Notes on permutation test

A permutation test provides a quantitative way of assessing the statistical significance of difference between groups. Take the following example where two groups, A and B (as shown on the left panel in Figure 1), are observed. Suppose that each of these groups includes two individuals (represented by identical colors) and we wish to know if the average height of individuals in group B is greater than the average height of individuals in group A. Let's call this difference in average height *Dobs* for observed difference. Suppose that we observed *Dobs > 0*, in which case it suggests that on average, group A is taller than group B. However, it is conceivable that this difference is simply due to chance, i.e. random fluctuations or variations in people's heights can generate a group difference that is merely a noisy observation, hence drawing inference purely based on the magnitude of difference could lead to the wrong conclusion. In principle, we would want to assess the extent to which this difference is “real,” or *statistically significant*.

Permutation test provides one way of getting at this question, and its basic idea is illustrated partially on the right panel in Figure 1. Here we construct a chance level, a.k.a. a null distribution, by permuting the data (or individuals in this case) across the two groups. The idea is that if the observed individuals in group B are really systematically higher than those in group A, we would expect that such a difference would be appreciably smaller for a random permutation of these individuals (as their group identities are distorted). It turns out that with 4 data points and 2 groups, we can enumerate all possible permutations exhaustively – the six possible permuted arrangements are shown on the right panel. Note that as the sample size grows, permuting at all possibilities entails a much larger space and computationally it becomes infeasible to carry out the test exhaustively. A typical solution is to permute many times, e.g. 10000, so at least we explore a fair amount of the space to get a good sense of the null, or chance distribution.

For each permuted set of group data, we can then compute the difference between groups just as we did with the observed grouping (with no permutation). Let's refer to these permuted differences as *Dperm* = {*Dperm1*, *Dperm2*, *…..*,*Dperm10000}*, assuming there are 10000 permutations. The statistical significance of *Dobs* can then be calculated via this equation:

p = sum (number of items in *Dperm* >= *Dobs*)

Total number of permutations

where “p” is typically referred to as the p value which is an indicator of significance. The equation above says that the degree of significance is determined by the fraction of values in *Dperm* that appears to be greater than or equal to the observed value *Dobs*. P-value is always between 0 and 1 (do you see why from the equation above?), and in general, a small p-value (e.g. p<0.05) is taken as evidence for establishing statistical significance that the null can be rejected, hence as an alternative way of saying that the observed difference is more extreme than chance.

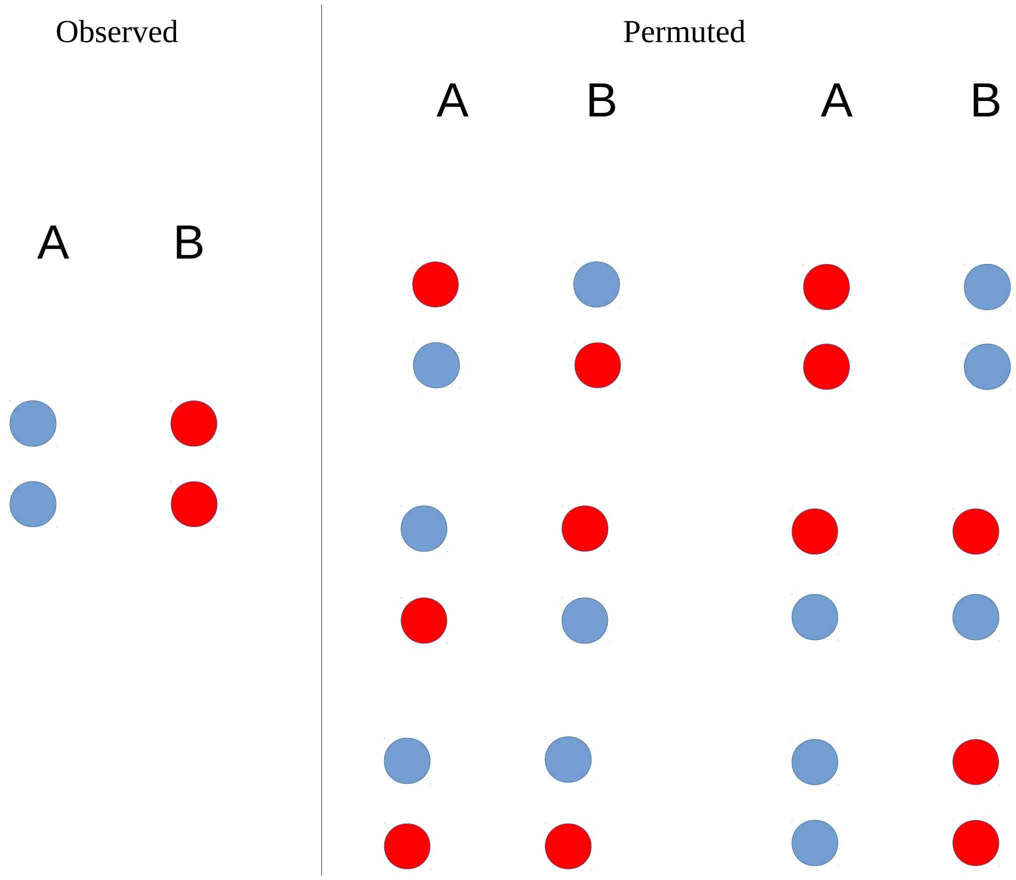

Figure 1: Illustration of the permutation procedure.