Nonparametric Survival Analysis: Cox-Mantel tests and Permutation tests

Stanley Sawyer — Washington University — September 21, 2005

1. Introduction. Assume that we have survival data $\{ X_{ia} \}$ for K different treatment groups ($1 \le i \le n_a, 1 \le a \le K$). The values X_{ia} may be uncensored or censored. If X_{ia} is *uncensored*, then $L_{ia} = X_{ia}$ is the true death or failure time. If X_{ia} is *censored*, then the true lifetime $L_{ia} > X_{ia}$ is otherwise unknown. Censoring is assumed to be independent of L_{ia} (given $L_{ia} > X_{ia}$) and independent of the sample.

We are interested in testing H_0 that the true survival times L_{ia} all have the same distribution, against the alternative that $E(L_a) \ne E(L_1)$ for some a.

We assume $K = 2$ for simplicity, which is the most common case of a treatment and a control group. In this case, we are testing H_0 against $H_1 : E(L_1) \ne E(L_2)$.

2. Permutation Tests in Survival Analysis. For a permutation test, we first assign numbers V_{ia} (called *scores*) to each observation X_{ia} in such a way that V_{ia} depends only on X_{ia}, its censoring status, and the set of values X_{ia} as a whole, but is independent of permutations of the data. The scores V_{ia} are typically ranks or midranks or functions of ranks or midranks within the entire sample, usually modified for censored observations.

The traditional nonparametric Wilcoxon rank-sum ($K = 2$) and Kruskal-Wallis ($K > 2$) tests without censored values use $V_{ia} = R_{ia}$, where R_{ia} is the rank or midrank of X_{ia} in the entire sample, so that $1 \le R_{ia} \le N$ for $N = \sum_{a=1}^{K} n_a$. *Gehan-Wilcoxon* scores are a generalization of the symmetrized ranks $R_{ia}^s = 2R_{ia} - (N + 1)$ that allow for censoring.

Scores for the K samples are defined by

$$V_a = \sum_{i=1}^{n_a} V_{ia}, \qquad a = 1, 2, \ldots, K \tag{2.1}$$

Statistical tests and P-values for H_0 depend on an assumed probability model for the observed values given H_0. Here the implicit probability model is that, given H_0, all possible permutations of the data X_{ia} among the different samples are equally likely, keeping the same number of values in each sample. Under our assumptions about the scores V_{ia}, the same is true for permutation of the scores.

For $K = 2$, permutation-test P-values are found by comparing the observed value V_1 for the first sample score with the randomized values of V_1

resulting from permutations of the scores V_{ia} among the K samples. The *upper and lower P-values* of V_1 are

$$P_{up} = \frac{\#\{\text{ Randomized } V_1 : V_1 \ge \text{Observed } V_1 \}}{\text{Total number of permutations}} \tag{2.2}$$

$$P_{lo} = \frac{\#\{\text{ Randomized } V_1 : V_1 \le \text{Observed } V_1 \}}{\text{Total number of permutations}}$$

where # means "number of". The *two-sided P-value* of the observed score V_1 is twice the smaller of P_{up} and P_{lo}, or

$$P = 2\min\{ P_{up}, P_{lo} \}$$

If $P_{up} < P_{lo}$, then the observed V_1 is on the "upper tail" of the randomized distribution of V_1, and $P = 2P_{up}$. This is usually equivalent to $V_1 \ge \overline{V}$ for

$$\overline{V} = \frac{1}{N}\sum_{a=1}^{A}\sum_{i=1}^{n_a} V_{ia} \tag{2.3}$$

If the sample sizes n_a are large and n_a/N are bounded from below, then a central limit theorem for permutations states that

$$Z_P = \frac{V_1 - n_1\overline{V}}{\sqrt{\mathrm{Var}(V_1)}} \approx N(0,1) \tag{2.4}$$

where $\overline{V}$ is as in (2.3) and

$$\mathrm{Var}(V_1) = \frac{n_1 n_2}{N(N-1)}\sum_{a=1}^{A}\sum_{i=1}^{n_a}(V_{ia} - \overline{V})^2 \tag{2.5}$$

In (2.5), "$\approx N(0,1)$" means distributed as a standard normal. The expressions $\overline{V}$ and $\mathrm{Var}(V_1)$ in (2.4)–(2.5) are mean and variance of V_1 under random permutations. The Gehan-Wilcoxon scores have $\overline{V} = 0$, which simplifies (2.4) and (2.5).

3. Cox-Mantel Tests in Survival Analysis. Another family of tests is based on viewing the observations X_{ia} for $K = 2$ as a series of contests between the two samples. Specifically, we consider a series of 2×2 contingency tables at each of the distinct observed death times t_i:

$$\begin{array}{cc} \begin{bmatrix} d_{i1} & N_{i1} - d_{i1} \\ d_{i2} & N_{i2} - d_{i1} \end{bmatrix} & \begin{matrix} N_{i1} \\ N_{i2} \end{matrix} \\ \hline \begin{matrix} d_i & N_i - d_i \end{matrix} & N_i \end{array} \tag{3.1}$$

In (3.1), the rows correspond to the two samples ($a = 1, 2$). The first column has the numbers of individuals in each sample who were observed to die or fail at time t_i. The row sums are the numbers of individuals in each sample who were "at risk" at time t_i, which is the same as

$$N_{ia} = \#\{\, j : X_{ja} \geq t_i \,\}$$

Using this definition, individuals who died at time t_i are considered to be "at risk" at time t_i, as well as those individuals who were last seen alive at (that is, were censored at) time t_i. The second column has the number of individuals who were "at risk" at time t_i but did not die.

Given the 2×2 tables (3.1), the (weighted) Cox-Mantel statistic for the first sample is

$$V_1 \;=\; \sum_{i=1}^{r} w_i \left(d_{i1} - \frac{d_i N_{i1}}{N_i} \right) \tag{3.2}$$

where

(i) $0 = t_0 \leq t_1 < t_2 < t_i < \ldots < t_r$ are the distinct times at which observed deaths (uncensored values) X_{ia} occur in either sample,

(ii) d_i is the total number of observed (uncensored) deaths at time t_i in all samples,

(iii) N_i is the size of the total "risk set" at time t_i,

(iv) d_{ia}, N_{ia} are the same as d_a, N_a but in the a^{th} sample only.

(v) $w_i \geq 0$ are arbitrary weights.

The usual Cox-Mantel or *log-rank* test has weights $w_i = 1$. The *Wilcoxon* form of the Cox-Mantel test has weights $w_i = N_i$ (see below).

The statistic (3.2) is the same as a (weighted) Mantel-Haenszel statistic for stratified 2×2 tables. The only difference is that the 2×2 tables are assumed independent in the Mantel-Haenszel test, whereas here the tables are slightly dependent in (3.2). This is because the risk-set sizes N_{i1}, N_{i2}, N_i depend on the number of deaths at previous times. However, the 2×2 tables are conditionally independent given the prior risk set sizes, which turns out to be sufficient to apply the large-sample approximation of the Mantel-Haenszel test.

The probability model for H_0 for the Cox-Mantel test is that, at each distinct observed failure time t_i, all individuals in the two risk sets of sizes N_{i1}, N_{i2} at time t_i are equally likely to die with some unknown probability p_i. Using the sample estimator $\widehat{p}_i = d_i/N_i$ for p_i, the expected number of individuals in sample #1 who die at time t_i is $\widehat{p}_i N_{i1} = (d_i/N_i)N_{i1}$. Thus $d_{i1} - (d_i/N_i)N_{i1}$ is the deviation of the observed deaths from its expected value given H_0. In particular, $E(V_1) = 0$ given H_0 for V_1 in (3.2). The expected value is conditional (in each term) on the risk set sizes N_{i1}, N_{i2} and the total number of deaths d_i at that time.

Under these assumptions, the observed counts d_{i1} can be assumed to have a hypergeometric distribution conditional on d_i, N_{i1}, and N_{i2}, exactly as in the Mantel-Haenszel test. The variance of V_1 is the sum of the conditional variances for each term, which is

$$\mathrm{Var}(V_1) = \sum_{i=1}^{r} w_i^2 \, \frac{d_i(N_i - d_i)N_{i1}N_{i2}}{N_i^2(N_i - 1)} \tag{3.4}$$

where the i^{th} term is w_i^2 times the variance of the corresponding hypergeometric distribution. Note that the numerator of the fractions in (3.4) are the products of the four row and column sums in (3.1). If N is large and n_a/N are bounded from below, the expression

$$Z_C \;=\; \frac{\displaystyle\sum_{i=1}^{r} w_i \left(d_{i1} - \frac{d_i N_{i1}}{N_i} \right)}{\sqrt{\displaystyle\sum_{i=1}^{r} w_i^2 \, \frac{d_i(N_i - d_i)N_{i1}N_{i2}}{N_i^2(N_i - 1)}}} \tag{3.5}$$

has a standard normal distribution.

4. Cox-Mantel Scores are Permutation Scores. We now prove that the weighted Cox-Mantel statistic (3.2) (that is, the numerator of (3.5)) can always be written in a natural way as a sum of permutation-test-like scores V_{ja} as in (2.1).

We first note that the risk set sizes N_{i1} can be found by summing over the times $t_j \geq t_i$:

$$N_{i1} = \sum_{j=i}^{r} (d_{j1} + c_{j1}) \tag{4.1}$$

where c_{j1} is the number of censored observations X_{1k} with $t_j \leq X_{1k} < t_{j+1}$, where $t_0 = 0$ and $t_{r+1} = \infty$ for convenience. Then by (3.2) and (4.1)

$$\begin{aligned} V_1 \;&=\; \sum_{i=1}^{r} w_i \left(d_{i1} - \frac{d_i N_{i1}}{N_i} \right) \qquad (4.2) \\ &=\; \sum_{i=1}^{r} w_i d_{i1} - \sum_{i=1}^{r} w_i \frac{d_i}{N_i} \sum_{j=i}^{r} (d_{j1} + c_{j1}) \end{aligned}$$

$$= \sum_{i=1}^{r} w_i d_{i1} - \sum_{i=1}^{r} (d_{i1} + c_{i1}) \sum_{j=1}^{i} w_j \frac{d_j}{N_j}$$

$$= \sum_{i=1}^{r} \left(w_i - \sum_{j=1}^{i} w_j \frac{d_j}{N_j} \right) d_{i1} \; - \; \sum_{i=1}^{r} \left(\sum_{j=1}^{i} w_j \frac{d_j}{N_j} \right) c_{i1}$$

Since the first sample size $n_1 = \sum_{j=1}^{r} (d_{j1} + c_{j1})$ as in (4.1), the last expression can be viewed as a sum over all of the individuals in the first sample. That is,

$$V_1 = \sum_{j=1}^{n_1} V_{j1}$$

where

$$V_{ja} \;=\; \begin{cases} w_i - \sum_{k=1}^{i} w_k \frac{d_k}{N_k} & \text{If } X_{ja} = t_i \text{ is observed} \\ -\sum_{k=1}^{i} w_k \frac{d_k}{N_k} & \text{If } X_{ja} \text{ is censored and } t_i \le X_{ja} < t_{i+1} \end{cases} \tag{4.3}$$

Note that the V_{ja} in (4.3) do not depend explicitly on the sample designator a. They are also constant within ties groups (including censoring state) of the values X_{ja}. This shows that the Cox-Mantel statistic (3.1) can be written as a sample permutation-like score (2.1).

5. Examples. (Example 1.) The *Wilcoxon* form of the Cox-Mantel statistic (3.1) uses weights $w_i = N_i$. Then by (4.3)

$$V_{ja} \;=\; \begin{cases} N_i - \sum_{k=1}^{i} d_k & \text{If } X_{ja} = t_i \text{ is observed} \\ -\sum_{k=1}^{i} d_k & \text{If } X_{ja} \text{ is censored and } t_i \le X_{ja} < t_{i+1} \end{cases}$$

If there are no ties and no censoring, then $d_k = 1$ and the risk set sizes are $N_i = N - i + 1$. Then

$$V_{ja} = N_i - i = N - 2i + 1 = -\big(2i - (N+1)\big)$$

which is exactly the negative of the symmetrized Wilcoxon rank-sum rank of V_{ja}. If there are ties but no censoring, then V_{ja} are minus the Wilcoxon midranks. (*Exercise*: Prove this.)

The difference in sign results from the fact that if (for example) sample #1 dies at a faster rate, then its Wilcoxon ranks will be smaller (and thus $V_1 < \overline{V}$) while the entries d_{i1} in the 2×2 tables in (3.3) will be larger than expected (and hence $V_1 > 0$). Except for the sign, the sample scores V_1 are the same. Of course, you could make the signs the same by using V_1 for one test and the analog of V_2 for the other test, but the difference in definitions may be more confusing than the difference in signs. It is not uncommon for signs to vary in survival analysis statistics due to different ways of counting deaths.

In general, with ties and censoring, V_{ja} with $w_i = N_i$ can be written

$$V_{ja} \;=\; \#\{\,(k,b) : X_{kb} \ge t_i\,\} \;-\; \#\{\,(k,b) : \text{Observed death } X_{kb} \le t_i\,\}$$

if $X_{ja} = t_i$ is observed and

$$V_{ja} \;=\; -\#\{\,(k,b) : \text{Observed death } X_{kb} \le X_{ja}\,\}$$

if X_{ja} is censored. This is exactly the Mantel form of the Gehan-Wilcoxon statistic. (*Exercise*: Prove that the two sets of formulas for V_{ja} are the same.)

Example 2. The *log-rank* form of the Cox-Mantel statistic (3.1) uses weights $w_i = 1$. Then

$$V_{ja} \;=\; \begin{cases} 1 - \sum_{k=1}^{i} \frac{d_k}{N_k} & \text{If } X_{ja} = t_i \text{ is observed} \\ -\sum_{k=1}^{i} \frac{d_k}{N_k} & \text{If } X_{ja} \text{ is censored and } t_i \le X_{ja} < t_{i+1} \end{cases}$$

If there are no ties and no censored observations, then $d_k = 1$ and $N_k = N - k + 1$ as before. Then

$$\sum_{k=1}^{i} \frac{d_k}{N_k} \;=\; \sum_{k=1}^{i} \frac{1}{N-k+1} \;=\; \sum_{k=N-i+1}^{N} \frac{1}{k} \;\approx\; \log\left(\frac{N}{N-i+1}\right)$$

The "log" in the log-rank test comes from this logarithmic approximation.

6. Conclusion. The *sample scores* for the Gehan-Wilcoxon test and the Wilcoxon form of the Cox-Mantel test — that is, the *numerators* of the test statistics (2.4) and (3.1) — are exactly the same except for the sign. However, the probability models for the two tests differ.

The means of the sample scores given H_0 are zero in both cases, but the variances — that is, the expression (2.5) (with $\overline{V} = 0$) and (3.4) — are generally different. In practice, the values of the large-sample normal statistics Z_P in (2.4) and Z_C in (3.5) are usually similar but slightly different.

7. Which is Best: Gehan-Wilcoxon or Log Rank? The Cox-Mantel statistic (3.1) with $w_i = 1$ puts equal weight on deaths at all observed death times while the Gehan-Wilcoxon test (in effect) uses weights $w_i = N_i$ for all deaths.

Consider a general probability model in which, for given observed death times t_i, each extant individual of sample #1 dies with probability p_{i1} at time t_i and each extant individual of sample #2 dies with probability p_{i2}. Note that p_{i1} and p_{i2} do not specify death *rates*, since the death times t_i can be spaced out or bunched in without affecting p_{i1} and p_{i2}.

If $p_{i1} = p_1$ and $p_{i2} = p_2$ are constant in time, then it can be shown that the log-rank test is more powerful than the Gehan-Wilcoxon test for detecting $H_1 : p_1 \neq p_2$. However, if $p_{ia} = C_i p_a$ where $C_i = N_i$, then the Gehan-Wilcoxon test can be shown to be more powerful. That is the Gehan-Wilcoxon test is more powerful if initial death rates are higher, but not if death rates are constant over time.

From the form of (3.1), if you want to put equal emphasis on all deaths, you should use the log-rank test. If, conversely, you want to put more emphasis on earlier deaths (perhaps later deaths are more likely to be due to unrelated causes), then the Gehan-Wilcoxon test may be preferable.