EMORY ROLLINS SCHOOL OF PUBLIC HEALTH | Department of Biostatistics and Bioinformatics

BIOS 522: Survival Analysis Methods

Reading 3:

The log-rank test

This week, we will study the log-rank test for comparing survival curves. We will also learn about the weighted log-rank and stratified log-rank tests. We will demonstrate how to implement these tests in R.

Part 1. The log-rank test

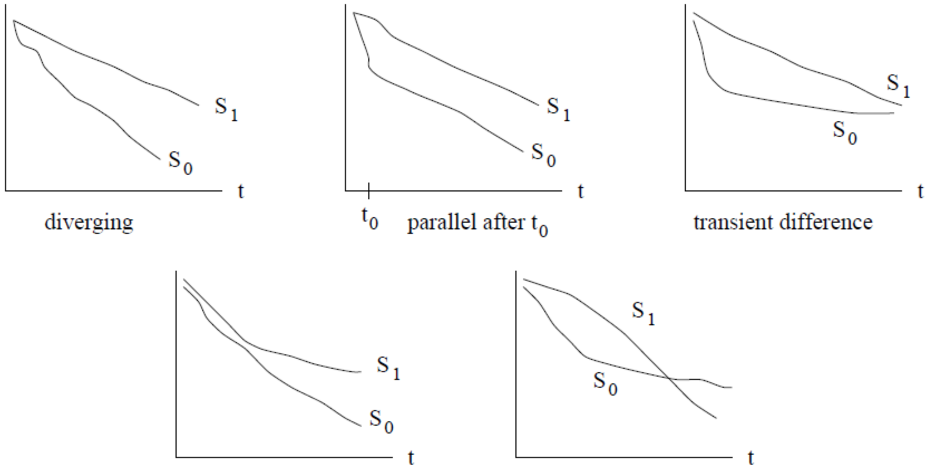
Evaluating differences across survival

Last week, we learned how to use the Kaplan-Meier estimator to estimate the survival function of a given population. Often, we are interested in comparing survival across multiple, independent populations, to assess which group has better survival. To compare groups, we can plot the survival functions and make a visual assessment of any differences, but this is not a formal statistical test.

Consider two independent groups: a group with covariate $X = 1$ and a group with covariate $X = 0$. The null hypothesis of interest is that the survival curve for group 1 is equal to the survival curve in group 0:

$$H_0: S_1(\cdot) = S_0(\cdot)$$

The null hypothesis above can be read as indicating that $S_1(t) = S_0(t)$ for any value of t. There are many ways in which $S_1(\cdot)$ and $S_0(\cdot)$ can differ:



diverging

parallel after t_0

transient difference

late emerging difference

not stochastically ordered

As we examine the shapes above, we can envision many different strategies for comparing the curves:

- Measure the largest difference between the two curves
- Compare the median survival for each group
- Add up the differences between the two survival estimates over time
- Compare the rates of failure across the two groups over time
- And so on…

The logic of the log-rank test

The **log-rank test** is the most commonly used statistical test for comparing the survival functions of two or more independent groups. At each distinct failure time, we compare the failure rates between the two groups. It considers the magnitude and direction of this difference. The differences are summed over time. If we tend to see a higher rate of failure in one group versus the other across all time points, this will be captured by the test.

The log-rank test is a nonparametric test whose validity does not depend on any parametric assumptions. Nonetheless, the log-rank test is not well suited for the setting where the survival curves cross and the direction of the difference changes (e.g., “not stochastically ordered” example in the previous figure). In that case, the positive differences can cancel out the negative differences. The test is best suited for the setting where one group has consistently better (or worse) survival than the other.

Calculating the log-rank test statistic

To calculate the log-rank test statistic and associated p-value, we construct a 2x2 table at each unique failure time in the data set.

Let $t_1 < t_2 < \cdots < t_J$ be the distinct times where failures occur in either group. We use the following notation:

- n_{0j} is the number at risk in group 0 at time t_j
- n_{1j} is the number at risk in group 1 at time t_j
- d_{0j} is the number of failures in group 0 at time t_j
- d_{1j} is the number of failures in group 1 at time t_j

The information at time t_j can be summarized in the following table, where the shaded part is our 2x2 table:

	At risk just before t_j	Fails at t_j	Survives past t_j
Group 0	n_{0j}	d_{0j}	$n_{0j} - d_{0j}$
Group 1	n_{1j}	d_{1j}	$n_{1j} - d_{1j}$
Total	$n_j = n_{0j} + n_{1j}$	$d_j = d_{0j} + d_{1j}$	$n_j - d_j$

At time t_j, there were d_j total failures across the two groups. We observed d_{0j} of these failures in group 0. In the log-rank test, we characterize if d_{0j} is higher (or lower) than would be expected assuming that the two groups have the same rate of failure. For example, if there are equal numbers at risk in both groups, and the groups had the same rate of failure, we would expect to see the same number of failures in both groups ($d_{0j} = d_{1j}$). If there are more people at risk in one group, but the groups have the same rate of failure, we would expect to see proportionally more failures in the larger group.

Under H_0, the expected number of failures in group 0 at time t_j is:

$$E_j = \frac{n_{0j}}{n_j} d_j$$

This is the proportion of people at risk who are in group 0, multiplied by the total number of events at that time.

In the log-rank test statistic, we calculate the difference between the observed number of failures in group 0 with its expected value E_j under the null hypothesis. The observed value is $O_j = d_{0j}$.

We standardize the difference between O_j and E_j by the variance V_j under H_0:

$$V_j = \frac{n_{0j} n_{1j} d_j \left(n_j - d_j\right)}{n_j^2 (n_j - 1)}$$

We sum O_j, E_j, and V_j over the J distinct failure times $t_1, \ldots, t_J$, yielding the following terms:

$$O = \sum\nolimits_{j=1}^{J} O_j$$

$$E = \sum_{j=1}^{J} E_j$$

$$V = \sum_{j=1}^{J} V_j$$

The log-rank test statistic Z is calculated as:

$$Z = \frac{O - E}{\sqrt{V}}$$

We reject the null hypothesis for large values of the test statistic.

The log-rank test statistic Z approximately follows a standard normal distribution, so we can use this fact to calculate an associated p-value. The two-sided p-value will be less than 0.05 if $Z < -1.96$ or $Z > 1.96$.

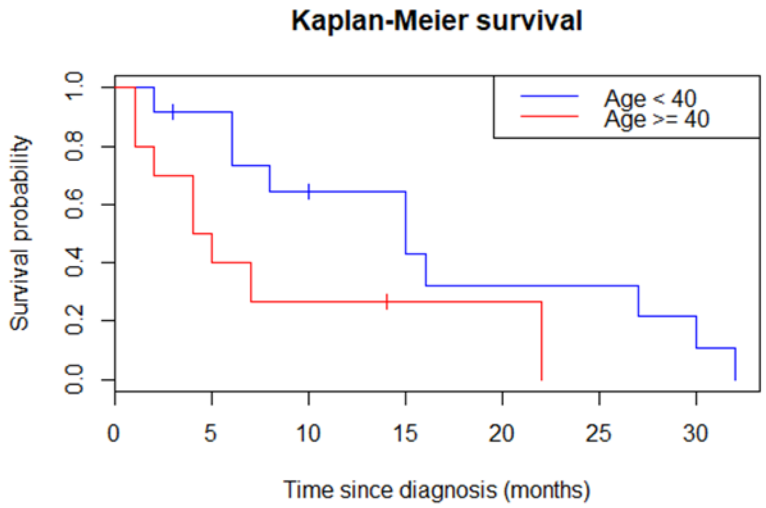
Sometimes the log-rank test statistic is reported as Z^2, which is compared to a chi-squared distribution with one degree of freedom. The two-sided p-value will be less than 0.05 if $Z^2 > 3.84$.

Example: *We examine times from primary AIDS diagnosis until death for hemophiliacs. We wish to compare survival for patients who were at most 40 years of age at the time of diagnosis to survival for hemophiliacs who were over the age of 40 at diagnosis.*

Consider the right-censored survival times (in months) for the two age groups:

Group 0 (Age <40):	2, 3+, 6, 6, 8, 10+, 15, 15, 16, 27, 30, 32
Group 1 (Age ≥ 40):	1, 1, 2, 4, 4, 5, 5+, 7, 14+, 22

Group-specific Kaplan-Meier curves are plotted below:



It appears that individuals who are younger at the time of AIDS diagnosis survive longer than those who are older at the time of diagnosis. Is the difference greater than might be expected by chance alone?

To calculate the log-rank test statistic, we first identify distinct *failure* times. (Note, this does not include censoring only times.) There are $J = 13$ times:

1, 2, 4, 5, 6, 7, 8, 15, 16, 22, 27, 30, 32

Starting with $t_1 = 1$:

	Fail	Survive	At risk
Group 0	0	12	12
Group 1	2	8	10
	2	20	22

$$O_1 = 0$$

$$E_1 = \left(\frac{12}{22}\right)(2) = 1.091$$

$$V_1 = \frac{(12)(10)(2)(20)}{22^2(22-1)} = 0.472$$

At time $t_1 = 1$, the observed number of failures in group 0 is lower than the expected number of failures in group 0 under H_0.

For $t_2 = 2$:

	Fail	Survive	At risk
Group 0	1	11	12
Group 1	1	7	8
	2	18	20

$$O_2 = 1$$

$$E_2 = \left(\frac{12}{20}\right)2 = 1.2$$

$$V_2 = \frac{(12)(8)2(18)}{20^2(20-1)} = 0.455$$

The observed number of failures in group 0 is (slightly) lower than the expected number of failures in group 0 under H_0.

The next distinct failure time is $t_3 = 4$. Note that one person in group 0 was censored at 3 months, so they are not at risk at time $t_3 = 4$.

	Fail	Survive	At risk
Group 0	0	10	10
Group 1	2	5	7
	2	15	17

$$O_3 = 0$$

$$E_3 = \left(\frac{10}{17}\right)2 = 1.176$$

$$V_3 = \frac{(10)(7)2(15)}{17^2(17-1)} = 0.454$$

The observed number of failures in group 0 is lower than the expected number of failures in group 0 under H_0. At time $t_3 = 4$, the failure rate is lower in group 0 than group 1.

We continue forming tables. Let's skip ahead to $t_{10} = 22$:

	Fail	Survive	At risk
Group 0	0	3	3
Group 1	1	0	1
	1	3	4

$$O_{10} = 0$$

$$E_{10} = \left(\frac{3}{4}\right)1 = 0.75$$

$$V_{10} = \frac{(3)(1)1(3)}{4^2(4-1)} = 0.188$$

When we proceed to $t_{11} = 27$, note that no one from group 1 is still at risk. All 10 participants have either been censored or failed.

	Fail	Survive	At risk
Group 0	1	2	3
Group 1	0	0	0
	1	2	3

$$O_{11} = 1$$

$$E_{11} = \left(\frac{3}{3}\right)1 = 1$$

$$V_{11} = \frac{(3)(0)1(3)}{3^2(3-1)} = 0$$

Because no one is at risk in group 1, the observed number of failures in group 0 is equal to the expected number of failures. These quantities will always be equal, and therefore these tables contribute no new information about group differences to the log-rank test. Thus, once one of the groups is empty, we do not need to construct additional tables.

We then sum across the failure times to calculate the log-rank test statistic:

$$O = \sum_{j=1}^{10} O_j = (0 + 1 + 0 + \cdots + 0) = 10$$
$$E = \sum_{j=1}^{10} E_j = (1.091 + 1.2 + 1.176 + \cdots + 0.75) = 13.46$$
$$V = \sum_{j=1}^{10} V_j = (0.472 + 0.455 + 0.454 + \cdots + 0.188) = 2.83$$

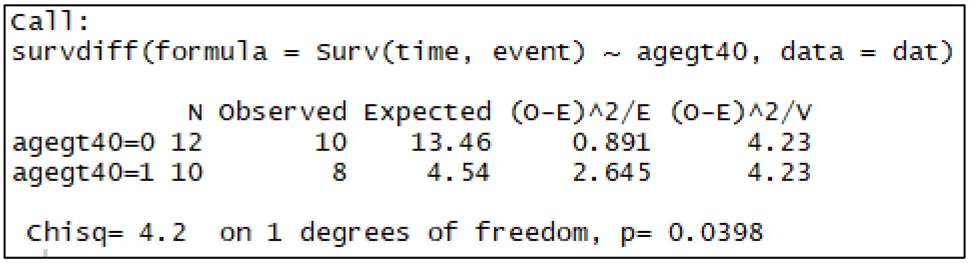
$$Z = \frac{O - E}{\sqrt{V}} = \frac{10 - 13.46}{\sqrt{2.83}} = -2.057$$

The associated two-sided p-value for this test statistic Z is 0.0398.

Alternatively, we can use the formula for a chi-square test statistic:

$$Z^2 = \frac{(O - E)^2}{V} = \frac{(10 - 13.46)^2}{2.83} = 4.23$$

In this week's activity, we will learn how to calculate the log-rank test in R. Until then, we can recognize the elements we calculated by hand in R output. In the results, find the group sample sizes (12 and 10), the observed value for group 0 (10), expected value (13.46), test statistic (4.23), and p-value (0.0398).



```
Call:
survdiff(formula = Surv(time, event) ~ agegt40, data = dat)

           N Observed Expected (O-E)^2/E (O-E)^2/V
agegt40=0 12       10    13.46     0.891      4.23
agegt40=1 10        8     4.54     2.645      4.23

 Chisq= 4.2  on 1 degrees of freedom, p= 0.0398
```

We conclude that there is a significant difference in survival between the two groups. Survival is significantly lower in hemophiliac patients who are older than 40 at the time of primary AIDS diagnosis when compared to patients who are younger than 40.

Note that the value of the test statistic (4.23) is the same regardless of which group we choose to construct our test.

The log-rank test can be extended to compare more than two independent samples. In this case, it tests the null hypothesis that all groups have the same failure time distribution.

Part 2. Weighted log-rank test

In the log-rank test, all time points are given equal weight, even though the population at-risk is largest at earlier time points. We can select weights to give some time points more weight than others.

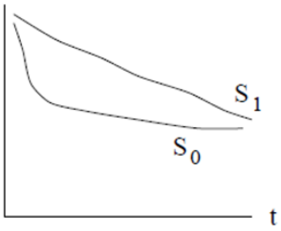
Let $w_1 \geq 0$, $w_2 \geq 0$, …, $w_J \geq 0$ be known constants (weights). Then the **weighted log-rank test** is given by:

$$Z_w = \frac{\sum_{j=1}^{J} w_j(O_j - E_j)}{\sqrt{\sum_{j=1}^{J} w_j^2 V_j}}$$

Under H_0, Z_w approximately follows a standard normal distribution. We reject the null hypothesis for large values of the test statistic. The two-sided p-value will be less than 0.05 if $Z_w < -1.96$ or $Z_w > 1.96$, or, where a chi-squared test statistic is reported, if $Z_w^2 > 3.84$.

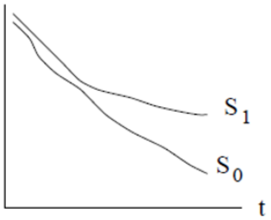
Selecting constant weights $W_j = w$ yields the standard log-rank test. Choosing $W_j = n_j$ yields the **generalized Wilcoxon test**. Since $n_1 > n_2 > \cdots$, the generalized Wilcoxon test places (relatively) greater emphasis on early differences between $S_0(\cdot)$ and $S_1(\cdot)$ than the log-rank test. The generalized Wilcoxon test is also sometimes referred to as the **Gehan-Breslow test**.

Consider a transient difference where the group difference is large early on but then decreases over time. We would expect the generalized Wilcoxon test statistic to be *larger* than the standard log-rank test statistic.



transient difference

Consider a late emerging difference where the group difference is small early on but then increases over time. We would expect the generalized Wilcoxon test statistic to be *smaller* than the standard log-rank test statistic.

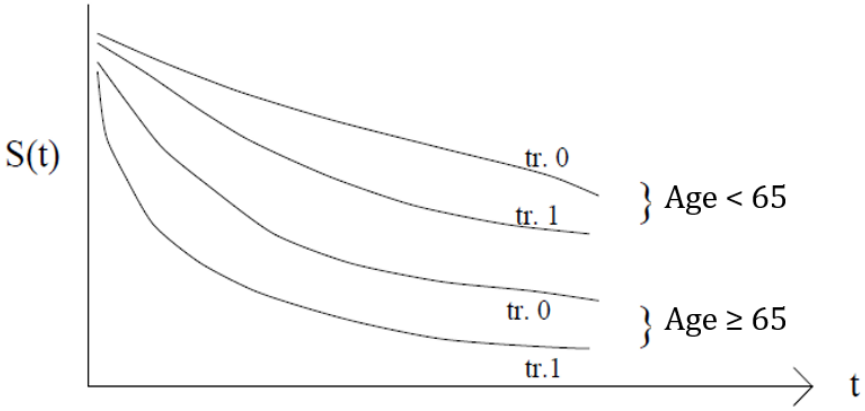


late emerging difference

Other weighted log-rank tests include the Tarone-Ware test, the Peto-Prentice test, the Efron test, and the Harrington-Fleming test. Each test uses a different weighting scheme and so yields a different test statistic.

Part 3. Stratified log-rank test

Suppose that we want to compare two groups, such as two treatments, but we also want to control (adjust) for a categorical covariate or confounder, such as age ≥65 or <65 years. In this example with two treatments and a binary age category, 4=2x2 types of individuals are monitored, and their respective survivor functions might be as shown below:



If we want to compare treatment groups but also 'adjust' for age group, a stratified log-rank test could be used. In this case, the null hypothesis is that the survival functions of the two groups of interest (e.g. treated and untreated) are the same *within each stratum* (e.g. <65 years, ≥65 years). Imagine the population is divided into $k = 1, \dots, K$ non-overlapping strata. The null hypothesis is:

$$H_0 \colon S_0^{(k)}(\cdot) = S_1^{(k)}(\cdot)$$

where $S_1^{(k)}$ is the survivor function for the treated group in stratum k.

To calculate the stratified log-rank test statistic, we calculate the observed and expected number of failures and the variance separately in each stratum and then combine these into a single test statistic. Within each stratum k, calculate the following quantities using the same strategy described for the standard log-rank test:

$$O_k = \sum_{j=1}^{J} O_{kj}$$

$$E_k = \sum_{j=1}^{J} E_{kj}$$

$$V_k = \sum_{j=1}^{J} V_{kj}$$

The stratified log-rank test statistic Z_S is calculated as:

$$Z_S = \frac{\sum_{k=1}^{K} O_k - \sum_{k=1}^{K} E_k}{\sqrt{\sum_{k=1}^{K} V_k}}$$

Under H_0, Z_S approximately follows a standard normal distribution. We reject the null hypothesis for large values of the test statistic. The two-sided p-value will be less than 0.05 if $Z_S < -1.96$ or $Z_S > 1.96$, or, where a chi-squared test statistic is reported, if $Z_S^2 > 3.84$.

The stratified log-rank test does not assume that the survival functions in different strata are the same, which is important because there might be differences in the shape of the survival function across strata (here, age groups). Instead we are estimating the treatment differences *within* each stratum and then pooling these differences *across* strata.

The stratified log-rank test is useful for two major reasons:

(1) Including a categorical covariate allows us to "explain" some of the variability observed between individuals. For example, in the above figure, older individuals have poorer survival than younger individuals. Looking within groups allows us to more precisely isolate the effect of treatment.

(2) There may be confounding. For example, older individuals are likely to have poorer survival, but there may be relatively more older individuals on treatment than younger individuals on treatment. When we do not account for this type of imbalance, it can lead to bias or, in extreme cases, reverse the direction of an effect. By looking within groups, we address these imbalances.

Part 4. Looking ahead

So far, we have discussed statistical inference for the survival function – a key function in survival analysis, but not the only one we are interested in. Next week, we will introduce the **hazard** and **cumulative hazard functions**, which help form the basis of our future regression models.