LOG-RANK TEST

Alan Hopkins
Theravance Inc.,
South San Francisco, California

Calculation of the log-rank test is described, and an example is provided. Conditions under which the test is valid are discussed and testing of assumptions is addressed. Generalizations of the log-rank test are described including its relationship to the Cox regression model. Sample size calculation for the log-rank test is discussed, and computing software is described.

Suppose T is a continuous, non-negative random variable representing survival times from point of randomization. The distribution of censored event times can be estimated by the survival function $S(t)$ The survival function estimates the probability of surviving to time t so $S(t) = \Pr(T > t)$. The hazard function is directly related to the survival function through the relationship $\lambda(t) = -d \log S(t)/dt$. The hazard rate is the conditional probability of death in a small interval $[t, t + dt)$ given survival to the beginning of the interval. With censored data, we cannot observe events on all subjects, so we only observe a censoring time. Thus, for each subject, we have data on observation time, an indicator whether an event was observed at the last observation time, and treatment group. With this information, we can compare hazard functions in intervals containing observed death times and calculate a global test statistic for comparing the observed survival distributions among several groups. The log-rank test was originally proposed by Mantel (1), and it is equivalent under certain circumstances to the Cox (2) regression model.

0.0.1 Hypothesis. Suppose we have K groups and we denote the survival functions associated with each group as $S_1(t), \ldots, S_K(t)$. The null hypothesis can be expressed as

$$H_0 : S_1(t) = S_2(t) = \ldots = S_K(t), \text{for } t \geq 0$$

The null hypothesis can also be stated using the hazard function. The null hypothesis is equivalent to comparing the hazard rates

$$H_0 : \lambda_1(t) = \lambda_2(t) = \ldots = \lambda_K(t), \text{for } t \geq 0 \quad (1)$$

The alternative hypothesis usually of interest is that the survival function for one group is stochastically larger or smaller than the survival functions for the other groups.

$H_a : S_k \geq S_{k'}(t)$, or $S_k(t) \leq S_{k'}(t)$, for some k, k' with strict inequality for some t.

0.0.2 Assumptions. The log-rank test is nonparametric in the sense that no underlying probability distribution is assumed for the survival function. The log-rank test assumes the censoring process is unrelated to survival times or to the treatment groups themselves (independent censoring) and that the survival times are from the same distribution for subjects recruited early or late in the clinical trial (stationarity). Observations are assumed to be independent. Special methods are required if recurrent events are observed on a single individual (e.g., multiple infections.)

0.0.3 Inference. Let $t_1 < t_2 < \cdots < t_L$ represent the ordered distinct failure times in the combined groups of observations. At time t_t, d_{ij} events are observed in the jth sample out of R_{ij} individuals at risk just prior to t_i. Here $d_i = \sum_{j=1}^{K} d_{ij}$ represents the total number of deaths at t_i and $R_i = \sum_{j=1}^{K} R_{ij}$ the number of subjects at risk at t_i. We can represent the data at time t_i as shown in Table 1.

The test of hypothesis (1) is based on weighted deviations of the estimated hazard functions for each group from the overall estimated hazard rate among all data combined. If the null hypothesis is true, then an estimator of the expected hazard rate in the jth population under H_0 is the pooled sample estimator of the hazard rate d_i/R_i. An estimate of the hazard rate for the jth sample

Wiley Encyclopedia of Clinical Trials, Copyright © 2008 John Wiley & Sons, Inc.

Table 1. Layout and Notation for the K-Group Log-Rank Test

Time t_i	Group 1	Group 2	...	Group K	At Risk
Deaths	d_{i1}	d_{i2}	...	d_{iK}	d_i
Survivors	$R_{i1} - d_{i1}$	$R_{i2} - d_{i2}$	...	$R_{iK} - d_{iK}$	$R_i - d_i$
Total	R_{i1}	R_{i2}	...	R_{iK}	R_i

is d_{ij}/R_{ij}. To compare survival distributions, we take a weighted average of deviations across all failure times. The test is based on statistics of the form:

$$v_j = \sum_{i=1}^{L} W(t_i) \left\{ d_{ij} - R_{ij} \left(\frac{d_i}{R_i} \right) \right\}, \; j = 1, \ldots, K$$

where $W(t)$ is a positive weight function. $v_j(t)$ is the sum over all event times of the difference in observed and conditionally expected events for group j. This quantity has a product hypermultinomial distribution with covariance matrix:

$$\mathbf{V}_{jg} = \sum_{i=1}^{L} W(t_i)^2 \frac{R_{ij}}{R_i} \left(\delta_{jg} - \frac{R_{ig}}{R_i} \right) \left(\frac{R_i - d_i}{R_i - 1} \right) d_i$$
$$j, g = 1, \ldots, K$$

where $\delta_{jg} = 1$ when $j = g$ and 0 otherwise. Let $\mathbf{v} = (v_1, v_2, \ldots, v_K)^T$. Then a test statistic for hypothesis (1) is the quadratic form

$$X^2 = \mathbf{v}^{\mathbf{T}} \mathbf{V}^{-} \mathbf{v} \tag{2}$$

where $\mathbf{V}^{-}$ is a generalized inverse. The components of $\mathbf{v}$ are linearly dependent and sum to zero so the variance-covariance matrix has maximum rank $K - 1$. The overall test statistic can be constructed using any $K - 1$ components of $\mathbf{v}$ and corresponding elements of the variance-covariance matrix. Therefore, if the last row and column of $\mathbf{V}$ is deleted to give $\mathbf{V}_{K-1}$ and $\mathbf{v}_{K-1} = (v_1, v_2, \ldots, v_{K-1})^T$, then the overall log-rank test statistic is

$$X^2 = \mathbf{v}_{K-1}^T \mathbf{V}_{K-1}^{-1} \mathbf{v}_{K-1} \tag{3}$$

where $\mathbf{V}_{K-1}{}^{-1}$ is an ordinary inverse. The distribution of the weighted log-rank statistic is chi-squared with $K - 1$ degrees of freedom. Using $W(t_i) = 1$ gives the widely used log-rank test. Alternative weights will be discussed in a later section.

Since the log-rank statistic as presented here sums across multiple failure times, the tables used are not independent, which precludes use of standard methods to derive the asymptotic distribution of the statistic. The asymptotic properties of the log-rank test were rigorously developed using counting process techniques. For details of this approach, see Fleming and Harrington (3) or Kalbfleisch and Prentice (4).

0.0.4 A Special Case ($K = 2$) and $W(t_i) = 1$. Often a clinical trial consists of only two treatment groups. In this case, the computations are simplified. We may write the two-sample log-rank test as

$$Z_{LR} = \frac{\sum_{i=1}^{L} \left\{ d_{i1} - R_{i1} \left(\frac{d_i}{R_i} \right) \right\}}{\sqrt{\sum_{i=1}^{L} \frac{R_{i1}}{R_i} \left(1 - \frac{R_{i1}}{R_i} \right) \left(\frac{R_i - d_i}{R_i - 1} \right) d_i}}$$

which has approximately a standard normal distribution under the null hypothesis for large samples.

0.0.5 Relationship of the Log-rank Statistic to the Cox Regression Model. The log-rank test is closely related to the Cox proportional hazards regression model. Let $\mathbf{z}^T = (z_1, \ldots, z_p)$ represent p covariates on a given subject. In the case of the log-rank test, $\mathbf{z}$ would be indicator variables for treatment groups. The proportional hazards regression model is $\lambda(t|\mathbf{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^{\mathbf{T}}\mathbf{z})$, where $\lambda_0(t)$ is the baseline hazard corresponding to $\mathbf{z}^{\mathbf{T}} = (0, \ldots, 0)$ and $\boldsymbol{\beta}$ is a vector of regression coefficients. The likelihood for the Cox regression model is simply

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{L} \frac{\exp(\boldsymbol{\beta}^{\boldsymbol{T}} \mathbf{z_i})}{\sum_{j \in D_i} \exp(\boldsymbol{\beta}^{\mathbf{T}} \mathbf{z_j})} \tag{4}$$

where D_i is the set of subjects at risk at time t_i. The efficient score for Equation (4) is given by $\mathbf{U}(\boldsymbol{\beta}) = \partial/\partial\boldsymbol{\beta} \log L(\boldsymbol{\beta})$, and its covariance by the inverse of $\mathbf{I}(\boldsymbol{\beta}) = -\ \partial^2/\partial\boldsymbol{\beta}^2 \log L(\boldsymbol{\beta})$. Then the score statistic is $\mathbf{U'(0)I^{-1}(0)U(0)}$, which has a chi-squared distribution with $p-1$ degrees of freedom. This statistic is equivalent to the log-rank test when there are no tied survival times.

0.0.6 Power. The log-rank test is most powerful when the survival curves are proportional. This occurs when one survival function is consistently greater than the other over the study period. The log-rank test is the most powerful nonparametric test to detect proportional hazards alternatives. If the hazard functions cross, then there may be very little power to detect differences between the survival curves. One easy way to assess the proportionality assumption is to plot the Kaplan-Meier survival curves. If the survival curves cross, then the proportionality assumption is not met. Alternatively, a plot the estimated survival curves on a log(-log) scale gives a constant vertical shift of the two curves by an amount equal to the log of the hazards if the hazards are proportional.

A more rigorous approach to checking the proportionality assumption is to use a statistical test based on a Cox regression model. Proportionality fails when there is an interaction between treatments and time. Introduction of a time-dependent interaction can be used to test formally for nonproportional hazards with the Cox regression model. Therneau and Grambsch (5) describe using residuals from Cox regressions to identify deviations from the proportional hazards assumption.

1 EXAMPLE: DISEASE-FREE SURVIVAL FOR ACUTE MYELOGENOUS LEUKEMIA AFTER BONE MARROW TRANSPLANTATION

Klein and Moeschberger (6) provide a dataset containing 101 patients who received bone marrow transplantation after chemotherapy for acute myelogenous leukemia. Transplants were either allogenic (from the patient's sibling) or autologous (from the patients own marrow harvested prior to chemotherapy). The event time was based on relapse or death, whichever occurred first. The R software (7) package `KMsurv` contains this dataset called `alloauto`.

Each patient in the dataset has a sequence number, a leukemia-free survival time (in months), an indicator for censoring (0 = yes, 1 = no), and an indicator for type of bone marrow transplant (1 = allogenic and 2 = autologous). There are 101 subjects in the dataset and 50 leukemia relapses. Of the 101 patients, 50 had allogenic transplants and 51 had autologous transplantation.

An R script for this example is in Table 2. `survfit` calculates the Kaplan-Meier curve. The `plot` command gives the Kaplan-Meier curves shown in Fig. 1. The allogenic transplant survival is initially higher than the autologous transplants. This trend reverses itself as the survival functions cross at about 12 months casting doubt on the proportional hazards assumption. Vertical dashes on the survival functions represent censored observations. Finally, the `survdiff` command calculates the log-rank test shown in Table 3.

Although there is separation of the two survival curves late in the time axis, the log-rank test does not yield a $P < 0.05$. Differences in the survival functions summed over time decrease the magnitude of the log-rank statistic when the survival functions cross. The sum of the quantities (O-E)^2/E in Table 3 is a conservative approximation to the actual log-rank chi-squared statistic and is produced in the output for information purposes only.

1.0.7 The Stratified Log-Rank Test. Sometimes one may know that survival is not proportional among all subjects but is related to a nuisance factor. Heterogeneous populations can sometimes be stratified into homogeneous groups for analysis purposes eliminating the nuisance source of variation. Stratified analysis can be applied to the log-rank test. This process is appropriate for situations where the proportional hazards assumption breaks down in known subgroups. For example, the hazard rate may be different for a disease outcome depending on disease burden at baseline. In that case, it may be possible to define strata within which the proportional hazards assumption is more

Table 2. R Script for Kaplan-Meier Plot and Log-Rank Test

```
library(survival)
data(alloauto, package = ``KMsurv'')
my.fit <- survfit( Surv(time,delta) ~ type, data = alloauto )
plot(my.fit, xlab = ``Time (Months)'', ylab = ``Leukemia Free
  Survival Probability'', lty = c(6, 1))
legend(5, 0.35, c(``allogenic'', ``autologous''), lty = c(1,
  3))
survdiff( Surv(time,delta) ~ type, data = alloauto )
```

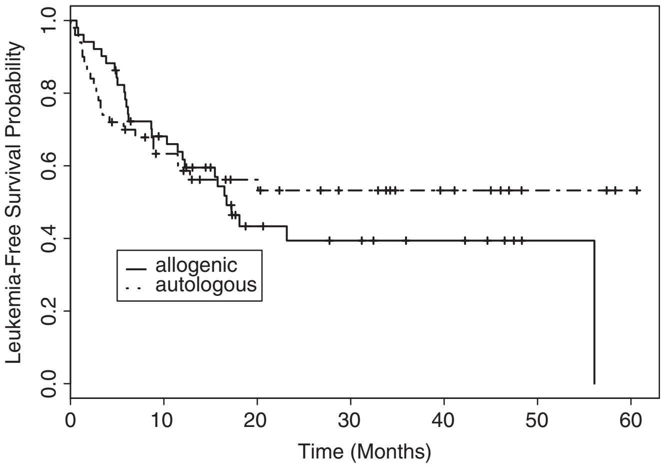

Figure 1. Kaplan-Meier survival curves for autologous and allogenic bone marrow transplants.

viable. The elements of the log-rank test can be computed separately for each stratum and then combined for an overall test. Let the strata be indexed $h = 1, \ldots, s$ and let $\mathbf{v}^{(h)}$ and $\mathbf{V}^{(h)}$ represent the stratum-specific components of the log-rank statistic. Then the stratified test statistic is expressed as (4):

$$\chi^2_{K-1} = \left(\sum_{h=1}^{s}\mathbf{v}^{(h)}\right)^T \left(\sum_{h=1}^{s}\mathbf{V}^{(h)}\right)^{-1} \left(\sum_{h=1}^{s}\mathbf{v}^{(h)}\right)$$

1.1 Choice of Weights

The log-rank test is a special case of more general methods for comparing survival distributions with different weights. Careful selection of weights can emphasize differences between certain regions of the survival curves. Several weighting schemes are summarized in Table 4. Weights equal to the number of subjects at risk R_i at each time t_i was proposed by Gehan (8) for the two group setting and by Breslow (9) for multiple groups. $W_j(t) = R_i$ weights early portions of the survival functions more heavily than later portions. Tarone and Ware (10) proposed using weights that are a function of R_i such as $\sqrt{R_i}$. Peto and Peto (11) proposed a weighting scheme based on an estimate of the common survival function. Anderson et al. (12) recommended slight modification to the Peto–Peto weights.

Table 3. Results of the Log-Rank Test for Bone Marrow Transplant Patients

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
type = 1	50	22	24.2	0.195	0.382
type = 2	51	28	25.8	0.182	0.382
Chisq = 0.4 on 1 degrees of freedom, p = 0.537					

Table 4. Common Weights for Weighted Log-Rank Statistics

Weight	Test
1.0	Log-rank (1)
R_i	Gehan-Breslow-Wilcoxon (8, 9)
$\sqrt{R_i}$	Tarone-Ware (10)
$\hat{S}(t) = \prod_{t_i<t}\left(1 - \frac{d_i}{R_i+1}\right)$	Peto-Peto-Wilcoxon (11)
$\hat{S}(t)R_i/(R_i+1)$	Modified Peto (12)
$\left\{\hat{S}(t)\right\}^{\rho}\left\{1-\hat{S}(t)\right\}^{\gamma}$	G^{ρ}, $G^{\rho,\gamma}$ class (13, 3) $\rho \geq 0$, $\gamma \geq 0$

Table 5. Log-Rank Statistics for Bone Marrow Transplant Data Calculated Using SAS PROC LIFETEST

Line No.	Test	Chi-Square	DF	P-Value
1	Log-rank	0.3816	1	0.5368
2	Wilcoxon	0.0969	1	0.7556
3	Tarone	0.0039	1	0.9501
4	Peto	0.0000	1	0.9956
5	Modified Peto	0.0007	1	0.9791
6	G (1,0)	0.0008	1	0.9771
7	G (0,1)	4.2026	1	0.0404
8	G (0,2)	5.9276	1	0.0149
9	G (1,1)	2.9600	1	0.0853

Fleming and Harrington (13) proposed a general class of tests $G^{\rho,\gamma}$ with weight function $W_{\rho,r}(t_i) = \hat{S}(t_{i-1})^{\rho}[1 - \hat{S}(t_{i-1})]^{\gamma}$, $\rho \geq 0$, $\gamma \geq 0$, where $\hat{S}(t)$ is the product-limit estimator based on the combined sample. Choice of ρ and γ can provide flexibility in selecting a region of the survival curves for weighting differences among the curves. Of course, $W_{\rho,\gamma}(t_i) = 1$ when $\rho = \gamma = 0$ and we have the ordinary log-rank test. For $\rho = 1$ and $\gamma = 0$, more weight is given to early differences between the survival functions. For $\rho = 0$ and $\gamma > 0$, then more weight is given to departures between survival functions observed later during the observation period.

Table 5 shows results for the bone marrow transplant data for various weighting schemes. Statistics were calculated using SAS procedure LIFETEST (14). The log-rank test lacks power with its equal weighting in this nonproportional hazards dataset. Tests in line numbers 2–6 are even less sensitive since they weight the early portion of the curves where there is the least difference. The G-class statistics that weight the later region of the survival curves (lines 7–9) give the smallest P-values for this dataset. The SAS code for calculating these statistics is given in Table 6. The code `test = (all)` calculates the statistics in lines 1–6 in Table 5.

1.2 Sample Size and Power

The power of the log-rank test depends on the number of events during the course of a clinical study and not on the total number of subjects enrolled. Typically one estimates the number of events required and then makes assumptions about the accrual rate, length of the intake period, and the length of follow-up to estimate the number of subjects required to observe a specific number of events for a given power requirement.

In the two-sample log-rank proportional hazards setting, assume the alternative hypothesis $H_A : S_1(t) = S(t)^{\theta}$. For the two-sided log-rank test with level $\alpha = 0.05$ and power $1 - \beta$, Schoenfeld (15) showed that the total sample size required is

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2}{(\ln \theta)^2\, p(1-p)P_d}$$

Table 6. SAS v9.1 Code to Calculate the *P*-Values for Weighted Log-Rank Statistics

```
proc lifetest notable data = alloauto;
   time time*delta(0);
   strata /group = type test = (all); run;
proc lifetest notable data = alloauto;
   time time*delta(0);
   strata /group = type test = ( fleming(0, 1) );
proc lifetest notable data = alloauto;
   time time*delta(0);
   strata /group = type test = ( fleming(1, 1) );
proc lifetest notable data = alloauto;
   time time*delta(0);
   strata /group = type test = ( fleming(0, 2) );
run;
```

where p is the proportion of subjects in group 1 and P_d represents the cumulative proportion of events in both groups by the end of the study. Generally the calculation of P_d will require assumptions about the survival distributions in each group. P_d could be estimated by assuming an exponential survival function or through simulation based on θ.

In the real world of study design, the number of events observed depends on the rate of accrual, the length of the intake period, the dropout rate, and the length of follow-up. Lakatos (16) developed a general method for calculating sample size for the log-rank test that relaxes the proportionality assumption and allows arbitrary specification of the survival functions. Lakatos and Lan (17) reviewed the literature on different methods for estimating sample sizes and compared them with computer simulations. See also "Sample Size Calculation for Comparing Time-to-Event Data" in this encyclopedia.

Use of a computer is recommended for complicated designs. The SAS procedure POWER (14) can calculate power and sample size for the log-rank test comparing two survival curves using the Lakatos method. The survival functions can be proportional hazards models, piecewise linear curves with proportional hazards, or arbitrary piecewise linear curves. The software allows specification of uniform accrual periods and a follow-up time.

The EastSurv (18) program allows for calculation of sample size and power along with interim analysis planning. EastSurv permits use of time-varying accrual patterns and modeling of dropout rates and piecewise exponential hazards.

REFERENCES

1. N. Mantel, Evaluation of survival data and two new rank order statistics arising in its construction. *Cancer Chemotherapy Rep.* 1966; **50**: 163–170.
2. D. R. Cox, Regression models and life tables. *J. Roy. Stat. Soc.* 1972; **B34**: 187–220.
3. T. R. Fleming and D. P. Harrington, *Counting Processes and Survival Analysis*. New York: Wiley, 1991.
4. J. D. Kalbfleisch and R. L. Prentice *Statistical Analysis of Failure Time Data, Second Edition*. New York: Wiley, 2002.
5. T. M. Therneau and P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. New York: Springer, 2000.
6. J. P. Klein and M. L. Moeschberger, *Survival analysis: Techniques for Censored and Truncated Data (Second Edition)*. New York: Springer, 2003.
7. R Development Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2007. Available: http://www.R-project.org.
8. E. A. Gehan, A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* 1965; **52**: 203–223.
9. N. Breslow, A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika*, 1970; **57**: 579–594.

10. R. E. Tarone and J. Ware, On distribution-free tests for equality of survival distributions. *Biometrika* 1977; **64**: 156–160.
11. R. Peto and J. Peto, Asymptotically efficient rank invariant test procedures (with discussion). *J. Roy. Stat. Soc. A*, 1972; **135**: 186–206.
12. P. K. Anderson, O. Borgan, R. D. Gill, and N. Keiding, Linear nonparametric tests for comparison of counting processes with application to censored survival data (with discussion). *Int. Stat. Rev*. 1982; **50**: 219–258.
13. D. P. Harrington and T. R. Fleming, A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**: 553–566.
14. SAS Institute Inc., *SAS/STAT® 9.1. User's Guide*. Cary, NC: SAS Institute Inc.
15. D. Schoenfeld, The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 1981; **68**: 316–319.
16. E. Lakatos, Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics* 1988; **44**: 229–241.
17. E. Lakatos and K. K. G. Lan, A comparison of sample size methods for the logrank statistic. *Stat Med.* 1992; **11**: 179–191.
18. Cytel Software, Inc. EastSurv, User Manual. Cambridge, MA: Cytel, Inc., 2005.

CROSS-REFERENCES

Censoring

Kaplan-Meier Plots

Cox proportional hazard model

Hazard rate

Hazard ratio

Sample Size Calculation for Comparing Time-to-Event Data

Stratified Analysis

Survival Analysis