

## Log-Rank Test for More Than Two Groups

Prepared by Harlan Sayles (SRAM)  
Revised by Julia Soulakova (Statistics)

The Log-Rank test for more than two groups allows for comparing rates of events of interest (e.g., death, remission from cancer, machine part failure) between more than two distinct groups. A population of interest may be cancer patients who undergo certain cancer treatments, a batch of machine parts, or criminals recently released from prison. The corresponding events of interest may be death, part failure, or recidivism (arrested for a new crime) in each of these respective cases. The groups within these populations may be formed from the race of the patient for the persons who have undergone cancer treatments, the aluminum alloy composition of the different machine parts to be tested, and the socioeconomic status of the recently released prisoners. The Log-Rank test can be used to answer questions similar to the following:

1. Are death rates for persons who have undergone an experimental treatment for colon cancer different based on a person's race (e.g., compare Blacks, Whites, Hispanics, and others)?
2. Does the rate or failure for machine parts differ based upon the composition of the aluminum alloy that they are made from (e.g., compare parts that are identical in size and function, but made from different metals)?
3. Do recently released criminals with low socioeconomic status tend to have higher recidivism rates than criminals with medium or high socioeconomic status?

### *Data Considerations*

The Log-Rank test allows for two types of non-perfect survival data, left-truncated data and right-censored data. Ideally, all subjects would be observed starting at time 0 and each subject would experience the event of interest prior to the conclusion of the study period. For practical reasons, however, this is not always possible.

For left-truncated data, the event of interest happens prior to the beginning of the study period and because of this, the subject is not included in the study. For example, if a patient dies of colon cancer prior to receiving the experimental treatment, he can not be included in the study and we say that our data are left-truncated. For right-censored data,

the subject's event of interest does not occur during the study period or before it, but the subject is at least observed for the duration of the study period, so some information is known about the subject's survival. For the example of the machine parts, if the time allowed for a researcher to complete the study is limited to six months, and at the end of six months some parts have not yet failed, then the researcher must end the test and accept that the failure time for the remaining parts is greater than six months (assuming that all parts will eventually fail). The time data corresponding to the parts that do not fail within six months are said to be right-censored because it is known that the failure time is greater than six months, but exactly how much greater is not known.

Several variables are required in order to perform a Log-Rank test for a comparison between more than two groups. First, the dataset must contain an event indicator variable as it is necessary in all survival analysis experiments. For ease of use, the indicator should be coded as a "1" if the event happens within the observation period, or a "0" if the event does not happen within the observation period, i.e., the case is right-censored. The data must also include a variable representing the amount of time on study. This variable should correspond to the time from the beginning of the observational period until either the event of interest (event indicator = 1) or censoring (event indicator = 0). Finally, the dataset must include a group indicator variable to indicate which comparison group each subject belongs to. The group indicator variable may be either numeric (e.g. 1, 2, 3, 4) or a string (e.g. "Black", "White", "Hispanic", "Other"). Alternatively, if the group variable is continuous and needs to be converted to a discrete variable, this can be done with the LIFETEST procedure in SAS by specifying a list of intervals following the variable name in the SAS syntax (see the SAS Help and Documentation for details).

The test statistics for the Log-Rank test are based on large-sample approximations. Larger sample sizes will lead to more reliable results, as is the case with many statistical procedures. The number of comparison groups should not be allowed to get too large to avoid having comparison groups with too few subjects. Each group should contain at least 30 subjects (and preferably more) for best results.

### *Main Ideas*

The basic concept behind the Log-Rank test is to allow for a test of the null hypothesis that the hazard functions for all groups are equal for all study time versus the alternative hypothesis that at least one hazard function is different from the others for some time during the study. To put it into mathematical terms for  $k$  groups

$$H_0 : h_1(t) = h_2(t) = \dots = h_k(t) \text{ for all } t \leq \tau$$

$$H_A : \text{at least one of the } h_j(t)\text{'s is different for some } t \leq \tau$$

where  $\tau$  is the largest time during which each group has at least one subject at risk (at least one subject in each group has not experienced the event of interest or been censored).

The Log-Rank test will compare the hazards from each group at each event time between 0 and  $\tau$ . Each event time is defined as each unique time when any subject from

any group experiences the event of interest, regardless of whether any other subject in any of the other groups experiences the event or not. The total number of unique event times in the population is represented by  $D$ . If all hazards are equal for all groups, then it is assumed that the proportion of each group experiencing the event at any given event time  $t_i$  will be equal to the proportion of the overall population experiencing the event at that same time. If  $d_{ij}$  is the number of events experienced by group  $j$  at event time  $t_i$ ,  $Y_{ij}$  is the number of persons at risk in group  $j$  just prior to time  $t_i$ ,  $d_i$  is the total number of events experienced by the entire study population at event time  $t_i$ , and  $Y_i$  is the total number of persons at risk in the entire study population just prior to time  $t_i$ , then it is assumed that

$$d_{ij}/Y_{ij} = d_i / Y_i \text{ for all event times } t_i, \quad i = 1, 2, \dots, D, \quad j = 1, 2, \dots, k$$

The Log-Rank test begins by calculating a statistic representing the sum of the weighted differences between  $d_{ij}/Y_{ij}$  and  $d_i/Y_i$  at each event time  $t_i$  for each group  $j = 1$  through  $k$ . For the Log-Rank test, the weights applied to these differences are all equal to 1, so each event time has an equal weighting on the value of the statistics. Variances and covariances for each of these  $k$  statistics are also calculated using slightly more advanced formulas (see Klein and Moeschberger, 2003, pg 207).

The statistics calculated for the  $k$  groups are linearly dependent and therefore, we may use only  $k-1$  of them to calculate a test statistic. To calculate the test statistic,  $k-1$  of the statistics are formed into a vector called  $\mathbf{Z}$ . The variances and covariances for these  $k-1$  statistics are placed into a variance-covariance matrix called  $\mathbf{\Sigma}$  (again, see Klein and Moeschberger, 2003, pg 207 for details). A test statistic is then calculated by equation 1

$$\chi^2 = \mathbf{Z}(\mathbf{\Sigma}^{-1})\mathbf{Z}^t \quad (1)$$

which has a chi-squared distribution with  $k-1$  degrees of freedom when the null hypothesis is true.

### *An Example*

The data for this example are empirical. They are for illustration purposes only! The data set is very small to allow for simplicity of illustration, and consists of three variables representing group membership (Race), time in the study in months (Months), and an event indicator (Status) that equals “1” if the person has experienced the event and “0” if the case is right-censored. Assume that this data represents a group of 10 persons who have undergone an experimental treatment for colon cancer. The data are shown in Table 1 below.

The SAS code below can be used to create the data set and perform the analyses. The “Data” command simply creates the data set. The “Proc Sort” command is necessary to sort the data set by the grouping variable prior to running the “Proc Lifetest” command. SAS will expect the data to be sorted by the grouping variable and may not function properly if the data are not sorted.

**Table 1.** Fictitious Colon Cancer Patient Data

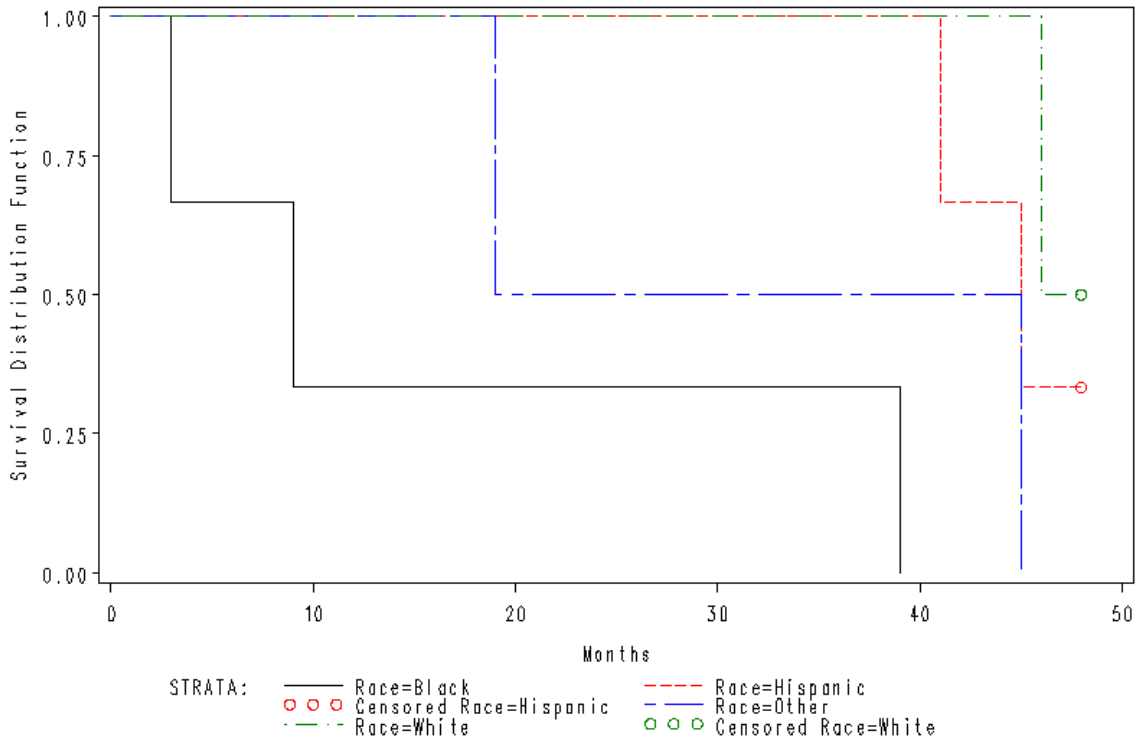
Race	Months	Status
Black	9	1
Black	15	1
Other	17	1
Black	20	1
Other	24	1
Hispanic	25	1
Hispanic	31	1
White	40	1
Hispanic	48	0
White	48	0

The “Proc Lifetest” command uses the Product-Limit estimator, by default, to estimate the survival curves (Figure 1). Unfortunately, this method does not allow for a graph of the hazard functions for each group to be printed directly, but it does give output for the Log-Rank test. The Log-Rank test output is requested as an option in the “strata” statement with the syntax “test=(Logrank)”. Incidentally, the Log-Rank test is the default test output for the “strata” statement, along with a couple of other tests, so the results of the Logrank test would be included in the output even if the “test=(Logrank)” option were left out of the syntax, but by including this option, only the Log-Rank results are given. SAS output for the Lifetest Procedure is presented after the figure.

### SAS Syntax

```
Data Cancer;  
input Race $ Months Status;  
datalines;  
Black 3 1  
Black 9 1  
Other 19 1  
Black 39 1  
Hispanic 41 1  
Other 45 1  
Hispanic 45 1  
White 46 1  
Hispanic 48 0  
White 48 0  
run;  
  
Proc Sort data=work.Cancer;  
By Race;  
Run;  
  
Proc Lifetest data=work.Cancer method=PL plots=(s(name=Survival))  
graphics;  
time Months*Status(0);  
strata Race / test=(Logrank);  
symbol1 v=none color=black line=1;  
symbol2 v=none color=red line=3;  
symbol3 v=none color=blue line=27;  
symbol4 v=none color=green line=41;  
run;
```

**Figure 1.** Survival Graphs for “method=PL”



**SAS Output**

The LIFETEST Procedure

Stratum 1: Race = Black  
Product-Limit Survival Estimates

Months	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	3
3.0000	0.6667	0.3333	0.2722	1	2
9.0000	0.3333	0.6667	0.2722	2	1
39.0000	0	1.0000	0	3	0

Summary Statistics for Time Variable Months

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval [Lower Upper)		
75	39.0000	3.0000	39.0000	
50	9.0000	3.0000	39.0000	
25	3.0000	3.0000	39.0000	
Mean		Standard Error		
	17.0000	11.1355		

Stratum 2: Race = Hispanic  
Product-Limit Survival Estimates

Months	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	3
41.0000	0.6667	0.3333	0.2722	1	2
45.0000	0.3333	0.6667	0.2722	2	1
48.0000*	.	.	.	2	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable Months  
Quartile Estimates

Percent	Point	95% Confidence Interval	
	Estimate	[Lower	Upper)
75	.	41.0000	.
50	45.0000	41.0000	.
25	41.0000	41.0000	.

Mean	Standard Error
43.6667	1.5396

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Stratum 3: Race = Other  
Product-Limit Survival Estimates

Months	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	2
19.0000	0.5000	0.5000	0.3536	1	1
45.0000	0	1.0000	0	2	0

Summary Statistics for Time Variable Months  
Quartile Estimates

Percent	Point	95% Confidence Interval	
	Estimate	[Lower	Upper)
75	45.0000	19.0000	45.0000
50	32.0000	19.0000	45.0000
25	19.0000	19.0000	45.0000

Mean	Standard Error
32.0000	13.0000

Stratum 4: Race = White  
Product-Limit Survival Estimates

Months	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	2
46.0000	0.5000	0.5000	0.3536	1	1
48.0000*	.	.	.	1	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable Months

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower Upper)	
75	.	46.0000	.
50	.	46.0000	.
25	46.0000	46.0000	.

Mean	Standard Error
46.0000	.

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

Stratum	Race	Total	Failed	Censored	Percent Censored
1	Black	3	3	0	0.00
2	Hispanic	3	2	1	33.33
3	Other	2	2	0	0.00
4	White	2	1	1	50.00
-----					
Total		10	8	2	20.00

Testing Homogeneity of Survival Curves for Months over Strata

Rank Statistics

Race	Log-Rank
Black	2.2099
Hispanic	-1.0702
Other	0.6183
White	-1.7579

Covariance Matrix for the Log-Rank Statistics

Race	Black	Hispanic	Other	White
Black	0.61466	-0.27217	-0.16104	-0.18145
Hispanic	-0.27217	1.74372	-0.49238	-0.97916
Other	-0.16104	-0.49238	1.02168	-0.36825
White	-0.18145	-0.97916	-0.36825	1.52887

Test of Equality over Strata

Test	Chi-Square	DF	Pr >
			Chi-Square
Log-Rank	9.6836	3	0.0215

The output shows results for each value (stratum) of the group identification variable, which is Race in this example. The values listed under “Months” for each stratum indicate event times or censoring times for cases within that stratum. If a time is a censoring time, it is marked with an asterisk. The next column labeled “Survival” is the Product-Limit Estimator of the survival function at that time point. The third column, labeled “Failure” gives the percent of cases in the stratum that have failed (experienced the event) up to that time point. Failure is equal to 1-Survival. The fourth column is the standard error of the Product-Limit survival estimation. The fifth and sixth columns are the total number in the stratum that have failed and the total number in the stratum that have not failed or been censored up to a given time point, respectively.

The next part of the output for each stratum calculates estimates of survival percentiles. These numbers for the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentiles are the estimates at which 25%, 50%, and 75% of the subjects in the stratum have NOT survived. The standard errors for these estimates are also presented. These estimates are fairly meaningless with such a small sample size, but can be somewhat useful with a more appropriately sized group.

The last statistic reported for each stratum is the mean survival time; that is the average amount of time that a subject in a given stratum is expected to survive. The standard error for this estimate is also given. This number is usually different from the 50% percentile, which is a median survival time as opposed to a mean survival time.

The “notes” under stratum 2 and 4 (Hispanic and White) indicate the statistical problem of trying to calculate an average when not all of the values are known (i.e. the highest one is censored). The estimates of mean survival time are calculated using all of the known values, but they are certain to be underestimates, because the censored survival times must be larger than the largest known survival times and when these larger values are added into the calculation, the calculated mean will necessarily increase.

Following the individual strata statistics is a summary table that presents the total number of subjects, failures, censored cases, and the percent of censored cases for each stratum. This information is redundant with information presented above for each individual stratum and is given only for convenience.

The final part of the output labeled “Testing Homogeneity of Survival Curves for Months over Strata” contains all of the information necessary for performing the Log-Rank test and also the result of the test. The first part, labeled “Rank Statistics” are the

sum of the weighted differences between the observed hazards  $d_{ij}/Y_{ij}$  for each of the four racial groups and the population hazard  $d_i/Y_i$  at each distinct event time. The second part is the variance-covariance matrix for the rank statistics, calculated using the equations on page 207 of Klein and Moeschberger (2003). Finally, using equation 1 from above, the test statistic for the Log-Rank test is calculated to be 9.6386, which is distributed as a chi-squared random variable with 3 degrees of freedom. The p-value for this statistic is 0.0215, which is less than the standard alpha cutoff value of 0.5, so the test has rejected the null hypothesis that all strata hazards are equal for all  $t \leq \tau$ .

The graph in Figure 1 shows that there is some evidence of differences between the survival functions across the four strata; early on, the survival for Blacks is much lower than the one for other three groups. Unfortunately, the SAS Lifetest Procedure with the Product-Limit estimator does not allow for a plotting of the hazard curves for each stratum. Such a graph would likely not be very useful anyways, as the hazard function will be equal to zero for all strata at times other than the  $D$  distinct event times. The Log-Rank test provides evidence that at least one of the four stratum hazard plots is significantly different from the others for some value of  $t \leq \tau$ , meaning that there is an effect of race on subject survival time. Hence, there is a significant difference among the survival rates of White, Black, Hispanic and others.