Computational Details for Choplump test: Test Motivated in "Choplump Tests for Vaccine Trials"

by Dean A. Follmann, Michael P. Fay, and Michael A. Proschan

September 4, 2007

1 General Chop-Lump Test

Suppose n_0 and n_1 subjects are randomized to control and vaccine respectively. Here we allow $n_0 \neq n_1$, which causes some notational complexity, although the chopping function simply removes zeros from both groups in approximately the same proportion within each group such that one group has no zeros.

Let m_0 and m_1 denote the number of positive responses in the control and vaccine group repectively. Let $k_i = n_i - m_i$, $N = n_0 + n_1$, $M = m_0 + m_1$, and $K = k_0 + k_1$. Let the responses be represented by the vector, $\mathbf{W} = [W_1, W_2, \ldots, W_N]$, where $\mathbf{W}$ are the responses of all N subjects and K of those responses are 0. Let the treatment randomization assignment be denoted by the vector, $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$, where $Z_i = 0$ for subjects randomized to control and $Z_i = 1$ for subjects randomized to vaccine. We order the indices by W_i first then by Z_i within tied W_i values, so that $Z_1, \ldots, Z_{k_0}$ are zeros and $Z_{k_0+1}, \ldots, Z_K$ are ones. Let $\mathbf{W}_a$ and $\mathbf{Z}_a$ be the last a values of $\mathbf{W}$ and $\mathbf{Z}$, respectively. Let $\mathbf{0}_a$ and $\mathbf{1}_a$ be vectors of zero or one of length a, where $a = 0$ denotes no vector (e.g., $[\mathbf{0}_3, \mathbf{1}_0]$ is a 3×1 vector of 0's). Let $C(\mathbf{W}, \mathbf{Z})$ be the chopping function which creates the "chopped" data set, specifically,

$$C(\mathbf{W}, \mathbf{Z}) = (\mathbf{W}_{M+a+b}, [\mathbf{0}_a, \mathbf{1}_b, \mathbf{Z}_M]),$$

where

$$\begin{array}{lll} \text{if } \frac{m_0}{n_0} \geq \frac{m_1}{n_1} & & \text{then } a = 0 \text{ and } b = k_1 - \lfloor \frac{n_1 k_0}{n_0} \rfloor \\ & \text{and} & \\ \text{if } \frac{m_0}{n_0} < \frac{m_1}{n_1} & & \text{then } a = k_0 - \lfloor \frac{n_0 k_1}{n_1} \rfloor \text{ and } b = 0 \end{array}$$

and $\lfloor x \rfloor$ is the largest integer less than or equal to x.

In the usual permutation test, we define a test statistic T which is a function of $\mathbf{W}$ and $\mathbf{Z}$. Let T_0 be the test statistic evaluated at the original data, and T_j be the test statistic evaluated at the jth permutation of the values of $\mathbf{Z}$. If lower values of the test statistic are more extreme, then a one-sided p-value is

$$p - value = \frac{\sum_{j=1}^{N!} I\{T_j \leq T_0\}}{N!} \tag{1}$$

where $I(a) = 1$ if a is true and 0 otherwise. A chop-lump test is simply a permutation test where the test statistic is of the form, $T_{CL}(\mathbf{W}, \mathbf{Z}) = T\{C(\mathbf{W}, \mathbf{Z})\}$.

2 Computational Issues: Exact Tests

In this section, we describe exact computation for any two-sample permutation test. There are computationally better ways to calculate the p-value than equation 1. First, we we need not enumerate all $N!$ permutations of $\mathbf{Z}$, since there are only $\begin{pmatrix} N \\ n_1 \end{pmatrix}$ unique permutations of $\mathbf{Z}$, and each has exactly $n_0!n_1!$ permutations which correspond to the same permuted $\mathbf{Z}$. We can obtain similar computational savings by partitioning the $\begin{pmatrix} N \\ n_1 \end{pmatrix}$ unique permutations into sets with equal numbers of zero responses in the vaccine

group. One can think of this partition as being derived from the hypergeometric distribution where we are sampling zeros in the vaccine group. The partition can be written as

$$\begin{pmatrix} N \\ n_1 \end{pmatrix} = \sum_{h=\max(0,n_1-M)}^{\min(n_1,K)} \begin{pmatrix} K \\ h \end{pmatrix} \begin{pmatrix} M \\ n_1 - h \end{pmatrix} \tag{2}$$

On the right-hand-side of equation 2 the first term in the sum represents the number of ways to permute the indices of the zero responses, while the second term represents the number of ways to permute the nonzero responses. Let Q_h be the proportion of the permutation test statistics less than or equal to the observed test statistic among permutations with h zeros in the vaccine group. Specifically,

$$Q_h \quad = \quad \frac{\sum_{j \in \Omega_h} I\left[T_j \leq T_0\right]}{\begin{pmatrix} M \\ n_1 - h \end{pmatrix}} \tag{3}$$

where Ω_h is the set of unique permutations of $\mathbf{Z}_M$ that induce h zeros in the vaccine group. In other words, Ω_h does not include two different permutations of $\mathbf{Z}$ if they only differ within the first $K = N - M$ elements, since those elements are all equal to zero.

The standard calculation groups the $N!$ permutations into $\begin{pmatrix} N \\ n_1 \end{pmatrix}$ sets of unique permutations of $\mathbf{Z}$, and each set has the same number of members. In the case of equation 2, each group with h zeros in the vaccine group does not have the same number of members. The one-sided p-value is a weighted average of the Q_h values:

$$\begin{aligned} p - value \quad &= \quad \sum_{h=\max(0,n_1-M)}^{\min(n_1,K)} Pr[\text{ a permutation has } h \text{ zeros in the vaccine group}] Q_h \\ &= \quad \sum_{h=\max(0,n_1-M)}^{\min(n_1,K)} \left\{ \frac{\begin{pmatrix} K \\ h \end{pmatrix} \begin{pmatrix} M \\ n_1 - h \end{pmatrix}}{\begin{pmatrix} N \\ n_1 \end{pmatrix}} \right\} Q_h \\ &= \quad \sum_{h=\max(0,n_1-M)}^{\min(n_1,K)} f(h; K, M, n_1) Q_h \end{aligned} \tag{4}$$

where $f(h; K, M, n_1)$ is the implicitly defined probability mass function of the hypergeometric distribution.

3 Computational Issues: Approximations for Q_h

3.1 Difference in Means Statistics on Scores

The key to the approximation is state Q_h in a form such that we can use the permutational central limit theorem (PCLT), which we give informally (see Sen [1985] for formal statement).

PCLT: *Consider a permutation where the test statistic is of the form $T_\ell(\mathbf{S}, \mathbf{R}) = \sum S_i R_i$, and where both of the $N \times 1$ vectors of constants, $\mathbf{S}$ and $\mathbf{R}$, meet some regularity conditions as N gets large. Under the assumption that each permutation of $\mathbf{R}$ is equally likely,*

$$(N-1)^{-1/2} \frac{T_\ell(\mathbf{S}, \mathbf{R}) - N\bar{S}\bar{R}}{\hat{\sigma}_S \hat{\sigma}_R} \dot{\sim} N(0,1) \tag{5}$$

where $\bar{R}, \bar{S}, \hat{\sigma}_R$ and $\hat{\sigma}_S$ are sample means and standard deviations and $\dot{\sim}$ denotes approximately distributed for large N.

Before we consider chop-lump tests, we first consider a simple test statistic, representing the standardized difference in means of a set of scores $S_1, \ldots, S_N$,

$$T_{DiM}(\mathbf{S}, \mathbf{Z}) = \frac{\left(\sum S_i Z_i - n_1 \bar{S}\right)}{\sqrt{V}}$$

where all unmarked summations go from $i = 1$ to N, and $V = (N-1)\hat{\sigma}_S^2 \hat{\sigma}_Z^2$, and as above $\hat{\sigma}_S^2$ and $\hat{\sigma}_Z^2 = (N-1)^{-1}\sum(Z_i - \bar{Z})^2 = \frac{n_0 n_1}{N(N-1)}$ are sample variances of the values of $\mathbf{S}$ and $\mathbf{Z}$. The permutation t-test results when $S_i = W_i$ and the Wilcoxon rank sum test results when $S_i = rank(W_i)$ (i.e., in the earlier notation, the permutation t-test has $T(\mathbf{W}, \mathbf{Z}) = T_{DiM}(\mathbf{W}, \mathbf{Z})$, while the Wilcoxon rank sum test has $T(\mathbf{W}, \mathbf{Z}) = T_{DiM}(rank(\mathbf{W}), \mathbf{Z})$). Since within the permutations of Q_h (see equation 3) we only permute within the last M values of $\mathbf{Z}$, we want to write $T_{DiM}(\mathbf{S}, \mathbf{Z}) = a_h + b_h T_\ell(\mathbf{S}_M, \mathbf{Z}_M)$, where a_h and b_h are constant throughout the permutations in Ω_h. If we let $S_i = S_0$ for $i \le K$ (i.e., scores when $W_i = 0$) then we get

$$\begin{aligned} a_h &= \frac{hS_0 - n_1\bar{S}}{\sqrt{V}} \\ &\text{and} \\ b_h &= \frac{1}{\sqrt{V}} \end{aligned}$$

Thus, within the permutations in Ω_h,

$$\begin{aligned} & T_{DiM}(\mathbf{S}, \mathbf{Z}) \le t \\ \Rightarrow \quad & T_\ell(\mathbf{S}_M, \mathbf{Z}_M) \le \frac{t - a_h}{b_h} = t\sqrt{V} - hS_0 + n_1\bar{S} \\ \Rightarrow \quad & \frac{T_\ell(\mathbf{S}_M, \mathbf{Z}_M) - M\bar{S}_M\bar{Z}_M}{\sqrt{V_M}} \le \frac{t\sqrt{V} - hS_0 + n_1\bar{S} - M\bar{S}_M\bar{Z}_M}{\sqrt{V_M}} \end{aligned}$$

where $V_M = (M-1)\hat{\sigma}_{S_M}^2 \hat{\sigma}_{Z_M}^2$. Then substituting $\bar{Z}_M = \frac{n_1 - h}{M}$ and using the PCLT we approximate Q_h for T_{DiM} with

$$\hat{Q}_h^{(DiM)} = \Phi\left\{T_0\sqrt{\frac{V}{V_M}} + C(h)\right\}, \tag{6}$$

where $\Phi()$ is the standard normal cumulative distribution, and

$$C(h) = \frac{-hS_0 + n_1\bar{S} - (n_1 - h)\bar{S}_M}{\sqrt{V_M}}$$

3.2 Chop-Lump Statistics

Now consider the chop-lump versions of T_{DiM}, i.e., $T_{CL}(\mathbf{S}, \mathbf{Z}) = T_{DiM}\{C(\mathbf{S}, \mathbf{Z})\}$. There is one slight complication with the Wilcoxon rank sum chop-lump test; the rankings are calculated after the chop, so that the scores will change for different permutations. To minimize this problem, we rank only the non-zero values of $\mathbf{W}$, (i.e., $\mathbf{X}$), then we define S_0, the score that goes with $W_i = 0$, according to how many total zeros in the chopped data. Specifically, if there are k total zeros in the chopped data set, then let $S_0^{(k)} = -(k-1)/2$. The resulting scores give equivalent tests to the usual ranks, since they are just shifted ranks, $S_i = R_i - k$.

Suppose that when there are h zeros in the vaccine group in a permutation, this induces h^* zeros in the vaccine group of the chopped data set, where

$$h^* = \begin{cases} h - \lfloor \frac{n_1(K-h)}{n_0} \rfloor & \text{if } \frac{M-n_1+h}{n_0} \geq \frac{n_1-h}{n_1} \\ 0 & \text{if } \frac{M-n_1+h}{n_0} < \frac{n_1-h}{n_1} \end{cases}$$

Then we proceed similar to as above. Write $T_{CL}(\mathbf{S}, \mathbf{Z}) = a_{h^*} + b_{h^*} T_\ell(\mathbf{S}_M, \mathbf{Z}_M)$, where a_{h^*} and b_{h^*} are constant throughout the permutations in Ω_h. Let a superscript asterisk denote the sample sizes in the chopped data set (e.g., K^* is the total number of zeros in the chopped data, so that $K^* = h^*$ if $h^* > 0$ and $K^* = K$ if $h^* = 0$). Further let $\sum^* = \sum_{i=N-N^*+1}^{N}$.

$$T_{DiM}^{(h^*)}(\mathbf{S}_{N^*}, \mathbf{Z}_{N^*}) = \frac{\sum^* S_i Z_i - n_1^* \bar{S}_{N^*}}{\sqrt{V_{N^*}}}.$$

Now rewrite $T_{DiM}^{(h^*)}(\mathbf{S}_{N^*}, \mathbf{Z}_{N^*})$ as $a_{h^*} + b_{h^*} T_\ell(\mathbf{S}_M, \mathbf{Z}_M)$, where

$$a_{h^*} = \frac{h^* S_0^{(K^*)} - n_1^* \bar{S}_{N^*}}{\sqrt{V_{N^*}}}$$

and

$$b_{h^*} = \frac{1}{\sqrt{V_{N^*}}}$$

Again using the PCLT we approximate Q_h for T_{DiM} with

$$\hat{Q}_h^{(CL)} = \Phi\left(\frac{T_0 \sqrt{V_{N^*}} - h^* S_0^{(K^*)} + n_1^* \bar{S}_{N^*} - (n_1^* - h^*) \bar{S}_M}{\sqrt{V_M}} \right).$$