K-sample methods

The permutation F test

The idea of the permutation test can be extended to many other situations, such as the K-sample problem. With observations randomly selected from k populations with cdfs $F_1(x), F_2(x), ..., F_k(x)$, the null hypothesis of interest is of equality of distributions,

$$H_0 : F_1(x) = F_2(x) = ... = F_k(x),$$

with an alternative of

$$H_a : F_i(x) \leq F_j(x) \text{ or } F_i(x) \geq F_j(x),$$

where at least one pair (i, j) has strict inequality holding for at least one x value. A special case of interest is for a shift alternative, so that

$$H_a : F_i(x) = F(x - \mu_i), \text{ or equivalently, } X_{ij} = \mu_i + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \sim i.i.d.\ F \quad .$$

For the usual analysis of variance (ANOVA) notation, refer to Table 3.1.1. Recall that in k-sample (or one-way) ANOVA, we partition the total sum of squares into two components,

$$SS_{total} = SST + SSE,$$

where

$$SST = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2, \text{ and } SSE = \sum_{i=1}^{k}(n_i - 1)S_i^2.$$

These sums of squares are divided by their degrees of freedom ($k - 1$ and $N - k$) to obtain mean squares MST and MSE, then the ratio is an F statistic: $F = MST/MSE$. Under the usual ANOVA assumptions (independence, normality, equal variances), the F statistic follows an F distribution with $k - 1$ and $N - k$ degrees of freedom.

However, we could instead use the random allocation of treatments to subjects to justify doing a permutation F test, as shown on section 3.1.2. It also turns out that the F distribution used for parametric ANOVA can be shown to give a good large sample approximation to the p value from the permutation F test. Also, similar to a result for two-sample tests, the F statistic can be rewritten

$$F = \frac{SST/(k-1)}{(C - SST)/(N-k)},$$

which is an increasing function of SST, so that the permutation F test can be based on SST or just SSX, a weighted sum of squared sample means.

The Kruskal-Wallis statistic

As we found in the two-sample situation, we can avoid the effect of outliers by using the ranks of the observations in the permutation test. If we use R_{ij} to denote the rank of observation X_{ij}, and denote the mean of the ranks in the i^{th} group as $\overline{R}_i$, since the overall mean of the ranks is $(N+1)/2$, then

$$\sum_{i=1}^{k} n_i(\overline{R}_i - \frac{N+1}{2})^2$$

is the treatment sum of squares applied to the ranked data. The Kruskal-Wallis statistic is just this statistic multiplied by a constant,

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i(\overline{R}_i - \frac{N+1}{2})^2 \ .$$

The constant allows the use of a large-sample chi-square approximation (for data without ties) to the permutation p value. Thus, we can either obtain a permutation p value or use the chi-square approximation which is generally conservative in small samples.