.

Notation

Recall our notation: Y'_{ijqk} denotes the rCBF (rA) measurement at voxel k =1,…,K, of scan j =1,…,M, under condition q = 0,1 (0="rest"), on subject i =1,…N; after normalisation by proportional scaling as described in §2.1.2. Let W be the set of (indices) of the voxels covering the region of interest, W={1,…,K}. Let $\boldsymbol{x}_k$ be the centre of voxel k.

Paired t-statistic image

The paired *t*-statistic image for the study is computed as described in §2.3.1.1. To recap, the statistic at voxel k, T_k, is given by:

$$T_k = \frac{\overline{\Delta}_{\bullet k}}{\sqrt{S_k^2/N}} \tag{21}$$

where $\Delta_{ik} = \overline{Y'}_{i\bullet 1k} - \overline{Y'}_{i\bullet 0k}$ (20)

$$\overline{\Delta}_{\bullet k} = \frac{1}{N}\sum_{i=1}^{N}\Delta_{ik} \tag{22}$$

and $$S_k^2 = \frac{1}{N\text{-}1}\sum_{i=1}^{N}\left(\Delta_{ik} - \overline{\Delta}_{\bullet k}\right)^2 \tag{23}$$

Variance "smoothing"

Assuming $\Delta_{ik} \sim N(\mu_k,\sigma_k^2)$; then under $H_k{:}\mu_k = 0$, $T_k \sim t_{N\text{-}1}$, a Student's *t* distribution with N-1 degrees of freedom. Since the number of subjects, N, is typically between six and twelve, the degrees of freedom are low, and the *t*-statistic image exhibits a high degree of (spatial) noise. As discussed in §3.3.6.5., and as seen for the "V5" study data (§2.6.1.), this noise is inherited from the sample variance image.

However, physical and physiological considerations would suggest that the true error variance image is smooth, being approximately constant over small localities. This suggests the use of a locally pooled variance estimate, formed by pooling variance estimates across neighbouring voxels, possibly weighting the contribution of voxels to reflect their displacement from the voxel where the estimate is sought. This effectively smoothes the variance image, giving smooth statistic images with no loss of resolution. The noise has been smoothed but not the signal. This idea is not new, but has not been pursued until now because the distribution of these locally pooled variance estimates is unknown, precluding any analysis in a parametric manner.

Consider a weighted, locally pooled estimate of the sample variance at voxel k, SS_k, obtained by convolving a Gaussian kernel of dispersion Σ with the voxel level sample variance image (eqn.69):

$$SS_k^2 = \frac{\sum_{k' \in W} f(\mathbf{x}_k - \mathbf{x}_{k'})S_{k'}^2}{\sum_{k' \in W} f(\mathbf{x}_k - \mathbf{x}_{k'})} \tag{69}$$

Here $f(\boldsymbol{x}) = \exp(-\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}/2) / \sqrt{(2\pi)^D |\Sigma|}$ is the Gaussian kernel. Since summation is over the intracerebral voxels, the filter kernel is truncated at the edge of the intracerebral volume.

"Pseudo" t-statistic image

Using this smoothed sample variance in the formula for the *t*-statistic (eqn.21), gives us a pseudo *t*-statistic image (eqn.70):

$$T_k = \frac{\overline{\Delta}_{\bullet k}}{\sqrt{SS_k^2/N}} \tag{70}$$

We shall use the same notation (T_k) for both the pseudo *t*-statistics and the "raw" ones, since the theory to be presented is applicable in either case.

6.2.2. Null hypothesis and labellings

If the two conditions of the experiment affect the brain equally, then, for any particular scan, the acquired image would have been the same had the scan been acquired under the other condition. This leads to voxel hypotheses for no activation at each voxel as:

H_k : Each subject would have given the same set of rCBF (rA) measurements at voxel *k*, were the conditions for that subject reversed

The hypotheses relate to the data, which are regarded as fixed. Under the omnibus hypothesis H_W, any of the possible allocations of conditions to scan times would have given us the same scans. Only the labels of the scans as A and B would be different, and under H_W, the labels of the scans as A or B are arbitrary. The possible labellings are those that could have arisen out of the initial randomisation. In this case the possible labellings are those with half the subjects scans labelled in AB order and half BA order, giving $L = {}_N C_{N/2} = N!/((N/2)!)^2$ possibilities. Thus, under H_W, if we re-randomise the labels on the scans to correspond to another possible labelling, the statistic image computed on the basis of these labels is just as likely as the statistic image computed using the labelling of the experiment, because the initial randomisation could equally well have allocated this alternative labelling.

Re-labelled t-statistic images

For the experimental situation under consideration, the possible labellings are of ABAB… or BABA… for each subjects scans. That is, each subjects scans are labelled either the same as in the actual experiment, or completely the opposite. Thus, each possible labelling can be specified by specifying for each subject whether the labelling is identical or opposite to the actual labelling of the experiment. Let ${}_l\boldsymbol{\delta} = ({}_l\delta_1,\ldots,{}_l\delta_N)$, $l = 1,\ldots,L$, be a set of N-vectors with elements ${}_l\delta_i = +1$ if under labelling l subject i is labelled identically to the actual experiment, and ${}_l\delta_i = -1$ if the labelling is the opposite.

Let ${}_lt_k$ be the value at voxel k of the t-statistic image computed for labelling l of the scans, for $l = 1,\ldots,L$. Then ${}_lt_k$ can be easily computed as:

$$ {}_lt_k = \frac{\overline{{}_l\Delta}_{\bullet k}}{\sqrt{{}_lS_k^2/N}} \tag{71} $$

where $\overline{{}_l\Delta}_{\bullet k} = \frac{1}{N}\sum_{i=1}^{N} {}_l\delta_i\,\Delta_{ik}$

with $\Delta_{ik} = \overline{Y'}_{i\bullet 1k} - \overline{Y'}_{i\bullet 0k}$ as before (20)

and $${}_lS_k^2 = \frac{1}{N\text{-}1}\sum_{i=1}^{N}\left({}_l\delta_i\,\Delta_{ik} - \overline{{}_l\Delta}_{\bullet k}\right)^2$$

$$ \frac{1}{N\text{-}1}\sum_{i=1}^{N}\Delta_{ik}^2 - \frac{N\left(\overline{{}_l\Delta}_{\bullet k}\right)^2}{N\text{-}1} \tag{72} $$

Thus, provided the sum of the squares of the subject differences is retained, the t-statistic for each labelling only requires the computation of the new study mean difference image, with voxel values $\overline{{}_l\Delta}_{\bullet k}$, and a few computations to derive the sample variance and t-statistic image. For "pseudo" t-statistic images, the sample variance image must be smoothed before computation of the t-statistic image.

A further computational "trick" arises from the observation that for each labelling of the scans, the opposite labelling is also possible, and the t-statistic images for these two opposite labellings are negatives of one another. Hence, t-statistic images need only be computed for half of the possible labellings.

Usually, labelling $l = 1$ is taken to be the labelling corresponding to the actual conditions of the experiment, so ${}_1\delta_i = +1$. In this case, $\overline{{}_l\Delta}_{\bullet k} = \overline{\Delta}_{\bullet k}$, ${}_1S_k^2 = S_k^2$, and ${}_1t_k = T_k$.

Randomisation distributions

For a single threshold test, rejection or acceptance of the omnibus hypothesis is determined by the maximum value in the statistic image. The consideration of a maximal statistic deals with the multiple comparisons problem. Let $T_{\max/W}$ denote the maximum of the observed statistic image T searched over voxels (with indices) in the set W; $T_{\max/W} = \max\{T_k : k \in W\}$. It is the distribution of this maximal statistic that is of interest.

Under H_W, the statistic images corresponding to all possible labellings are equally likely, so the maximal statistics of these images are also equally likely. Let ${}_lt_{\max/W}$ be the maximum value (searched over the intracerebral voxels W) of the statistic image computed for labelling l; $l = 1,\ldots,L$. This set of statistics, each corresponding to a possible randomisation of the labels, we call the randomisation values for the maximal statistic. When H_W is true, $T_{\max/W}$ is as likely as any of the randomisation values, because the corresponding labellings were equally likely to have been allocated in the initial selection of labels for the experiment. This gives the randomisation distribution of the maximal statistic, given the data and the assumption that the omnibus hypothesis H_W is true, as $\Pr(T_{\max/W} = {}_lt_{\max/W} \mid H_W) = 1/L$ (assuming that the ${}_lt_{\max/W}$ are distinct).

6.2.3. Single threshold test

From the above, the probability (under H_W) of observing a statistic image with maximum intracerebral value as, or more extreme than, the observed value $T_{max/W}$, is simply the proportion of randomisation values greater than or equal to $T_{max/W}$. This gives a *p*-value for a one sided test of the omnibus null hypothesis.

This *p*-value will be less than 0.05 if $T_{max/W}$ is in the largest 5% of the randomisation values, which it is if and only if it is greater than the 95th percentile of the randomisation values. Thus for a test with weak control over FWE at level 0.05, a suitable critical value is this 95th percentile. The probability of rejecting a true omnibus null hypothesis is the probability that any voxels in the observed statistic image have values exceeding the critical threshold. If any voxel values exceed the threshold then the maximal one does, and the probability of this is at most 0.05 when the omnibus null hypothesis is true.

In general, for a level α test, let $c = \lfloor \alpha L \rfloor$, αL rounded down. The appropriate critical value is then the c+1 th largest of the ${}_l t_{max/W}$, which we denote by ${}_{(c+1)} t_{max/W}$. The observed statistic image is thresholded, declaring as activated those voxels with value strictly greater than this critical value. There are c randomisation values strictly greater than ${}_{(c+1)} t_{max/W}$ (less if ${}_{(c+1)} t_{max/W} = {}_{(c)} t_{max/W}$), so the probability of type I error is:

$$\Pr(T_{max/W} > {}_{(c+1)} t_{max/W} \mid H_W) \le c/L = \lfloor \alpha L \rfloor / L \le \alpha \qquad (73)$$

This becomes an equality if there are no ties in the sampling distribution, and if αL is an integer. Ties occur with probability zero for the maxima of statistic images from continuous data. The size of the test is less than $1/L$ smaller than α, depending on the rounding of αL. Weak control over Type I error is maintained. Thus, the test is (almost) exact, with size (almost) equal to the given level α. Further, this test has strong control over FWE:

Proof of strong control for single threshold test

To prove strong control the test has to be shown to be valid for an arbitrary subset of the intracerebral voxels.

Consider a subset U of the intracerebral voxels, $U \subseteq W$. A randomisation test for the omnibus null hypothesis H_U for this region would proceed as above, but using the randomisation distribution of the maximal statistic searched over the voxels in U. Denote this maximal statistic by $T_{max/U} = \max\{T_k : k \in U\}$, and the randomisation values by ${}_l t_{max/U}$. Then, in notation corresponding to that above, the critical value is ${}_{(c+1)} t_{max/U}$, the c+1 th largest member of the sampling distribution of the maximal statistic searched over voxels in the set U.

Clearly ${}_l t_{max/U} \le {}_l t_{max/W}$ $l = 1,\ldots,L$; since $U \subseteq W$. This inequality also remains true once the two randomisation distributions are ordered (appendix F). In particular ${}_{(c+1)} t_{max/U} \le {}_{(c+1)} t_{max/W}$. That is, the appropriate threshold for the test applied to volume U is at most the critical value for the threshold test for the whole intracerebral volume W. Therefore:

$$\begin{aligned} \Pr(T_{max/U} > {}_{(c+1)} t_{max/W} \mid H_U) &\le \Pr(T_{max/U} > {}_{(c+1)} t_{max/U} \mid H_U) \\ &= c/L = \lfloor \alpha L \rfloor / L \\ &\le \alpha \end{aligned} \qquad (74)$$

In words: Considering voxels in the set U, the threshold computed for all the intracerebral voxels is greater than (or equal to) that appropriate for testing H_U alone,

resulting in a valid (but possibly conservative) test for this subset of the intracerebral voxels. Thus, a test thresholding the observed statistic image T at critical value ${}_{(c+1)}t_{\max/W}$ derived as above, has strong control over type I error.

Strong control with smoothed variance

The above proof relies on subset pivotality of the ${}_{l}t_k$. That is, that the t-statistic images for all possible labellings are identical under the restrictions H_U and H_W, for all voxels in U. This is assured under H_U at voxel $k \in U$ only for statistics computed locally at that voxel. If a smoothed variance estimate is used, then this condition is not maintained, the above proof breaks down, and strong control cannot be claimed.

However, since the effect of variance smoothing is local, it may be intuitively claimed that strong control is maintained in a broad sense. Consider a subset U of W, and let U' consist of the voxels in U and those voxels surrounding U whose sample variances contribute to the smoothed variance at voxels in U. Then, given $H_{U'}$, the test of H_U is valid.

An alternative solution to this predicament is to redefine the voxel hypotheses in terms of the computed statistic image, as follows:

$$H_k: \begin{array}{l}\text{The computed statistic at voxel } k \text{ would have been the same,}\\ \text{were the conditions for any of the subjects reversed}\end{array}$$

Two-sided test

For a two sided test to detect activation and deactivation, the statistic image is thresholded in absolute value. The randomisation distribution for the maximal absolute intracerebral value in the statistic image is computed exactly as above, with maximum value replaced by maximum absolute value. For every possible labelling the exact opposite labelling is also possible, giving statistic images that are the negatives of each other, and hence with the same maximum absolute intracerebral value. Thus the randomisation values are tied in pairs, effectively halving the number of possible labellings.

Single-step adjusted p-value image

A p-value for the observed maximal statistic has already been derived. For other voxels, p-values can be similarly computed. The p-value is the proportion of the randomisation values for the maximal statistic which are greater than or equal to the voxels value. These p-values are known as *single step adjusted p-values* (Westfall & Young, 1993, §2.3.3), giving single step adjusted p-value image $\tilde{\boldsymbol{P}}^{\boldsymbol{ss}}$ for these data:

$$\tilde{P}^{ss}{}_k = \begin{array}{l}\text{proportion of randomisation distribution of maximal}\\ \text{statistic greater than or equal to } T_k\end{array} \qquad (75)$$

Proof of validity of single-step adjusted p-values

A voxel with associated p-value $\tilde{P}^{ss}{}_k \le \alpha$, must have value ($T_k$) exceeded or equalled by at most αL randomisation values for the maximal statistic, by the definition of the adjusted p-values. There are $c+1$ members of the sampling distribution of the maximal statistic greater or equal to the critical threshold ${}_{(c+1)}t_{\max/W}$, and since $c+1 = \lfloor \alpha L \rfloor + 1 > \alpha L$, T_k must exceed ${}_{(c+1)}t_{\max/W}$.

Similarly, if a voxel has value T_k exceeding the critical threshold ${}_{(c+1)}t_{\max/W}$, then T_k must be exceeded or equalled by at most c randomisation values for the maximal statistic, so the single step adjusted p-value at this voxel must be at most α.

Hence, thresholding the single step adjusted p-value image at α is equivalent to thresholding the observed statistic image at ${}_{(c+1)}t_{\max/W}$, for $c = \lfloor \alpha L \rfloor$.

6.2.4. Multi-Step tests

So far, we have been considering a single threshold test, a single-step method in the language of multiple comparisons (Hochberg & Tamhane, 1987). The critical value is obtained from the randomisation distribution of the maximal statistic over the whole intracerebral volume, either directly, or via adjusted *p*-value images. It is somewhat disconcerting that all voxel values are compared with the distribution of the maximal one. Shouldn't only the observed maximal statistic be compared with this distribution?

Secondary activation conservativeness

An additional cause for concern is the observation that an activation that dominates the statistic image, affects the randomisation distribution of the maximal statistic. A strong activation will influence the statistic images for re-labellings, particularly those for which the labelling is close to that of the experiment: The activation may possibly dominate the re-labelled statistic images for some re-labellings, leading to higher values in the randomisation distribution than were there no activation. Thus, a strong activation could increase the critical value of the test. This does not affect the validity of the test, but makes it more conservative for voxels other than that with maximum observed statistic, as indicated by eqn.74. In particular, the test would be less powerful for a secondary activation in the presence of a large primary activation, than for the secondary activation alone. Shouldn't regions identified as activated be disregarded, and the sampling distribution for the maximal statistic over the remaining region used?

We now consider step-down tests, extensions of the single step procedures. These are designed to address the issues raised above, and are likely to be more powerful.

6.2.4.1. Step-down test

The step-down test described here is a sequentially rejective test (Holm, 1979), adapted to the current application of randomisation testing. Starting with the intracerebral voxels, the *p*-value for the maximal statistic is computed as described above. If this *p*-value is greater than α the omnibus hypothesis is accepted. If not, then the voxel with maximum statistic is declared as activated, and the test is repeated on the remaining voxels, possibly rejecting the null hypotheses for the voxel with maximal value over this reduced set of voxels. This is repeated until a step rejects no further voxel hypothesis, when the sequence of tests stops. Thus, activated voxels are identified, cut out, and the remainder of the volume of interest analysed, the process iterating until no more activated voxels are found.

The algorithm is as follows:

1)	Let $k^{(1)},\ldots,k^{(K)}$ be the indices of the intracerebral voxels, ordered such that the corresponding voxel values in the observed *t*-statistic image go from largest to smallest. That is, $T_{k^{(1)}} = T_{\max/W}$, the maximum intracerebral statistic value, and $T_{k^{(K)}}$ the minimum. (Voxels with tied values may be ordered arbitrarily.)
2)	Set $i = 1$, $R = \phi$ (the empty set), $c = \lfloor \alpha L \rfloor$
3)	Compute $\{ {}_l t_{\max/W_i} \}_{l=1}^{L}$, the L randomisation values for the maximal value of the statistic image searched over voxels $W_i = W \setminus R = \{k^{(i)},\ldots,k^{(K)}\}$.
4)	Compute *p*-value $P'_{k^{(i)}}$, as the proportion of the randomisation distribution just computed greater than or equal to $T_{k^{(i)}}$
5)	If $P'_{k^{(i)}}$ is less than or equal to α, then $H_{k^{(i)}}$ can be rejected: Add $k^{(i)}$ to set R, increase i by one, and return to step (3). If $H_{k^{(i)}}$ cannot be rejected, or if there are no more voxels to test, then continue to step (6).
6)	Reject voxel hypotheses H_k for voxels $k \in R$. If voxel hypotheses have been rejected then the omnibus hypothesis H_W is also rejected.
7)	The corresponding threshold is ${}_{(c+1)}t_{\max/\overline{R}}$, for $\overline{R} = W \setminus R$. This is the $c+1$ th largest member of the last sampling distribution calculated.

Algorithm (a)

Points (3)–(5) constitute a "step". The test proceeds one voxel per step, from the voxel with largest value in the observed statistic image, towards that with the smallest value. The set of voxels with rejected hypotheses, R, is added to one voxel per step until a non-significant voxel hypothesis is found, when the algorithm stops. $H_{k^{(i)}}$ is tested in step i, at which point W_i is the set of voxels not already rejected.

This defines a protected sequence of tests. Each test "protects" those following it in that the omnibus null hypothesis for the remaining region must be rejected in order to proceed to subsequent tests. In particular, the first test protects the entire sequence. This first test is simply the test of the overall omnibus hypothesis, discussed in §6.2.3. above. Therefore, the multi-step and single-step tests come to the same conclusion regarding the omnibus hypothesis. Hence, the multi-step test maintains weak control over FWE. Strong control is also maintained:

Proof of strong control of FWE for multi-step test

Consider a subset U of the intracerebral voxels, $U \subseteq W$, with H_U true. Let r be the rank of the maximum statistic for voxels in U, so that $T_{\max/U} = T_{k^{(r)}}$, in the notation of part (1). Clearly H_U is rejected by the step-down method if and only if $H_{k^{(r)}}$ is.

$H_{k^{(r)}}$ is tested if and only if, in preceding steps (with $i < r$), all $H_{k^{(i)}}$ are rejected. At step r, at which $H_{k^{(r)}}$ is tested, $W_r = \{k^{(r)},\ldots,k^{(K)}\}$ is the set of voxels not already rejected, so $U \subseteq W_r$ (by construction of the $k^{(i)}$ in part (1), assuming any ties for the voxel with maximum value in U are broken as they are in part (1)). $H_{k^{(r)}}$ is rejected if $T_{k^{(r)}}$ is in the top $100\alpha\%$ of the sampling distribution of the maximal statistic, computed over voxels in W_i. When H_U is true, the probability of this is at most α, since the situation is the same

as in eqn.74, with W replaced by W_i. Thus the probability of falsely rejecting H_U, given that the step-down test reaches voxel $k^{(r)}$, is at most α. Therefore, for any set of voxels U with H_U true, the probability of false rejection is at most α. Thus, the step-down test controls Type I FWE in the strong sense.

Step-down adjusted p-values

An adjusted *p*-value image corresponding to the test is computed by enforcing monotonicity of the *p*-values computed in part 4, so that once the adjusted *p*-value exceeds α no further voxels are declared significant. This adds a further part to the algorithm:

8) Enforce monotonicity of the *p*-values to obtain step-down adjusted *p*-values:

$$
\begin{aligned}
\tilde{P}^{sd}_{k^{(1)}} &= P'_{k^{(1)}} \\
\tilde{P}^{sd}_{k^{(2)}} &= \max\{\tilde{P}^{sd}_{k^{(1)}}, P'_{k^{(2)}}\} \\
&\vdots \qquad\qquad \vdots \\
\tilde{P}^{sd}_{k^{(i)}} &= \max\{\tilde{P}^{sd}_{k^{(i-1)}}, P'_{k^{(i)}}\} \\
&\vdots \qquad\qquad \vdots \\
\tilde{P}^{sd}_{k^{(K)}} &= \max\{\tilde{P}^{sd}_{k^{(K-1)}}, P'_{k^{(K)}}\}
\end{aligned}
$$

Algorithm (b)

Note that it will only be possible to compute these adjusted *p*-values for voxels for which the P'_k were computed in algorithm (a). Since computation halts at the first non-significant voxel hypothesis, these voxels are those whose null hypotheses are rejected, plus the voxel with largest value whose hypothesis is accepted. The advantage of forming the full adjusted *p*-value image would be that the test level α need not be specified in advance of computations. Voxels declared activated by the step-down test are precisely those with step-down adjusted *p*-value less than or equal to α.

6.2.4.2. Step-down in jumps variant

In this form, the test is too computationally intensive to be useful, involving the computation of a new randomisation distribution at each step. It is possible to accelerate the algorithm, by using the single threshold test at each step to identify any activated voxels and then reject them *en masse*, rather than considering only the voxel with maximal statistic at each step. This "jump-down" variant provides a computationally feasible equivalent test. The algorithm is as follows:

1)	As above
2)	Set $i = 1$, $R = \phi$ (the empty set), $c = \lfloor \alpha L \rfloor$
3)	Compute $\{ {}_l t_{\max/W_i} \}_{l=1}^{L}$, the L randomisation values for the maximal value of the statistic image searched over voxels $W_i = W \backslash R = \{k^{(i)},\ldots,k^{(K)}\}$.
4)	Compute the critical threshold for this step as ${}_{(c+1)}t_{\max/W_i}$.
5)	If $T_{k^{(i)}} > {}_{(c+1)}t_{\max/W_i}$, then $H_{k^{(i)}}$ can be rejected: Let r be the largest member of $\{i,\ldots,K\}$ such that $T_{k^{(r)}} > {}_{(c+1)}t_{\max/W_i}$. (So $\{k^{(i)},\ldots,k^{(r)}\}$ is the set of remaining voxels whose values exceed the critical threshold of this step.) Add $\{k^{(i)},\ldots,k^{(r)}\}$ to set R, set $i = r+1$, and return to step (3). If $H_{k^{(i)}}$ cannot be rejected, or if there are no more voxels to test, then continue to step (6).
6)	Reject voxel hypotheses H_k for voxels $k \in R$. If voxel hypotheses have been rejected then the omnibus hypothesis H_W is also rejected.
7)	The corresponding threshold is ${}_{(c+1)}t_{\max/\overline{R}}$, for $\overline{R} = W \backslash R$. This is the $c+1$ th largest member of the last sampling distribution calculated.

Algorithm (c)

Points (3)–(5) constitute a "step".

Equivalence of step-down and jump-down algorithms

We now prove that the step-down and the jump-down algorithms are equivalent.

($\Rightarrow$) Step-down rejects $H_k \Rightarrow$ jump-down rejects H_k (by contradiction).

Suppose H_k is rejected by the step-down algorithm. Let r be the rank of T_k, so that $k = k^{(r)}$, in the notation of point (1). Since the step-down test rejects $H_{k^{(r)}}$, $H_{k^{(i)}}$ must also be rejected, for $i = 1,\ldots,r$. Let $W_i = \{k^{(i)},\ldots,k^{(K)}\}$ for $i = 1,\ldots,r$. Then, since the step-down procedure rejects, $P'_{k^{(i)}} \leq \alpha$, equivalent to $T_{k^{(i)}} > {}_{(c+1)}t_{\max/W_i}$ for all $i = 1,\ldots,r$. Suppose that the jump-down algorithm does not reject $H_k = H_{k^{(r)}}$. Then, since the test stops short; for some voxel $k^{(i)}$ with $i \leq r$, $T_{k^{(i)}} \leq {}_{(c+1)}t_{\max/W_i}$, in contradiction to the above. Thus, the supposition that the jump down test does not reject H_k must be incorrect. *Reductio ad absurdum*, the assertion is proved.

($\Leftarrow$) Jump-down rejects $H_k \Rightarrow$ step-down rejects H_k (by construction)

Suppose now that H_k is rejected by the jump-down algorithm. Again, let r be the rank of T_k, so that $k = k^{(r)}$. Since $H_{k^{(r)}}$ is rejected, so must $H_{k^{(i)}}$, for $i = 1,\ldots,r$. Therefore, for each i, at the step at which $H_{k^{(i)}}$ is rejected, $T_{k^{(i)}} > {}_{(c+1)}t_{\max/W_j}$, where $W_j = \{k^{(j)},\ldots,k^{(K)}\}$, $j < i$, is the set of voxels under consideration at this step. Clearly $W_i \subseteq W_j$, and therefore ${}_{(c+1)}t_{\max/W_j} > {}_{(c+1)}t_{\max/W_i}$. Hence $T_{k^{(i)}} \leq {}_{(c+1)}t_{\max/W_i}$. But this is

precisely the condition for the step-down test to reject $H_{k^{(i)}}$, given that it proceeds to step i to test it. Since this is true for all $i = 1,\ldots,r$, the step-down test will reject the hypotheses for voxels $\{k^{(1)},\ldots,k^{(r)}\}$. Recall $k = k^{(r)}$.

6.2.4.3. Direct computation of adjusted p-values

A more efficient approach to the step-down test is to accumulate proportions for the adjusted *p*-values as each statistic image in the randomisation distribution is computed. Adapted from Westfall and Young (1993), the algorithm is as follows:

1)	As above
2)	Initialise counting variables $C_i = 0$; $i = 1,\ldots,K$. Set $l=1$
3)	Generate the statistic image ${}_l\boldsymbol{t} = ({}_lt_1,\ldots,{}_lt_K)$ corresponding to randomisation l of the labels.
4)	Form the successive maxima: $v_K = {}_lt_{k^{(K)}}$ $v_{K-1} = \max(v_K, {}_lt_{k^{(K-1)}})$ $(= {}_lt_{\max/W_{K-1}})$ $\vdots \qquad\qquad \vdots$ $v_2 = \max(v_3, {}_lt_{k^{(2)}})$ $(= {}_lt_{\max/W_2})$ $v_1 = \max(v_2, {}_lt_{k^{(1)}})$ $(= {}_lt_{\max/W})$
5)	If $v_i \geq T_{k^{(i)}}$, then increment C_i by one, for each $i = 1,\ldots,K$.
6)	Repeat steps (3)–(5) for each remaining possible labelling $l = 2,\ldots,L$
7)	Compute *p*-values $P'_{k^{(i)}} = C_i / L$
8)	As above (monotonicity enforcement)

Algorithm (d)

6.3. Exemplary Application

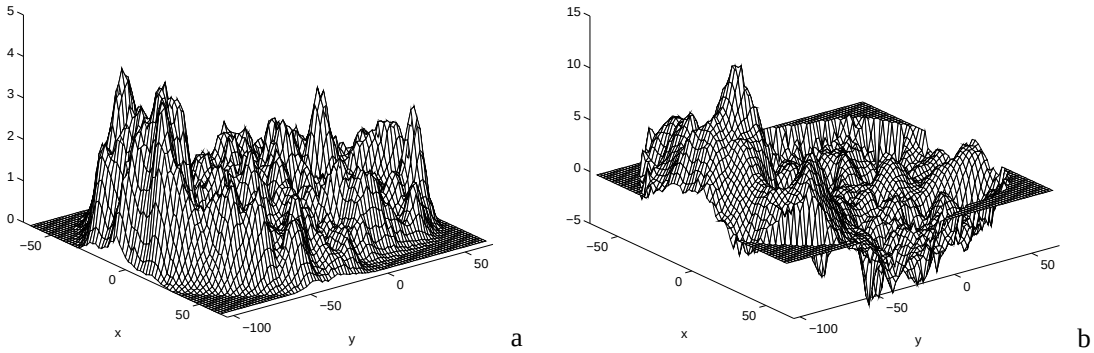
The randomisation tests described above were applied to the "V5" study data, for both the *t*-statistic image, and the "pseudo" *t*-statistic image.

There are $N = 12$ subjects in the "V5" study, giving $L = {}_{12}C_6 = 924$ possible labellings. The whole three-dimensional intracerebral volume was analysed.

6.3.1. Raw *t*-statistic

Statistic images

The study *mean difference image*, $\overline{\Delta}_{\bullet}$, the *sample variance image*, and the paired *t*-statistic image for the "V5" study (for the true labelling) were presented in §2.6.1. The latter two are repeated here for reference (fig.80).



a b

Figure 80
(a) Sample variance and (b) paired *t*-statistic image for "V5" study. These were previously presented in §2.6.1. The AC-PC plane is shown

Randomisation distribution for T_{max}

For each of the possible labellings $l = 1,\ldots,L$ the t-statistic image ${}_l\mathbf{t} = ({}_lt_1,\ldots,{}_lt_K)$ was computed (eqn.71), and the maxima, ${}_lt_{\max/W}$, retained, giving the randomisation distribution of $T_{\max/W}$ (fig.81). The largest of these is the observed maximum t-statistic, $T_{\max/W}$, the maxima of the t-statistic image computed with labellings corresponding to the actual conditions of the experiment. Therefore, a p-value for the omnibus hypothesis H_W is 1/924.

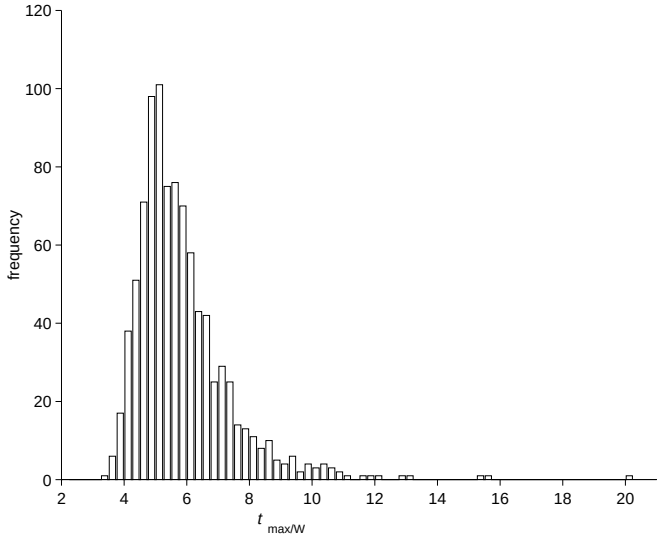


Figure 81
Histogram of randomisation values for the maximum intracerebral t-statistic for the "V5" study.

Single threshold test

For a level $\alpha = 0.05$ test, the appropriate critical value is the $c+1 = \lfloor \alpha L \rfloor + 1 = 47^{\text{th}}$ largest randomisation value of the maximal statistic, ${}_{(47)}t_{\max/W} = 8.6571$ (to 4dp). Values of the observed *t*-statistic image greater than this threshold indicate significant evidence against the corresponding voxel null hypothesis, at the 5% level (fig.82b). The locations of these 632 voxels in Talairach space can be declared as the activation region. The single step adjusted *p*-value image shows the significance of the activation at each voxel (fig.82a).

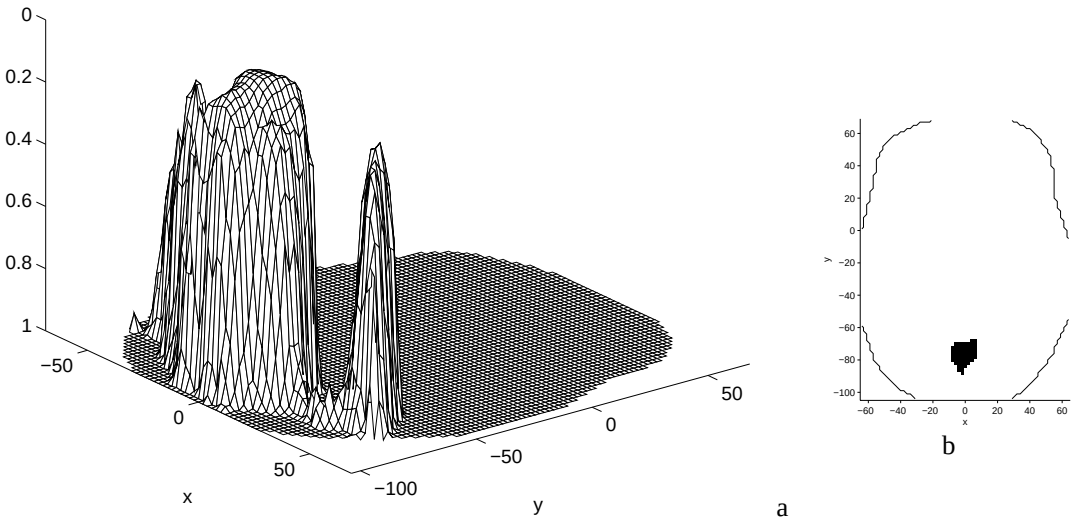


Figure 82

(a) Single-step adjusted *p*-values for the voxel hypotheses H_k, assessed by randomisation test of the *t*-statistic image. (b) Voxels with single-step adjusted *p*-values below level $\alpha = 0.05$, whose observed *t*-statistics exceed the critical threshold of 8.6571. The AC-PC plane is shown, but computations were performed over the whole intracerebral volume. The large activated region at the posterior of the brain corresponds to the primary visual cortex, visual area V1. The two (insignificant) activations either side of V1 are now known to be the motion centre, visual area V5, which this study was designed to locate.

Compare this result with those of the parametric approaches applied to the "V5" *t*-statistic image in §3.6. The single-step adjusted *p*-values for the randomisation test on the raw *t*-statistic image are smaller than those of the Bonferroni approach, Worsley's expected Euler characteristic method for *t*-fields, Worsley's expected Euler characteristic method for Gaussian fields (applied to the Gaussianised *t*-statistic image).

Step-down methods

The step-down test, implemented by the jump-down algorithm (algorithm c), gives a final critical value of 8.6566 (to 4dp) for a level $\alpha = 0.05$ test. This is reached in the second step, and is only a slight reduction over the single step critical threshold. An examination of the re-labellings and maxima shows that the ${}_{(l)}t_{\max/W}$ for $l = 2,\ldots 47$ all lie outside the region rejected by the single step test, and are therefore not excluded by the first step of the jump-down algorithm. No further voxel hypotheses were rejected using this reduced critical threshold. The step-down method gives no improvement over the single-step method.

Step-down adjusted *p*-values were computed using the direct algorithm (algorithm d), and are shown in fig.83a. These were found to differ from the single-step

adjusted *p*-values at only a few pixels, where the step-down *p*-values were 1/924 less than the single step ones (fig.90).

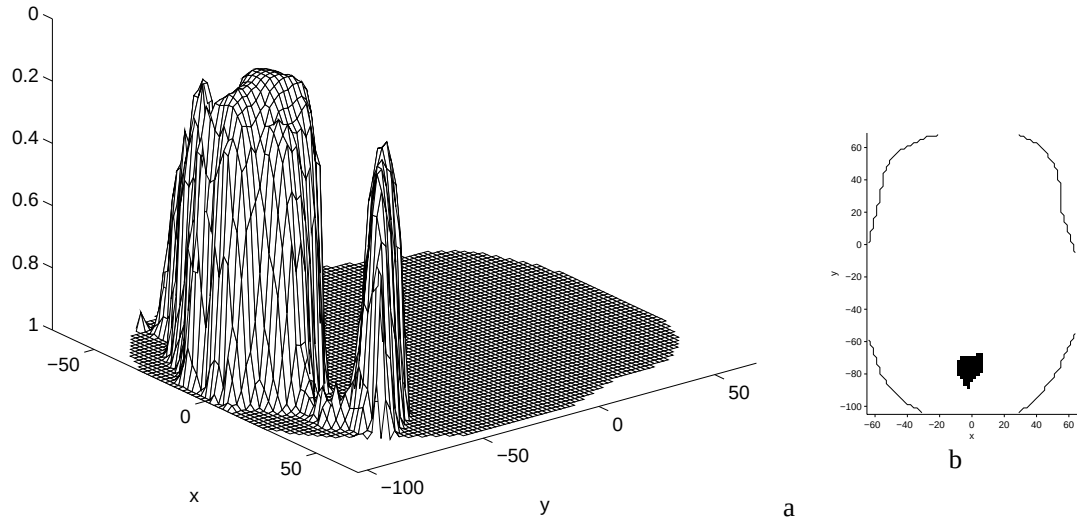


Figure 83

(a) Step-down adjusted *p*-values for the voxel hypotheses H_k, assessed by randomisation test of the *t*-statistic image. (b) Voxels with step-down adjusted *p*-values below level $\alpha = 0.05$, whose observed *t*-statistics exceed the critical threshold of 8.6566. The AC-PC plane is shown.

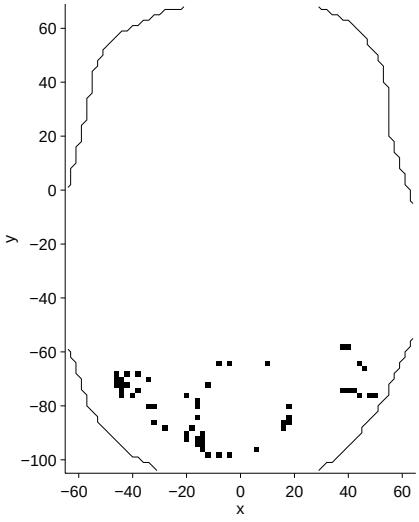


Figure 84

Difference in single-step and step-down adjusted *p*-values. The AC-PC plane is shown, with the edge of the intracerebral area superimposed for orientation. Pixels shaded black had step-down adjusted *p*-values 1/924 less than the corresponding single-step adjusted *p*-value.

6.3.2. Pseudo *t*-statistic image

Statistic images

Pseudo *t*-statistic images were computed with smoothed sample variance images (eqn.69). The Gaussian kernel used was orthogonal, with FWHM of 10mm×10mm×6mm in the X, Y, and Z directions. The variance-covariance matrix Σ of the kernel is therefore (appendix B:4):

$$\Sigma = \begin{pmatrix} 10^2 & 0 & 0 \\ 0 & 10^2 & 0 \\ 0 & 0 & 6^2 \end{pmatrix} \frac{1}{8\ln(2)}$$

The smoothing was implemented as a moving average filter, with weights computed by evaluating the kernel on a regular 17×17×7 lattice of points centred at the origin, separated by 2mm in the X and Y directions, and 4mm in the Z direction, these distances being the distances between voxel centres. This gives the numerator of eqn.69 for all the voxels. The denominator was computed by smoothing a mask image whose value was one for intracerebral voxels, and zero elsewhere.

The smoothed sample variance image for the "V5" study is shown in fig.85, and the pseudo *t*-statistic (eqn.70) computed with this variance in fig.86.

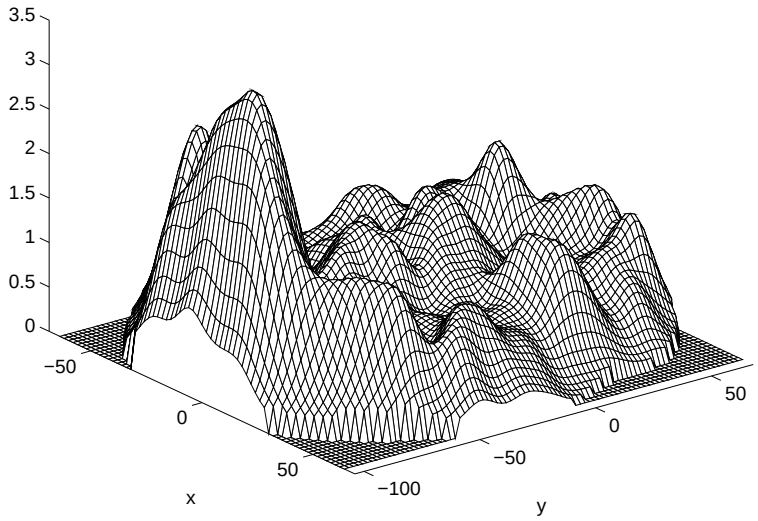


Figure 85
Smoothed sample variance of "V5" study subject difference images (eqn.69). Compare this with the (raw) sample variance image (fig.80).

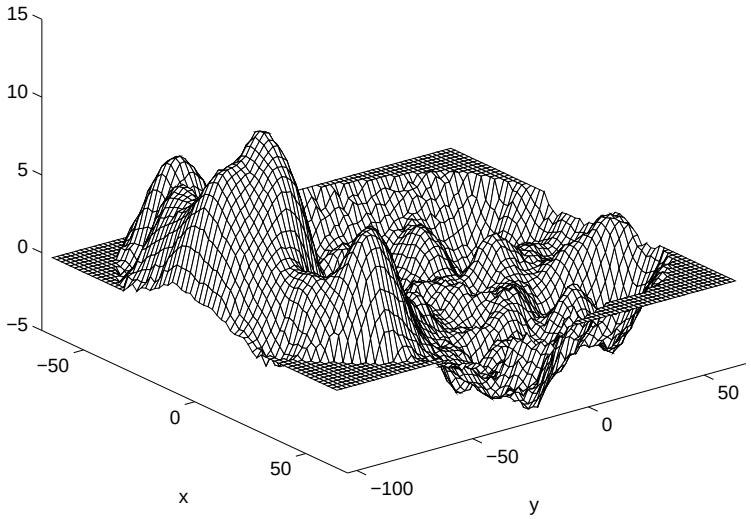


Figure 86
Paired "pseudo" *t*-statistic image for "V5" study. Compare this with the paired *t*-statistic image (fig.80).

Randomisation distribution for T_{max}

Once again, pseudo *t*-statistic images were computed for each possible labelling, and the maxima, ${}_{l}t_{\max/W}$, retained (fig.87). The observed maximum pseudo *t*-statistic, $T_{\max/W}$, is again the largest of the randomisation values for the maximum intracerebral pseudo *t*, giving *p*-value for the omnibus hypothesis H_W of 1/924.

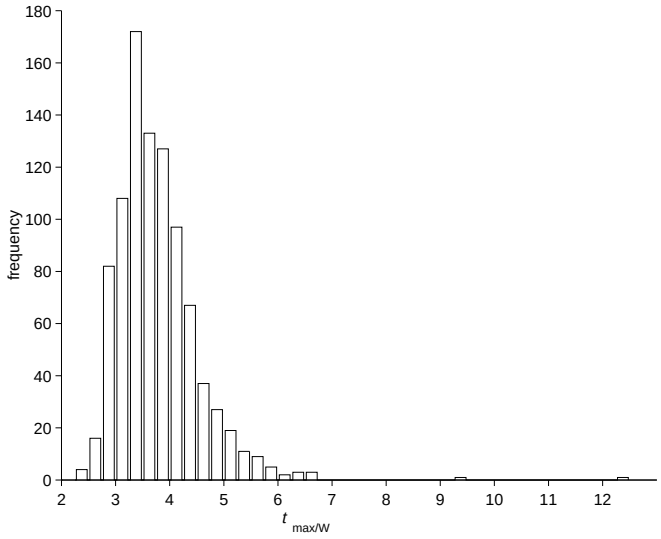


Figure 87
Histogram of randomisation values for the maximum intracerebral pseudo *t*-statistic for the "V5" study.

Single threshold test

The critical threshold for a single threshold test on the pseudo *t*-statistic image at level $\alpha = 0.05$ is the ${}_{(47)}t_{max/W} = 5.0830$ (to 4dp). This is exceeded by the pseudo *t*-values of 2779 voxels (fig.88b). The single step adjusted *p*-value image shows the significance of the activation at each voxel (fig.88a).

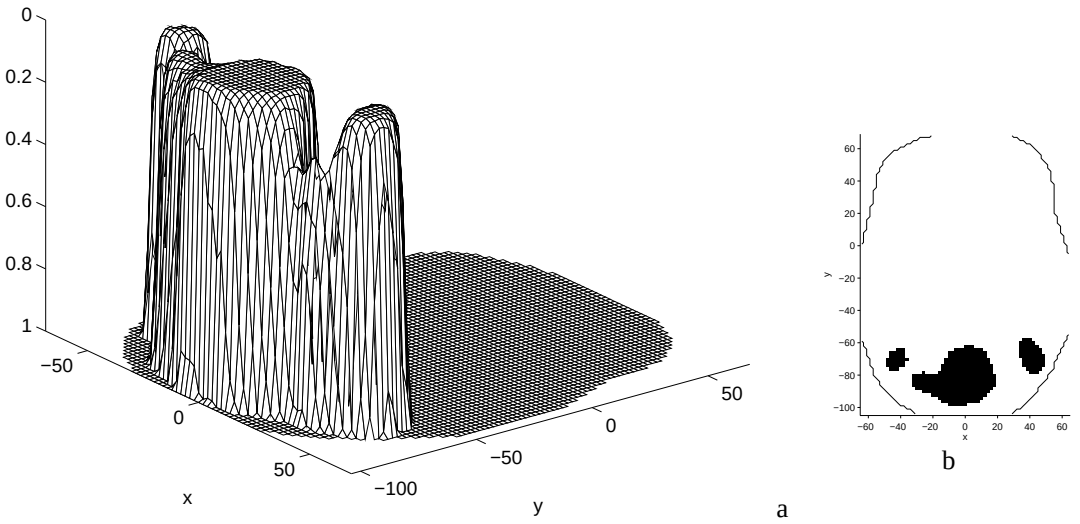


Figure 88

(a) Single-step adjusted *p*-values for the voxel hypotheses H_k, assessed by randomisation test of the pseudo *t*-statistic image. (b) Voxels with single-step adjusted *p*-values below level $\alpha = 0.05$, whose observed *t*-statistics exceed the critical threshold of 5.0830. The AC-PC plane is shown, but computations were performed over the whole intracerebral volume. Compare these figures with the corresponding ones for the "raw" *t*-statistic (fig.82). The two V5 activations either side of V1 are now significant.

Compare this result with those of the parametric methods (3.6.). The single-step adjusted *p*-values for the randomisation test on the pseudo *t*-statistic image are smaller than those of the Bonferroni approach (applied to the raw *t*-statistic image), Worsley's expected Euler characteristic method for *t*-fields (applied to the raw *t*-statistic image), Worsley's expected Euler characteristic method for Gaussian fields (applied to the Gaussianised *t*-statistic image), and Friston's "Bonferroni" approach (applied to the AC-PC plane of the Gaussianised *t*-statistic image). The activated region identified is also larger than that from Friston's suprathreshold cluster size test at threshold $\Phi^{-1}(1-0.0001)$.

Step-down methods

The jump-down algorithm (algorithm c) gives a final critical value of 5.0399 (to 4dp) for a level $\alpha = 0.05$ test. The first step is the simple threshold test, which picks out 2779 activated voxels. The algorithm took three further steps, these picking out an additional 37, 9 and 0 activated voxels. Therefore, thresholding the pseudo *t*-statistic image at this level picks out 2825 voxels as activated, a mere 46 more than the single-step test (fig.89b). The step-down test again provides little improvement.

Step-down adjusted *p*-values were computed using the direct algorithm (algorithm d), and are shown in fig.89a. The reduction in adjusted *p*-value attained by using the step-down test over those from the single-step test are shown in fig.90.

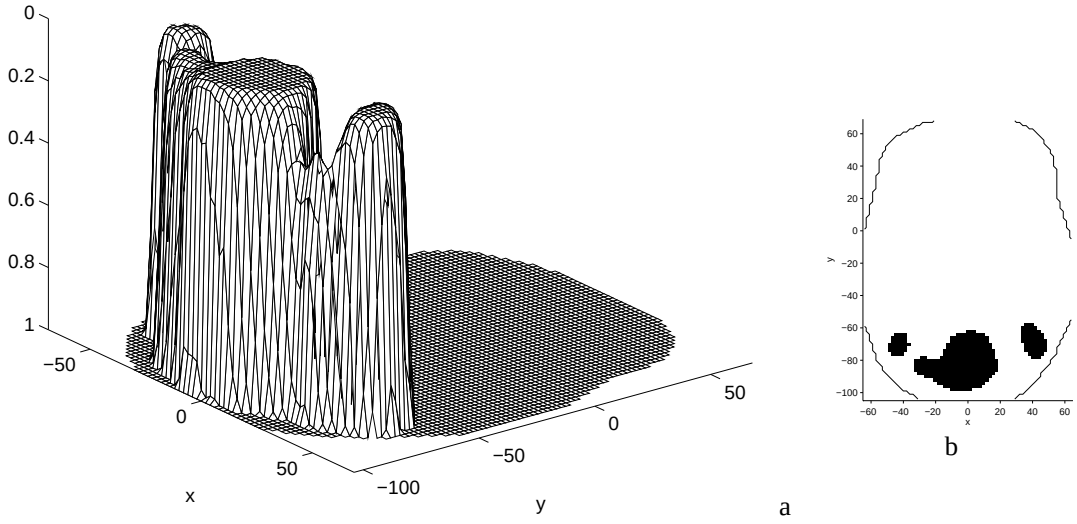
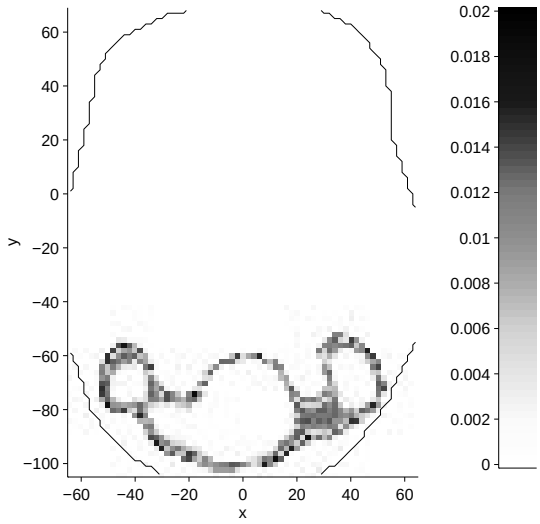


Fig 89

(a) Step-down adjusted *p*-values for the voxel hypotheses H_k, assessed by randomisation test of the pseudo *t*-statistic image. (b) Voxels with step-down adjusted *p*-values below level $\alpha = 0.05$, whose observed *t*-statistics exceed the critical threshold of 8.6566. The AC-PC plane is shown.



Figure

Image showing reduction in adjusted *p*-values attained by using the step-down test over those from the single-step test. The AC-PC plane is shown.

6.4. Discussion

6.4.1. Permutation tests

So far, we have been considering a randomisation test for a particular activation study, where the only assumption is that of an initial random assignment of conditions to subjects. This assumption gives a random element to the experiment, permitting hypothesis testing. It replaces the assumption of parametric analyses that the scans are sampled from some population of known form. However, these non-parametric methods are not limited to experiments where there was an initial randomisation. Permutation tests are computationally the same as randomisation tests, but the justification for re-labelling and computing sampling distributions comes from weak distributional assumptions. For the simple activation study experiment, assume that for each voxel k, the values of the subject mean difference images, Δ_{ik}, are drawn from a symmetric distribution centred at μ_k. Under H_k:$\mu_k = 0$, the subject mean difference image at k is as likely as its negative, corresponding to the opposite labelling of the subjects scans. Thus, under H_W, the 2^N statistic images corresponding to all possible labellings of the subjects scans to AB or BA presentation order are equally likely. However, as six subjects were given the opposite condition presentation order to protect against detecting time effects, only the ${}_N C_{N/2}$ permutations of the labels retaining this balance should be considered (Hochberg and Tamhane, 1987, p267). This gives the same test as the randomisation approach presented above. Indeed many authors neglect the theoretical distinction between randomisation tests and permutation tests and refer to both as permutation tests. The advantage of the randomisation test over the permutation test, is that the random allocation assumption of the former is clearly true, yielding a valid test.

6.4.2. Other applications

Clearly these non-parametric methods can be applied to many paradigms in functional neuroimaging. For statistic images where large values indicate evidence against the null hypothesis, we have developed the theory for a single threshold test, and a step-down extension. All that is required is the concept of a label for the data, on the basis of which a statistic image can be computed, and a null hypothesis specified. For instance, region of interest analyses can easily be accommodated, as can parallel group designs, comparing a control group with a disease group. Indeed, a randomisation alternative is available for most experimental designs. See Edgington (1980) for a thorough exposition. Consider the following examples:

Single Subject Correlation

It is wished to locate the regions of an individual's brain in which activity increases linearly with the difficulty of a certain task. During each of M scans the subject is given the task at a set level of difficulty, the order of presentation having been randomly chosen from the $M!$ possible allocations. The evidence for linear association of rCBF and difficulty level at a given voxel may be summarised by the correlation of difficulty level and rCBF (after some form of global flow correction), or by any suitable statistic. The "labels" here are the actual task difficulty levels. Under the omnibus null hypothesis H_W:[the individual would have had the same rCBF whatever the task difficulty], the statistic images corresponding to all $M!$ possible randomisations of the task difficulties are equally likely, and a randomisation test follows as above.

Single subject activation experiment

It is wished to locate the region of an individual's brain that is activated by a given task over a given baseline condition. The subject is scanned an even number of times, with each successive pair of scans consisting of one under each condition, giving M successive rest-activation pairs. The order of the conditions within each pair of scans is randomly chosen.

A statistic image is formed whose large values indicate evidence of an increase in rCBF between the baseline and activation condition. A suitable candidate would be a paired pseudo t-statistic. Under H_W:[each scan would have given the same rCBF image had it been acquired under the other condition], statistic images formed by permuting the labels within individual pairs of scans, are as likely as the observed one. The labels here are the scan condition, "baseline" or "activation", and the possible labellings are those for which each successive pair of scans are labelled AB or BA. There are M such pairs, so the randomisation distribution of the statistic image will consist of 2^M images.

6.4.3. Other statistics

As we have seen, these non-parametric tests can be applied to any form of voxel statistic image, unlike parametric analyses where the choice of statistic is limited to those of known distributional form (for suitable assumptions on the data). This enables the consideration of the pseudo t-statistic image formed with a smoothed variance estimate, for which no distributional results are available. The researcher has a free hand to invent experimental designs and devise statistics. However it is advisable to use standard experimental designs and statistics if these exist, possibly modifying the statistic to take advantage of its setting in an image, for example by smoothing. Standard test statistics are usually those that give the most powerful test in the parametric setting, when the assumptions are true, and can therefore be expected to retain their sensitivity when the assumptions are slightly in doubt.

The application of these randomisation and permutation tests is also not limited to the single threshold and step-down tests discussed above. The procedure gives the randomisation distribution of the whole statistic image, and hence of any statistic summarising a statistic image. By computing the exceedence proportion for each statistic image in the randomisation distribution, a non-parametric version of the omnibus tests described in §3.4.1. & §3.4.2. could be obtained.

Non-parametric suprathreshold cluster tests

Suprathreshold cluster tests were discussed in §3.5. Recall that the statistic image is thresholded at a predetermined level, clusters of voxels with suprathreshold values identified, and the significance of these clusters assessed using some statistic summarising the statistic image over the cluster, for instance its size:

Randomisation values for the maximum suprathreshold cluster size can be obtained by computing the maximum suprathreshold cluster size for each of the statistic images in the randomisation distribution. These randomisation values can then be used as those for the maximal voxel statistic were in the single-step methods, to obtain the critical cluster size. A cluster in the observed statistic image larger than the critical cluster size indicates significant evidence against the (omnibus) null hypothesis for the voxels in that cluster. Single-step adjusted p-values for the cluster omnibus hypotheses can be computed. Strong control follows on a cluster by cluster basis. The step-down method can also be used, where each step tests the omnibus hypothesis over a cluster of voxels identified from the actual statistic image. If a cluster is found to be significant, then all the voxels in

the cluster are omitted in the next step. Thus, non-parametric suprathreshold cluster size tests are possible.

6.4.4. Number of possible labellings, size and power.

A key issue for randomisation and permutation tests is the number of possible labellings (L), as this dictates the number of observations from which the randomisation distribution is formed. For the possibility of rejecting a null hypothesis at the 0.05 level there must be at least 20 possible labellings, in which case the observed labelling must give the most extreme statistic for significance. Since the observed statistic is always one of the randomisation values, the smallest p-value that can be obtained from these methods is $1/L$. This is shown in the adjusted p-value images for the pseudo t-statistic (figs.88 & 89), where the smallest possible p-value of 1/924 is attained for most of the activated region, even though the observed pseudo t-statistic image has a clear maximum (fig.86). To demonstrate stronger evidence against the null hypothesis, via smaller p-values, larger numbers of permutations are required. As the possible labellings are determined by the design of the experiment, this limits the application of non-parametric tests to designs with moderate subject and/or scan per subject numbers. The single subject activation experiment described above has only 2^M possible labellings, 64 for a twelve scan session, or 128 for a 14 scan session. Clearly the greater L, the greater the power, as this implies more data.

As L tends to infinity, the randomisation distribution tends to the sampling distribution of the test statistic for a parametric analysis, under suitable assumptions (Hoeffding 1952). Thus, if all the assumptions of a parametric analysis are correct, the randomisation distribution computed from a large number of re-labellings is close to the hypothetical sampling distribution of the test statistic, were data randomly sampled from the appropriate model. For smaller L, non-parametric methods are, in general, not as powerful as parametric methods when the assumptions of the latter are true. In a sense, assumptions provide extra information to the parametric tests giving them the edge. Comparison of parametric and non-parametric analyses of the "V5" data presented here, would suggest that this discrepancy is not too great. The attraction of the non-parametric methods is that they give valid tests when assumptions are dubious, when distributional results are not available, or when only approximate theory is available for a parametric test.

6.4.5. Approximate tests

In many cases the number of possible labellings is very large, and the computation involved makes generating all the randomisation values impractical. For instance the single subject correlation study described above has $M!$ possible labellings, 479 001 600 for a twelve scan experiment. Here, approximate tests can be used (Edgington, 1969b). Rather than compute all the randomisation values, an approximate randomisation distribution is computed using randomisation values corresponding to a subset of the possible labellings. The subset consists of the true labelling, and L'-1 labellings randomly chosen from the set of possible labellings (usually without replacement). The tests then proceed as before, using this subset of the randomisation values. A common choice for L' is 1000, so the randomisation values are those formed with the observed labelling and 999 re-labellings randomly chosen from the set of those possible.

Despite the name, the approximate tests are still (almost) exact. Only the randomisation distribution is approximate. The randomisation values used can be thought

of as a random sample of size L' taken from the full set of randomisation values, one of which corresponds to the observed labelling. Under the null hypothesis, all members of the subset of randomisation values are equally likely, so the theory develops as before. As the critical threshold is estimated, it has some variability about the true value that would be obtained from the full sampling distribution. There is a loss of power because of this, but the loss is very small for approximate sampling distributions of size 1000. See Edgington (1969a & 1969b) for a full discussion.

6.4.6. Step-down tests

Step-down tests ineffectual

The possibility of a large activation causing these tests to be conservative has been discussed, and the step-down tests proposed to counter the problem. From the results presented here, the step-down methods appear not to merit the computation they involve. Examination of the randomisation *t*-statistic images that yield the largest maximum values, shows that many correspond to labellings that are almost the opposite of the true labelling of the experiment. For this data set it appears that the negatively activated background (giving large statistics for labellings almost the opposite of the true labelling) influences the extreme of the randomisation distribution more than the positively activated region does in labellings close to the true labelling. These negative activations are not "cut out" by a step-down procedure.

Two-sided step-down tests

One solution to this problem would be to consider a two-sided step-down test. However, if only a one-sided test is required, the two-sided approach may not present an improvement over the one-sided step-down method, since the test level is effectively halved. This appears to be the case for the "V5" data. A two-sided single-step randomisation test at level $\alpha = 0.05$, using the pseudo *t*-statistic, gives a critical threshold of 5.3502 (4dp), which 6560 voxels exceed in absolute value. Omitting these voxels from the search set, the second step of the jump-down algorithm finds no evidence against additional voxel hypotheses, and gives a final step-down threshold of 5.3501 (4dp). Recall that the one-sided step-down randomisation test, using the pseudo *t*-statistic, gave critical value 5.0399 (4dp) for level $\alpha = 0.05$. A two-sided test using the "raw" *t*-statistic image behaves similarly.

In conclusion, it appears that the step-down methods are not worth the computational effort involved.

Artefactual deactivations

There remains the issue of whether a deactivation is real, or an artefact of global normalisation. This was discussed previously in §2.4., where it was concluded that for the "V5" study, g_{CBF} was not condition dependent, and that the depressed background of the "V5" statistic images represents a true inhibition of r_{CBF} induced by the large increase in the visual cortex.

If decreases are artefactual, then it may be worth investigating measures of g_{CBF} more sophisticated than the mean voxel value, measures which effectively measure the background g_{CBF}. Another possibility would be to use the jump-down method and re-normalise the data after every step, omitting the "activated" regions from the computation of g_{CBF} (g_A). This latter approach requires further investigation, since it is not clear whether strong control over FWE is maintained.

6.4.7. Computational burden

The main drawback of the non-parametric tests presented here are the vast amounts of computation involved. The current work was undertaken in MATLAB (The MathWorks Inc., Natick), a matrix manipulation package with extensive high level programming features. The test was implemented as a suite of functions. The platform used was a SUN Microsystems SPARC2 workstation, with 48MB of RAM and 160MB of virtual memory. Computing times for the "V5" data are presented in the table below. More efficient coding in a lower level compiled language, such as C, should greatly reduce the running times. Even so, considering the vast amounts of time and money spent on a typical functional mapping experiment, a day of computer time seems a small price to pay for an analysis whose validity is guaranteed.

Computing times

	"raw" *t*-statistic	pseudo *t*-statistic
Single-step (randomisation values only)	8 hours	14 hours
Jump-down (algorithm c)	16 hours (two steps @ 8 hours)	56 hours (four steps @ 14 hours)
Step-down *p*-value (algorithm d)	18 hours	24 hours

Table 91

Computing times for the randomisation tests on the "V5". All times include time spent computing the randomisation *t*-statistic images from the 12 subject difference images. Each subject difference image consists of $K = 77189$ intracerebral voxel values, held in double precision. The computational "tricks" of considering only half the labellings, and of storing the sum of squares of the subject difference images, were utilised in the code (recall §6.2.2.). Computations were undertaken in MATLAB on a SUN Microsystems SPARC2 workstation.

6.5. Conclusions

In this chapter a new method has been presented for the analysis of functional mapping experiments. Established non-parametric techniques have been extended to this special multiple comparisons problem, to produce a test with localising power, that is (almost) exact, relatively simple, and which has numerous advantages over existing parametric methods.

These non-parametric tests are valid and (almost) exact given minimal assumptions on the mechanisms creating the data, in contrast to existing parametric analyses, which rely on approximations and strong assumptions. The randomisation test assumes only an initial random allocation in the design of the experiment, an assumption that can clearly be verified. Permutation tests assume vague properties about the distribution of the data, such as symmetry about some location. Because the assumptions are true, and no approximations are made, there is no need to assess specificity using simulated data or rest-rest comparisons. That the tests force practitioners to think carefully about randomisation and experimental design is no bad thing.

The tests are also very flexible. They can be applied to any paradigm where there is a concept of a label for the data, from which a statistic can be formed, and a null hypothesis specified. As the distribution of the statistic image is not required to be known, images of non-standard test statistics can be analysed, for example the pseudo *t*-statistic computed with smoothed variance considered here. Further, any sensible statistic summarising a statistic image may be used to assess evidence of a signal. For example, the maximum voxel value or maximum suprathreshold cluster size may be taken as statistics for summarising statistic images, leading to exact non-parametric single threshold and suprathreshold cluster size tests respectively.

The disadvantages of the method are the computation involved, and the need for experiments with enough replications or subjects to give a workable number of possible labellings.

The power of these methods remains to be examined thoroughly. In general, non-parametric methods are outperformed by parametric methods when the assumptions of the latter are true. For fairly large experiments (with many possible labellings) the discrepancy may not be great, particularly if the parametric methods are conservative. Present experience suggests that the non-parametric methods and the current parametric methods give similar results for studies where the assumptions of the latter are reasonable. This could be examined (at great computational expense) by simulation. However, there are many situations where the assumptions of current parametric methods are in doubt. In these situations the non-parametric methods provide the only valid method of analysis. In addition, the ability to analyse non-standard statistic images, such as the pseudo *t*-statistic considered here, appears to afford the non-parametric tests additional power over the parametric methods, which are constrained to statistics for which distributional results are available. Experience of applying both methods to a wide range of data sets would be valuable.