Appendix C. Interpretation of α and p in SPM analysis

In univariate testing, interpretations of α and p are tightly coupled. Consider a single two-sample t test at α=0.05 which yields p=0.04. The null hypothesis would be rejected because p is less than α.

This α-p coupling, we argue, is an illusion. In univariate testing α and p appear to be coupled only because univariate data are 0D, and more precisely because supra-threshold univariate data have 0D geometry. When generalizing hypothesis testing to nD continuum data, this coupling is destroyed because suprathreshold data have nD geometry.

Consider the computational process for a univariate two-sample t test:

1. Set α.
2. Estimate group means and standard deviations.
3. Compute the test statistic (t) value.
4. Compute the critical value t^* above which only α% of repeated random samplings of the two groups would be expected to traverse, if there were truly no mean difference between those groups.
5. If $t > t^*$, then reject the null hypothesis.
6. Compute the precise probability p that the observed t value would be observed in repeated random samplings, if there were truly no mean difference between those groups.

For formal hypothesis testing, the only step that matters is #5 because the precise p value computed in #6 is irrelevant to the binary hypothesis rejection decision. A skeptic may argue that an alternative and equivalent computational process is: delete #4 and #5, and then rewrite the null hypothesis rejection decision following #6 as: "$p < \alpha$". The skeptic's process tightly couples p and α, and we acknowledge that the two processes are apparently equivalent for univariate data. However, we'd also suggest that the skeptic's process is fundamentally flawed because it cannot be generalized to arbitrary (i.e. nD) data.

To clarify, consider the SPM process for conducting a two-sample t test on 1D (scalar trajectory) data:

1. Set α.
2. Estimate group mean trajectories and standard deviation trajectories.
3. Compute the test statistic (t) trajectory.
4. Compute the critical value t^* above which only $\alpha\%$ of repeated random samplings of the two groups' trajectories would be expected to traverse (anywhere in the trajectory), if there were truly no mean trajectory difference between those groups.
5. If $t > t^*$ (anywhere in the trajectory), then reject the null hypothesis.
6. For each suprathreshold cluster, compute the precise probability p that a suprathreshold cluster of the given size would be observed in repeated random samplings, if there were truly no mean difference between those groups' trajectories.

The skeptic could not describe an equivalent hypothesis testing procedure based on p, because p values are not computable until the t trajectory is thresholded at t^*. In other words, thresholding a smooth nD continuum generally produces multiple suprathreshold clusters, and each cluster has arbitrary nD geometry. Most importantly, a suprathreshold cluster's geometry is irrelevant to the binary hypothesis rejection decision; that decision is driven simply by the presence/absence of suprathreshold clusters.

Back in the world of univariate testing, it is easy to understand that suprathreshold geometry is collapsed into 0D space (i.e. into a single scalar). In 0D space there can be neither multiple suprathreshold clusters nor suprathreshold clusters with arbitrary geometry. This 0D geometry produces the illusion that α, t^* and p are all tightly bound together, but in the more general sense: α determines t^*, and p can only be computed from suprathreshold nD geometry.

The perspective that $p < \alpha$ drives hypothesis testing is therefore not a generalizable one. P values are more accurately described as descriptive statistics which compliment the binary hypothesis testing decision.

Noting that 'tensors' generalize scalars and vectors, we offer the following generalized definitions of α and p:

α is the scalar that determines the critical test statistic value, above which only $\alpha\%$ of experimentally observed nD test statistic continua would traverse if the null hypothesis were true, and if the underlying data were generated by a Gaussian process with the observed nD tensor smoothness and covariance.

Each cluster-specific p value represents the likelihood that repeated random samplings from the same population(s) would produce an nD suprathreshold cluster as large[4] as the observed suprathreshold cluster, if the null hypothesis were true, and if the underlying data were generated by a Gaussian process with the observed tensor smoothness and variance.

[4]We note that "large" is imprecisely defined: it could refer to cluster breadth across the continuum, or it could refer to some integral of breadth and height, or to some other geometrical feature. Most commonly SPM-based inference is based on breadth, but cluster integral-based inference procedures have also been analytically derived in the literature. The latter are much less common because their additional computational complexity provides little or no practical benefit to the binary hypothesis testing decision.