Large Sample Approximations to Two-Sample Test Statistics

We can compute a large-sample approximation to the permutation distribution for any statistic that can be computed as a sum of scores associated with one of two treatments. A general such expression is:

$$T_1 = \sum a(R(X_i))I[X_i \text{ in group } 1] = \sum A_i I[X_i \text{ in group } 1],$$

where m and n are the two group sample sizes (with $N = m + n$) and $A_i, i = 1, 2, ..., N$ are the general scores for the observations.

Under the null hypothesis of no difference between treatments, each score A_i is as likely to occur in treatment 1 as any other score. Thus, the m scores associated with treatment 1 occur as if they had been randomly selected without replacement from the set of combined scores from both groups. Thus we can calculate the expected value and variance of T_1, the sum of the scores from treatment 1, and use a normal approximation. From the theory of sampling from finite populations you can then show that:

$$E(T_1) = m\mu = m\frac{\sum_{i=1}^{N} A_i}{N}, \text{ and}$$

$$\text{Var}(T_1) = \frac{mn}{N-1}\sigma^2 = \frac{mn}{N-1}\frac{\sum_{i=1}^{N}(A_i-\mu)^2}{N} = \frac{mn}{N-1}\left(\frac{\sum_{i=1}^{N} A_i^2}{N} - \mu^2\right).$$

For m and n sufficiently large, T_1 is normally distributed, so

$$\frac{T_1 - E(T_1)}{\sqrt{\text{Var}(T_1)}} \sim Z.$$

Use with the Wilcoxon Rank-Sum Test

For the Wilcoxon test, the A_i values are just the ranks $1, 2, ..., N$. It is shown in the text that $\mu = (N+1)/2$ and $\sigma^2 = (N-1)(N+1)/12$ in this case, so in the case without ties, the expected value and variance of the rank-sum statistic W are:

$$E(W) = \frac{m(N+1)}{2} \text{ and } \mathrm{Var}(W) = \frac{mn(N+1)}{12}.$$

If there are ties in the data, then $E(W)$ is unaffected, but $\mathrm{Var}(W)$ must be adjusted downward. In this case, $\mathrm{Var}(W)$ can be directly calculated (shown on page 68), or can be calculated via the formula:

$$\mathrm{Var}(W) = \frac{mn(N+1)}{12} - AF = \frac{mn(N+1)}{12} - \frac{mn\sum(t_i^3 - t_i)}{12N(N-1)},$$

where the t_i is the number of values tied in the ith tied group of values.

Use with a confidence interval based on the Mann-Whitney test

From our earlier expression relating an interval for Δ to the U distribution:

$$P(\mathrm{pwd}(k_a) < \Delta \le \mathrm{pwd}(k_b)) = P(k_a \le U \le k_b - 1)$$

we can apply a normal approximation:

$$P(k_a \le U \le k_b - 1) = P(\frac{k_a - E(U)}{\sqrt{\mathrm{Var}(U)}} \le Z \le \frac{k_b - 1 - E(U)}{\sqrt{\mathrm{Var}(U)}}),$$

and then solve to find the desired order statistic values k_a and k_b via:

$$k_a \approx E(U) - z_{(1-\alpha/2)}\sqrt{\mathrm{Var}(U)} \text{ and } k_b \approx 1 + E(U) + z_{(1-\alpha/2)}\sqrt{\mathrm{Var}(U)}.$$

Use with the permutation distribution of T_1 with the raw data

The text discusses obtaining a normal approximation for T_1 for the raw data, but this approximation would generally require a larger sample size than the ones above to be accurate.