Notes for Biostatistics 140.641.01 'Survival Analysis'

First Term, 2022

Chapter 6. Hypothesis Testing

Goal of testing: Determine if there is a difference between two groups.

- Some of the "traditional methods" are appropriate for complete survival times but not applicable to censored data.

Proper use of test statistics: Statistical tests together with 'p-value' are widely used for data analysis even though the results can be manipulated by small or large sample size. So, it is important to make sure that test analysis results are accompanied by interval estimation (e.g., point estimate together with the corresponding confidence interval).

6.1 Complete Survival Times

Suppose there is no censoring and the data include $t_1, t_2, \ldots, t_n$. For fixed t, we are interested in the t-year survival rate, $S(t)$, and observe

		D	$\bar{D}$	
Treatment	A	e	f	n_A
	B	g	h	n_B
		m_D	$m_{\bar{D}}$	n

D : Failing in t years, $T_i < t$

$\bar{D}$: Surviving beyond t years, $T_i \geq t$

$p_A = \mathsf{P}(D|A) \quad p_B = \mathsf{P}(D|B)$

Consider the following way to construct a χ^2 test statistic:

		D	$\bar{D}$	
Treatment	A	e	f	n_A
	B	g	h	n_B
		m_D	$m_{\bar{D}}$	n

Null hypothesis $H_0 : p_A = p_B$ or, equivalently, $S_A(t) = S_B(t)$.
Conditional on $n_A, n_B, m_D, m_{\bar{D}}$, the count "E" follows a hypergeometric distribution (under H_0) with

$$\begin{aligned} \mathsf{E}_0(E) &= m_D \left(\frac{n_A}{n}\right) \\ \mathsf{Var}_0(E) &= \frac{n_A n_B m_D m_{\bar{D}}}{n^2(n-1)} \end{aligned}$$

Construct a test statistic

$$W = \left(\frac{E - m_D \left(\frac{n_A}{n} \right)}{\sqrt{\frac{n_A n_B m_D m_{\bar{D}}}{n^2(n-1)}}} \right)^2$$

when n is large, $W \sim \chi^2(1)$.

6.2 A Test for Right Censored Data

Suppose t-year survival rate is of interest

$$H_0 : S_A(t) = S_B(t) \qquad H_1 : S_A(t) > S_B(t).$$

As the survival data could be censored before t, we use the K-M estimate to estimate $S_A(t)$ and $S_B(t)$ and construct a test statistic

$$T = \frac{\hat{S}_A(t) - \hat{S}_B(t)}{\widehat{SD}[\hat{S}_A(t) - \hat{S}_B(t)]} \sim N(0,1).$$

Here $SD[\hat{S}_A(t) - \hat{S}_B(t)]$ can be estimated by Greenwood's formula,

$$\begin{aligned} \text{Var}[\hat{S}_A(t) - \hat{S}_B(t)] &= \text{Var}(\hat{S}_A(t)) + \text{Var}(\hat{S}_B(t)) \\ \widehat{SD}[\hat{S}_A(t) - \hat{S}_B(t)] &= \sqrt{\widehat{\text{Var}}(\hat{S}_A(t)) + \widehat{\text{Var}}(\hat{S}_B(t))}, \end{aligned}$$

where $\widehat{\text{Var}}$ is derived by by Greenwood's formula.

This test is designed to test the survival difference at a specified time, t. It can be extended to test the overall difference of two survival functions:

$$H_0 : S_A(t) = S_B(t)$$
$$H_1 : S_A(t) > S_B(t) \text{ or } S_A(t) < S_B(t) \text{ for all } t$$

In this course we shall focus on a more popular statistical test, namely the log-rank test, which serves the purpose to test the overall difference between two hazard functions.

6.3 Log-rank Test for Right Censored Data

Null hypothesis is $H_0 : \lambda_A(t) = \lambda_B(t)$ for all t

(Note: In H_0, "for all t" might be replaced by "for observed t" in practice.)

What should alternative hypothesis be like? There is not parametric distributional assumption on H_1, but the hazard functions from the two groups are **not** allowed to cross-over each other.

Steps for constructing log-rank test

1. Create a 2×2 table at an uncensored survival time on the basis of the corresponding risk set.
2. Construct a test statistic based on each 2×2 table.
3. Combine all the test statistics from tables to construct a final test statistic (log-rank test)

The first step is to construct a 2×2 table at each **uncensored** survival time. At an uncensored time y ($y = y_{(i)}$ for some i),

		D	$\bar{D}$	
Treatment	A	e	$n_A - e$	n_A
Treatment	B	$m_D - e$	$n_B - (m_D - e)$	n_B
		m_D	$m_{\bar{D}}$	N

N: number of individuals in the risk set $R(y)$ from pooled data

e: number of failures at y from group A

m_D: number of failures at y from pooled data

n_A: number of individuals in the risk set $R(y)$ from group A

n_B: number of individuals in the risk set $R(y)$ from group B

$m_{\bar{D}} = N - m_D$

The second step is to construct a test statistic from the 2×2 table at an uncensored survival time:

Conditioning on $n_A, n_B, m_D, m_{\bar{D}}$, the random number e follows a hypergeometric distribution (under H_0) with probability

$$\frac{\binom{n_A}{e}\binom{n_B}{m_D - e}}{\binom{N}{m_D}}, \qquad \max(0, m_D - n_B) \leq e \leq \min(n_A, m_D).$$

Under H_0, $\mathsf{E}_0(E) = m_D\left(\frac{n_A}{N}\right)$ and $\mathsf{var}_0(E) = \frac{n_A n_B m_D m_{\bar{D}}}{N^2(N-1)}$

A test statistic is $E - m_D\left(\frac{n_A}{n}\right)$ (i.e., observed - expected).

The third step is to combine information from all the tables to construct the log-rank test statistic:

$$Z = \frac{\sum_{i=1}^{K}(E_{(i)} - \mathsf{E}_0[E_{(i)}])}{\sqrt{\sum_{i=1}^{K} \mathsf{var}_0(E_{(i)})}} \underset{n \text{ large}}{\sim} N(0,1)$$

where the variance of $\sum_{i=1}^{K}(E_{(i)} - \mathsf{E}_0[E_{(i)}])$ is $\sum_{i=1}^{K} \mathsf{var}_0(E_{(i)})$.

From a real data set, $Z = z$ is calculated as

$$z = \frac{\sum_{i=1}^{k}\left(e_{(i)} - \frac{m_{D_{(i)}} \cdot n_{A_{(i)}}}{N_{(i)}}\right)}{\sqrt{\sum_{i=1}^{k} \frac{n_{A_{(i)}} n_{B_{(i)}} m_{D_{(i)}} m_{\bar{D}_{(i)}}}{N_{(i)}^2 (N_{(i)} - 1)}}}$$

When do we reject H_0?
The null hypothesis is $H_0 : \lambda_A(t) = \lambda_B(t)$ for all t.

Consider three different kinds of alternatives under the assumption that there is **no cross-over** between hazard functions:

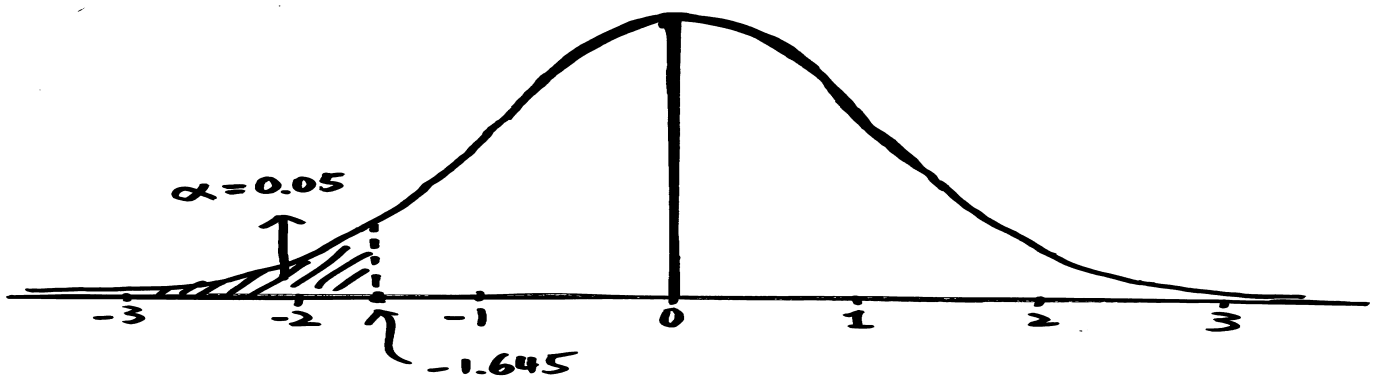
(A1) $H_1 : \lambda_A(t) < \lambda_B(t)$ for all t (treatment A is better)
(A2) $H_1 : \lambda_A(t) > \lambda_B(t)$ for all t (treatment B is better)
(A3) H_1 : either $\lambda_A(t) > \lambda_B(t)$ for all t, or $\lambda_A(t) < \lambda_B(t)$ for all t (no prior knowledge)

Usually the significance level of a test is set up to be 0.05.

For (A1),

When H_1 is true, Z is likely to be negative, so reject H_0 when z is small, that is, $z < -1.645$.

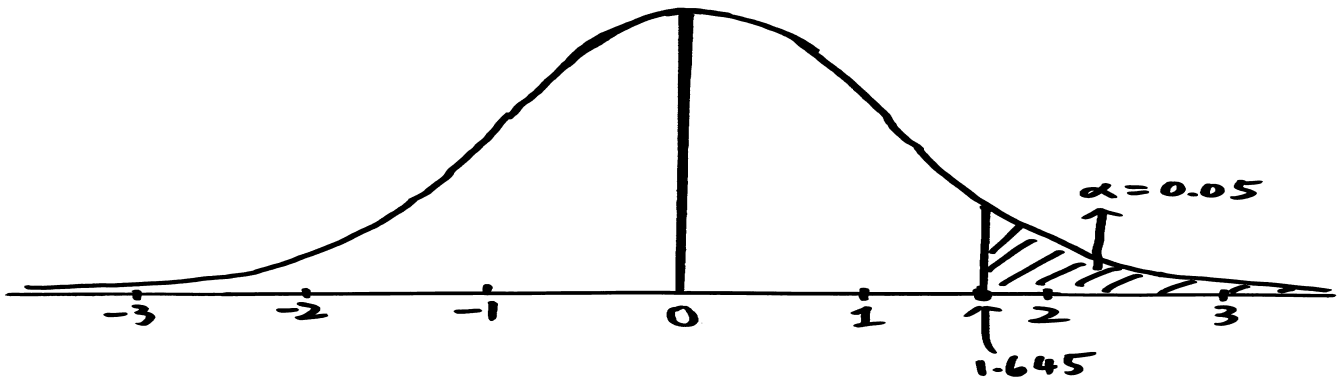
P-value $=$ Probability for values smaller than z.



For (A2)

When H_1 is true, Z is likely to be positive, so reject H_0 when z is large, that is, $z > 1.645$

P-value $=$ Probability for values larger than z.

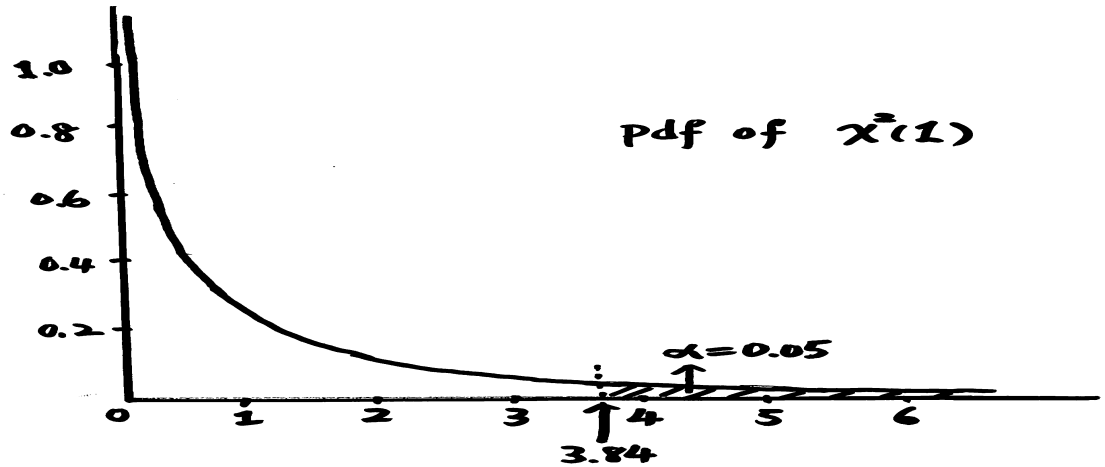


For (A3), use

$$Z^2 = \left[\frac{\sum_{i=1}^{K} (E_{(i)} - \mathsf{E}_0[E_{(i)}])}{\sqrt{\sum_{i=1}^{K} \mathsf{Var}_0(E_{(i)})}} \right]^2 \underset{n \text{ large}}{\sim} \chi^2(1)$$

Reject H_0 when $z^2 > 3.84$ ($|z| > 1.96$)

p-value = Probability for values larger than z^2.



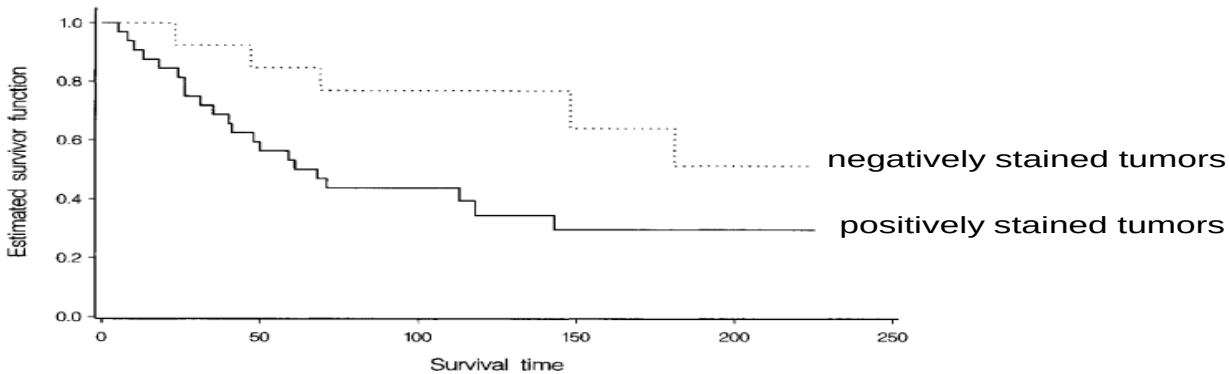
A real data example

Do two groups have significantly different survival performance? In many data applications, log-rank statistic is used to more formally address the question.

In this data example, two-sided test log-rank test: z^2=3.515, p=0.061

Remark: Sample size from one of the two groups might be too small. The p-value could become smaller when sample size of that group increases.

Example. Prognosis for breast cancer using HPA



Recall: $H_0 : \lambda_A(t) = \lambda_B(t)$ for all t.

Possible alternatives under the assumption that there is **no cross-over** between hazard functions:

(A1) $H_1 : \lambda_A(t) < \lambda_B(t)$ for all t (treatment A is better)
(A2) $H_1 : \lambda_A(t) > \lambda_B(t)$ for all t (treatment B is better)
(A3) H_1 : either $\lambda_A(t) > \lambda_B(t)$ for all t, or $\lambda_A(t) < \lambda_B(t)$ for all t
(no prior knowledge)

When 'cross-overs in hazards' occur, the log-rank test no longer holds!

Property. Suppose T_0 and T_1 are continuous survival time variables. Denote by $\lambda_0(t)$ and $\lambda_1(t)$ the hazard function of T_0 and T_1, respectively. Then $\lambda_0(t) \geq \lambda_1(t)$ for all $t > 0$ implies $S_0(t) \leq S_1(t)$ for all $t > 0$.

Proof. Note that $\lambda_0(t) \geq \lambda_1(t)$ for all $t > 0$ implies

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du \geq \int_0^t \lambda_1(u)du = \Lambda_1(t)$$

and

$$S_0(t) = e^{-\Lambda_0(t)} \leq e^{-\Lambda_1(t)} = S_1(t) \ .$$

Remark 1: $S_0(t) \leq S_1(t)$ for all $t > 0$ does NOT imply $\lambda_0(t) \geq \lambda_1(t)$ for all $t > 0$. So, it is possible that $S_0(t) \leq S_1(t)$ holds while there are cross-overs between hazards $\lambda_0(t)$ and $\lambda_1(t)$.

Remark 2: When the 'no-cross in hazards' assumption is violated, statisticians might take an alternative test which is designed to test the difference between two survival functions, $S_0(t)$ and $S_1(t)$.

Example.

Group A 3, 5, 7, 9^+, 18

Group B 12, 18, 20, 20, 20^+, 33^+

Uncesored times: 3, 5, 7, 12, 18, 20

Null hypothesis $H_0 : \lambda_A(t) = \lambda_B(t)$

$y_{(1)} = 3$

	D	$\bar{D}$	
A	1	4	5
B	0	6	6
	1	10	11

$y_{(2)} = 5$

	D	$\bar{D}$	
A	1	3	4
B	0	6	6
	1	9	10

$y_{(3)} = 7$

	D	$\bar{D}$	
A	1	2	3
B	0	6	6
	1	8	9

$y_{(4)} = 12$

	D	$\bar{D}$	
A	0	1	1
B	1	5	6
	1	6	7

$y_{(5)} = 18$

	D	$\bar{D}$	
A	1	0	1
B	1	4	5
	2	4	6

$y_{(7)} = 20$

	D	$\bar{D}$	
A	0	0	0
B	2	2	4
	2	2	4

$y_{(i)}$	$E_{(i)}$	$\mathsf{E}_0[E_{(i)}]$	$\mathsf{Var}_0[E_{(i)}]$
3	1	$1 \times \frac{5}{11} = 0.45$	$\frac{5 \times 6 \times 1 \times 10}{11^2 \cdot 10} = 0.30$
5	1	$1 \times \frac{4}{10} = 0.40$	$\frac{4 \times 6 \times 1 \times 9}{10^2 \cdot 9} = 0.24$
7	1	$1 \times \frac{3}{9} = 0.33$	0.22
12	0	$1 \times \frac{1}{7} = 0.14$	0.12
18	1	$2 \times \frac{1}{6} = 0.33$	0.22
20	0	$2 \times \frac{0}{4} = 0$	0

$$\sum_{1}^{6}(e_{(i)} - \mathsf{E}_0(E_{(i)})) = 0.55 + 0.60 + 0.67 - 0.14 + 0.67 = 2.35$$

$$\sum_{1}^{6}\mathsf{Var}_0(E_{(i)}) = 0.30 + 0.24 + 0.22 + 0.12 + 0.22 = 1.10$$

$$z = \frac{2.35}{\sqrt{1.10}} = 2.24$$

Now if $H_1 : \lambda_A \neq \lambda_B$ (two-sided)

$$z^2 = (2.24)^2 = 5.02 > 3.84$$

p-value $\approx 0.025 \Rightarrow$ reject H_0.

if $H_1 : \lambda_A > \lambda_B$ (one-sided)

$$z = 2.24 > 1.645$$

p-value $\approx 0.013 \Rightarrow$ reject H_0.

Warning: Sample size might be too small for the validity of χ^2 and normal distribution approximation!

6.4 Generalization of Log-Rank Test

After constructing a sequence of 2×2 tables at uncensored times, we consider the statistic $T = \sum_{i=1}^{K} w_{(i)}(E_{(i)} - \mathsf{E}_0[E_{(i)}])$ where $w_{(i)}$ is the "weight" on the table at $y_{(i)}$. The variance of T is $\sum_{i=1}^{K} w_{(i)}^2 \mathsf{Var}(E_{(i)})$, and the weighted log-rank test statistic is

$$Z = \frac{\sum_{(i)} w_{(i)}(E_{(i)} - \mathsf{E}_0(E_{(i)}))}{\sqrt{\sum_{(i)} w_{(i)}^2 \mathsf{Var}_0(E_{(i)})}} .$$

For a real data set, the realization value of Z is

$$z = \frac{\sum_{(i)} w_{(i)} \left(e_{(i)} - \frac{m_{D_{(i)}} n_{A_{(i)}}}{N_{(i)}} \right)}{\sqrt{\sum_{(i)} \frac{w_{(i)}^2 n_{A_{(i)}} n_{B_{(i)}} m_{D_{(i)}} m_{\bar{D}_{(i)}}}{N_{(i)}(N_{(i)}-1)}}} \quad \begin{matrix} \text{approx} \\ \sim \\ n \text{ large} \end{matrix} \quad N(0,1)$$

Three cases of the weighted log-rank test:

(i) $w_{(i)} = 1$, $Z =$ log-rank test

(ii) $w_{(i)} = N_{(i)}$, $Z =$ Gehan's test (1965, Biometrika)

(iii) $w_{(i)} = \sqrt{N_{(i)}}$, $Z =$ Tarone and Ware test (Biometrika, 1977)

The tests of (ii) and (iii) are motivated by examining the risk set size and giving weights to tables according to the risk set sizes. In general, the log-rank test is more efficient under the proportional hazards model, and (ii) and (iii) are more efficient under other classes of models.

For example, if the underlying model is a two-sample PHM, $\lambda_B(t) = \lambda_A(t)e^{\beta}$. The hypotheses can be set as

$$\begin{cases} H_0 : \beta = 0 \quad (\lambda_A(t) = \lambda_B(t)) \\ \\ H_1 : \beta \neq 0 \ \text{ or } \ H_1 : \beta > 0 \ \text{ or } \ H_1 : \beta < 0 \end{cases}$$

then the log-rank test coincides with the partial likelihood score test and is the most powerful test.

Remark: The partial likelihood score test is $W = \frac{\partial}{\partial \beta} \log L_p \mid_{\beta=0}$.

Further Remarks:

1. If the relative hazard is large at earlier times, then Gehan's test and Tarone and Ware's test might be more powerful than the log-rank test.

2. When cross-over in hazards occurs, the weighted or unweighted log-rank tests would not be proper tests in general.

3. Gehan's test is closely related to the Mann-Whitney-Wilcoxon test for complete data. It can be regarded as a generalization of the Mann-Whitney-Wilcoxon test.

6.5 Gehan's test as a generalization of Mann-Whitney-Wilcoxon Test

Nonparametric Wilcoxon Test for Complete Data

$$\begin{aligned} \text{Data from} \quad & \text{treatment A} = t_1, \ldots, t_m \sim S_A \\ & \text{treatment B} = z_1, \ldots, z_n \sim S_B \end{aligned}$$

Here $t_1, \ldots, t_m, z_1, \ldots, z_n$ are survival times (uncensored). $H_0 : \ S_A = S_B$.
The general idea is the following. Pool the data from treatments A and B. Rank the data. Calculate the sum of ranks from treatment-A data. If the rank-sum is large or small, then reject the null hypothesis.

Example. $A: 3,\ 7,\ 2 \qquad m = 3$

$B: 1,\ 4, \qquad n = 2$

Ordered data (1, 2, 3, 4, 7)

Ranks for 3, 7, 2 are (3, 5, 2)

Rank sum is $3 + 5 + 2 = 10$. Is "10" large or small? We will discuss it.

Order the pooled data and define

$$\begin{aligned} \gamma_i &= \text{rank of } t_i,\ t = 1, \ldots, m \\ R &= \sum_{i=1}^{m} \gamma_i \end{aligned}$$

Under $H_0 :\ S_A = S_B$,

$$\begin{aligned} \mathsf{E}_0[R] &= m\left(\frac{1+(m+n)}{2}\right) \\ \mathsf{Var}_0(R) &= \frac{mn(m+n+1)}{12} \text{ from permutation theory} \end{aligned}$$

Testing statistics is

$$W = \frac{R - E_0(R)}{\sqrt{\text{Var}_0(R)}}$$

When m, n are small $\Rightarrow$ Use small sample tables. Reject H_0 when W is far away from 0.

When m, n are large, use approximation result

$$W = \frac{R - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \overset{\text{approx}}{\sim} N(0,1)$$

Reject H_0 when W is very different from 0 (that is, R is very large or small).

To use the Wilcoxon test, the usual underlying models we have in mind are likely to be either $S_A(t) \geq S_B(t)$ or $S_A(t) \leq S_B(t)$. Examples:

- Location-shift model: $f_A(t)$, $f_B(t)$ are the pdf for the survival time T from group A and B respectively. Location-shift model assumes $f_A(t) = f_B(t - \theta)$
- Proportional hazards model $\lambda_B(t) = \lambda_A(t)e^{\beta}$

Extension of M-W-W Test: Gehan's Test for Right Censored Data

For complete and continuous data, an alternative way to write the rank sum is

$$R = \frac{m(m+n+1)}{2} + \frac{1}{2}U \qquad (*)$$

and U is defined as

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} U_{ij} \text{ where } U_{ij} = \begin{cases} 1 & \text{if } t_i > z_j \\ 0 & \text{if } t_i = z_j \\ -1 & \text{if } t_i < z_j \end{cases}$$

The statistic "U" is also called the Mann-Whitney statistic, which explains why the test is called the Mann-Whitney-Wilcoxon test (they are essentiall equivalent from statistical perspective). Reject H_0 if U is away from 0. Gehan (*Biometrika*, 1965) modified U_{ij} subject to right censored data.

Remark: See Appendix A.2 for a detailed proof of $(*)$.

Now the data are

A-sample $(y_1, \delta_1), \ldots, (y_m, \delta_m)$

B-sample $(y_1^*, \delta_1^*), \ldots, (y_n^*, \delta_n^*)$ δ_i, δ_j^* = censoring indicator.

Define

$$U_{ij} = \begin{cases} 1 & \text{if} \quad t_i > z_j \\ 0 & \text{either} \quad \text{“}t_i = z_j\text{”} \text{ or “don't know”} \\ -1 & \text{if} \quad t_i < z_j \end{cases}$$

Note: t_i and z_j may not be observable!

The Gehan statistic is

$$G = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij} \overset{\text{approx}}{\sim} N(0, \sigma^2)$$ Reject H_0 if G is large or small

Example.

$$A = \quad 3, 5, 7, 9^+, 18$$
$$B = \quad 12, 19, 20, 20^+, 33^+$$

$$G = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij}$$

$$i = 1, \quad \sum_{j=1}^{5} U_{1j} = (-1) + (-1) + (-1) + (-1) + (-1) = -5$$

$$i = 2, \quad \sum_{j=1}^{5} U_{2j} = -5$$

$$i = 3, \quad \sum_{j=1}^{5} U_{3j} = -5$$

$$i = 4, \quad \sum_{j=1}^{5} u_{4j} = 0$$

$$i = 5, \quad \sum_{j=1}^{5} U_{ij} = 1 + (-1) + (-1) + (-1) + (-1) = -3$$

The Gehan statistic is $G = -5 \ -5 \ -5 \ +0 \ -3 = -18$.

Remark: We can prove that Gehan's test is a weighted log-rank test with weight $w_{(i)} = N_{(i)}$ at uncensored times. The variance of Gehan's test can then be formulated using the variance formula for weighted log-rank test on page 24. See Appendix A.2 for more detailed discussion.

*Appendix

A.1 Log-rank test for three or more groups

Suppose we are interested in testing the null hypothesis that the hazard functions are the same for $K > 2$ groups. For example, $K > 2$ treatments are evaluated in a randomized clinical trial. The null hypothesis of no treatment difference is formulated as

$$H_0: \quad \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_K(t) \text{ for all } t$$
$$H_1: \quad \lambda_i(\cdot) \neq \lambda_j(\cdot) \text{ for } i \neq j \quad \text{(at least one equality fails to hold)}$$

A required assumption is that there is no cross-over between any two hazard functions from $(\lambda_1, \lambda_2, \ldots, \lambda_K)$.

Let $y_{(i)}$, $i = 1, 2, \ldots, m$, be the i^{th} uncensored time in the pooled sample. At $y_{(i)}$, the data can be viewed as a $2 \times K$ table

	1	2		k		K	
D	d_{1i}	d_{2i}	$\cdots$	d_{ki}	$\cdots$	d_{Ki}	d_i
$\bar{D}$	$\bar{d}_{1i}$	$\bar{d}_{2i}$		$\bar{d}_{ki}$		$\bar{d}_{Ki}$	$\bar{d}_i$
	n_{1i}	n_{2i}		n_{ki}		n_{Ki}	n_i

We now consider a vector of observed number of deaths minus their expected number of deaths under the null hypothesis for each treatment group $(O_1 - E_1, O_2 - E_2, \ldots, O_K - E_K)^t$. Define

Note: The sum of the elements in this vector is equal to zero, which means one element is redundant. Define

$$O_{ki} = d_{ki}, \qquad E_{ki} = \frac{n_{ki} d_i}{n_i}, \quad O_k - E_k = \sum_{i=1}^{m} (O_{ki} - E_{ki})$$

- Test statistic

$$X^2 = \begin{pmatrix} O_1 - E_1 \\ O_2 - E_2 \\ \vdots \\ O_{K-1} - E_{K-1} \end{pmatrix}^t V^{-1} \begin{pmatrix} O_1 - E_1 \\ O_2 - E_2 \\ \vdots \\ O_{K-1} - E_{K-1} \end{pmatrix} \underset{H_0}{\overset{\text{large } n}{\sim}} \chi^2_{K-1}$$

- V: the corresponding $(K-1) \times (K-1)$ covariance matrix of the vector.
- Under H_0, this is distributed asymptotically as a χ^2 distribution with (K-1) degrees of freedom.

Variance-covariance components in V:

- $V_{jj,i} = \frac{n_{ji}(n_i - n_{ji})d_i\bar{d}_i}{n_i^2(n_i-1)}$, $V_{jj',i} = \frac{-n_{ji}n_{j'i}d_i\bar{d}_i}{n_i^2(n_i-1)}$
- $V_{jj} = \sum_{i=1}^{m} V_{jj,i}$ $V_{jj'} = \sum_{i=1}^{m} V_{jj',i}$

  Note: The covariance between group j and j' is negative because, conditioning on the marginal totals, the increase of number of deaths in group k implies the decrease of number of deaths in other groups.

- Hence, a level α test would reject the null hypothesis whenever $X^2 \geq \chi^2_{\alpha,K-1}$, where $\chi^2_{\alpha,K-1}$ is the number satisfying $P(W \geq \chi^2_{\alpha,K-1}) = \alpha$, where W follows χ^2_{K-1} distribution.
- This test statistic numerically remains the same for any specific choice of the K-1 groups.
- The test statistic is a generalization of the two-sample log-rank statistic.
- SAS proc lifetest, STRATA statement can help you compute the test statistics.

A.2 Gehan's test as a generalization of Wilcoxon Test

To see the validity of (*), consider the condition when we have the total separation

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)} < z_{(1)} < \ldots < z_{(n)},$$

then $R = \frac{m(m+1)}{2}$. For every interchange of a consecutive (t, z) pair, R is increased by 1, and the number of interchanges is

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{1}{2} [U_{ij} + 1].$$

Thus

$$\begin{aligned} R &= \frac{m(m+1)}{2} + \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}\left[U_{ij}+1\right] \\ &= \frac{m(m+1)}{2} + \frac{m\cdot n}{2} + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}U_{ij} \\ &= \frac{m(m+n+1)}{2} + \frac{1}{2}U. \end{aligned}$$

Example.

$$A = \quad 3, 5, 7, 9^+, 18$$
$$B = \quad 12, 19, 20, 20^+, 33^+$$

$$G = \sum_i \sum_j U_{ij}$$

$$i = 1, \quad \sum_{j=1}^{5} U_{1j} = (-1) + (-1) + (-1) + (-1) + (-1) = -5$$

$$i = 2, \quad \sum_{j=1}^{5} U_{2j} = -5$$

$$i = 3, \quad \sum_{j=1}^{5} U_{3j} = -5$$

$$i = 4, \quad \sum_{j=1}^{5} u_{4j} = 0$$

$$i = 5, \quad \sum_{j=1}^{5} U_{ij} = 1 + (-1) + (-1) + (-1) + (-1) = -3$$

The Gehan statistic is

$$G = -5 \ -5 \ -5 \ +0 \ -3 = -18.$$

To get p-value, we need to estimate σ^2. Gehan provided a complicated formula (*Biometrika*, 1965). For your calculation, just use the "weighted" formula (ii) introduced earlier. Because

$$\begin{aligned} G &= -\sum_{(i)} N_{(i)} \left[d_{(i)} - \mathsf{E}_0 \left(d_{(i)} \right) \right] \\ &= -\sum_{(i)} N_{(i)} \left[d_{(i)} - \frac{m_{D_{(i)}} n_{A_{(i)}}}{N_{(i)}} \right] , \end{aligned}$$

we may derive the variance of the Gehan statistic by the previous formula. To see the equivalence, note that

$$\begin{aligned} G &= \sum_{y_i \text{ censored}} \sum_{j \in R_i} U_{ij} + \sum_{y_{(i)}} \sum_{j \in R_{(i)}} U_{ij} \\ &= I + II \end{aligned}$$

Clearly, $I = 0$. For II, if the failure at $y_{(i)}$ is from group "A", then the score is

$$-\{(N_{(i)} - n_{A_{(i)}}) - (m_{D_{(i)}} - d_{(i)})\}$$

$$\searrow$$

$$\#\text{ of failure at } y_{(i)} \text{ from "}B\text{"}$$

and $n_{A_{(i)}} - d_{(i)}$ otherwise. Thus the total score evaluated $y_{(i)}$ is

$$-\left[d_{(i)}\left(N_{(i)} - n_{A_{(i)}} - m_{D_{(i)}} + d_{(i)}\right) - \left(m_{D_{(i)}} - d_{(i)}\right)\left(n_{A_{(i)}} - d_{(i)}\right)\right]$$

$$= \quad -\left[d_{(i)}N_{(i)} - m_{D_{(i)}}n_{A_{(i)}}\right].$$

Thus

$$\begin{aligned} G &= -\sum_{y_{(i)}} \left[d_{(i)} N_{(i)} - m_{D_{(i)}} n_{A_{(i)}} \right] \\ &= -\sum_{(i)} N_{(i)} \left[d_{(i)} - \frac{m_{D_{(i)}} n_{A_{(i)}}}{N_{(i)}} \right], \end{aligned}$$

and

$$\frac{G}{\sqrt{\sum_{(i)} \frac{N_{(i)}^2 n_{A_{(i)}} n_{B_{(i)}} m_{D_{(i)}} m_{\bar{D}_{(i)}}}{N_{(i)}^2 (N_{(i)} - 1)}}} \overset{\text{approx}}{\underset{n \text{ large}}{\sim}} N(0,1)$$