# ON ROBUSTNESS OF LOG-RANK TEST AND AN ASSUMPTION-FREE ALTERNATIVE

**Lubomír Štěpánek – Filip Habarta – Ivana Malá – Luboš Marek**

**Abstract**

A situation of comparing two time-event survival curves is very common in applied statistics. Although the log-rank test is the first weapon of choice, it used to be limited by relatively rigorous statistical assumptions. Based on the context, there is a larger statistical toolbox used to exceed the efficiency of the log-rank test. However, each of its approaches has some limitations and violates statistical assumptions in different ways. In this work, we discuss selected issues of the robustness of the log-rank test. Furthermore, we also propose a bit different, assumption-free framework on how to model individual time-event survival curves that are depicted in a discrete combinatorial way as orthogonal paths in a grid of survival plot, which, besides others, enables by their counting up a direct estimation of the *p*-value using its original definition. Finally, using simulated time-event data, we check the robustness of both the log-rank test and the introduced method. Based on the simulations, the robustness of the log-rank test could be sometimes limited, while the sketched alternative seems to be promising on how to compare time-event curves regardless of any assumptions are met.

**Keywords:** survival analysis, log-rank test, robust alternative, time-event survival curve, graphical surface analysis

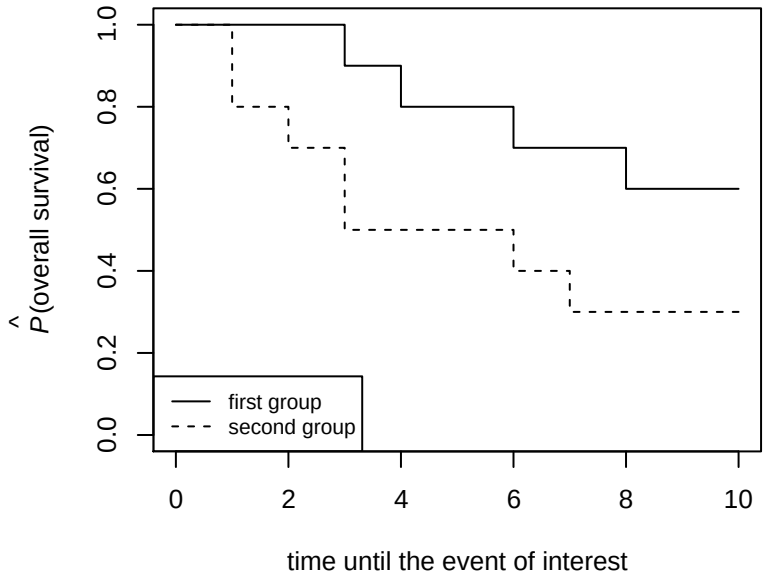
**JEL code:** C12, C14, C18

## Introduction

Regardless of its name, survival analysis includes various events of possible interests, not only – historically spoken – statuses such as a death or disease of treated individuals. Distributions of time points when the event of the interest in each individual either does occur or does not, and if not, such a subject is so-called censored, is one of the things that makes survival calculations somewhat different from other common kinds of statistical analyses. Consequently, the target variable is two-dimensional, both the time of the event and whether

the event or the censoring occurred does matter. Intuitively, such a variable suggests being plotted in a plane (two-dimensional) chart; usually, a number of subjects not experiencing the event of interest to a number of all subjects is plotted on a vertical axis at a given time point while the horizontal axis stands for the time until the event of interest (or until the censoring) (see Fig. 1). This is how Kaplan-Meier estimator is usually illustrated (Kaplan and Meier, 1958). Thus, the variable is represented as an orthogonal path (i. e., a polygonal path of horizontal and vertical segments) in the Cartesian two-dimensional chart. The variable deals both with the events and their times, it is commonly called the time-event survival curve.

**Fig. 1: Two time-event survival curves in a survival plot**




Source: authors' own research

When there are two time-event survival curves for two disjunctive groups of subjects supposed to be compared, a log-rank test could be used (Mantel, 1966). Under some special conditions, mainly when the data of the time-event survival curves are not censored, a simple Wilcoxon rank-sum test might be performed. If more than two groups are assumed to be compared, we can battle the problem using a score-rank test or even a Cox proportional hazards model (Li, 2015). All the methods mentioned above are easily available in both commercial and open-source software, including R language and environment (R Core Team, 2018), where they could be employed by using a pure R package `stats` or a package `survival` (Therneau, 2020.) Although, the relatively tough statistical assumptions limit the previously mentioned methods.

Usage of the log-rank test comparing two non-crossing time-event survival curves (similar as in Fig 1.) is limited when the censoring affects the events or is not balanced in the compared groups. Various modifications increasing the efficiency of the log-rank test or its robustness against the assumptions' violation were reported. While Kong (1997) improved the log-rank test efficiency by adjusting the hazard functions, Song et al. (2008) inspected covariate matrix decomposition and proposed formulas estimating minimal sample sizes that legitimize usage of the log-rank test. Peto and Peto (1972), Yang and Prentice (2010), and Li (2018), respectively, introduced the use of observation weights, usually higher for earlier events when there are larger numbers of observations, to improve the correctness of the log-rank test.

Other articles deal with exact computations when compare two survival curves. Whereas Thomas (1975) relied on fixed total numbers in the compared groups and Mehta et al. (1985) improved his algorithm computationally, Heinze et al. (2003) introduced a weighting scheme into the calculations.

All the listed papers work with a hazard function, which is a rate of the events of interest in a given time point conditional on survival until the time point or assume fixed total numbers of individuals in the compared groups. Unlike them, in this contribution, besides theoretical discussion handling with limitations of the log-rank test, we model the time-event survival curves using a discrete combinatorial approach and taking into account the mutual grid distances of the time-event curves as orthogonal paths in a two-dimensional plot (as shown in Fig. 1 and Fig. 2). That indicates how the *p*-value of the log-rank test could be calculated using its original definition as a conditional probability. Finally, with the employment of simulations of artificially generated survival curves, the first type errors are calculated for both the log-rank test and our proposed alternative, mutually compared and discussed within the scope of their robustness.

# 1 Logic, assumptions and limitations of the log-rank test

By introduction of principles of the widely used log-rank test, we can better understand both its assumptions and limitations.

## 1.1 Logic of the log-rank test

The log-rank test compares the expected and observed numbers of the events of interest in both groups of subjects across all time points where there is an event.

Consider two groups of individuals (marked by indices 1, and 2, respectively) and $k \in \mathbb{N}$ distinct event times. At each event time, we can construct a $2 \times 2$ contingency table and compare the event rates between the two groups, conditional on the number at risk in the groups. Let the $\{t_1, t_2, \cdots, t_k\}$ be an ordered tuple of the event times, then at the $j$-th event time $t_j$, we have the table Tab. 1, where $d_{1,j}$ and $d_{2,j}$ are the numbers of individuals who experienced the events in group 1 and 2, respectively, at the $j$-th event time, and $r_{1,j}$ and $r_{2,j}$ are the numbers of subjects at risk (who have not yet had the event or been censored) at that time in groups 1 and 2, respectively.

**Tab. 1: Numbers of the events of interest in both groups of subjects at the event time $t_j$**

| | event of interest at the event time $t_j$ | | |
|---|---|---|---|
| group | yes | no | total |
| 1 | $d_{1,j}$ | $r_{1,j} - d_{1,j}$ | $r_{1,j}$ |
| 2 | $d_{2,j}$ | $r_{2,j} - d_{2,j}$ | $r_{2,j}$ |
| total | $d_j$ | $r_j - d_j$ | $r_j$ |

Source: authors' own research

The log-rank test checks the null hypothesis $H_0$ that both groups have identical hazard functions. The values of the hazard functions are empirically estimated using the contingency tables similar to Tab 1. Under the null hypothesis $H_0$, the observed numbers of the events as random variables $D_{1,j}$ and $D_{2,j}$ follow a hypergeometric distribution with parameters $(r_j, r_{i,j}, d_j)$ for both $i \in \{1, 2\}$. Thus, the expected value of such a number is

$$\mathrm{E}(D_{i,j}) = r_{i,j}\frac{d_j}{r_j}, \tag{1}$$

and the variance is

$$\mathrm{var}(D_{i,j}) = \frac{r_{1,j}r_{2,j}d_j}{r_j^2}\left(\frac{r_j - d_j}{r_j - 1}\right), \tag{2}$$

for both $i \in \{1, 2\}$. For all $j \in \{1, 2, \cdots, k\}$ we compare the observed numbers of events $d_{i,j}$ to their expected values $\mathrm{E}(D_i) = r_{i,j}\frac{d_j}{r_j}$ under $H_0$. So, the test statistic for both $i \in \{1, 2\}$ is

$$\chi^2_{\text{log-rank}} = \frac{\left(\sum_{j=1}^{k} d_{i,j} - \mathrm{E}(D_{i,j})\right)^2}{\sum_{j=1}^{k}\mathrm{var}(D_{i,j})} = \frac{\left(\sum_{j=1}^{k} d_{i,j} - r_{i,j}\frac{d_j}{r_j}\right)^2}{\sum_{j=1}^{k}\frac{r_{1,j}r_{2,j}d_j}{r_j^2}\left(\frac{r_j - d_j}{r_j - 1}\right)} \tag{3}$$

and under $H_0$ follows a $\chi^2$ distribution with 1 degree of freedom, thus $\chi^2_{\text{log-rank}} \sim \chi^2(1)$. For feasible large $r_j$, a square root of $\chi^2_{\text{log-rank}}$ follows a standard normal distribution, $\sqrt{\chi^2_{\text{log-rank}}} \sim N(0,1)$.

### 1.2 Assumptions and limitations of the log-rank test

Besides the assumption that censoring should not affect the event of interest anyhow, the proportion of censored data should be of nearly equal size in both the groups. Otherwise, the statistic $\chi^2_{\text{log-rank}}$ calculated using (3) separately for $i = 1$, and for $i = 2$, respectively, could differ. That may affect the robustness of the log-rank test.

There is also a statistical assumption based on that the test statistic $\chi^2_{\text{log-rank}}$ follows a $\chi^2$ distribution. If the initial total number of individuals $r_0$ and the number of all event times $k$ is relatively small, than both the numerator and denominator of the fraction in the formula (3) is relatively small, too, and, consequently, we could expect that the $\chi^2_{\text{log-rank}}$ statistic (or the $\sqrt{\chi^2_{\text{log-rank}}}$. statistic) does not fulfill its supposed asymptotic properties, and its estimate could be biased. That influences both the robustness and the power of the log-rank test.

It is worth mentioning that by inspecting the denominator of the equation (3), we can realize the $\chi^2_{\text{log-rank}}$ is the highest when the denominator $\sum_{j=1}^{k} \text{var}(D_{i,j})$ is as low as possible. It could be proved this holds exactly when the proportions $\frac{r_{1,j}}{r_j} = \frac{r_{1,j}}{r_{1,j}+r_{2,j}}$ and $\frac{r_{2,j}}{r_j} = \frac{r_{2,j}}{r_{1,j}+r_{2,j}}$ are both constant (and mutually different enough) across all the time points $j \in \{1, 2, \cdots, k\}$, and then the log-rank test is the most powerful (i. e. its ability to reject the null hypothesis $H_0$ when it is not true is maximal possible). This is the main issue that impacts on the power of the log-rank test. The proportions are typically not constant when the time-event survival curves cross themselves one or more times. However, the power of the test is decreased by any deviations from the constant values of the proportion $\frac{r_{1,j}}{r_j}$, and $\frac{r_{2,j}}{r_j}$, respectively.

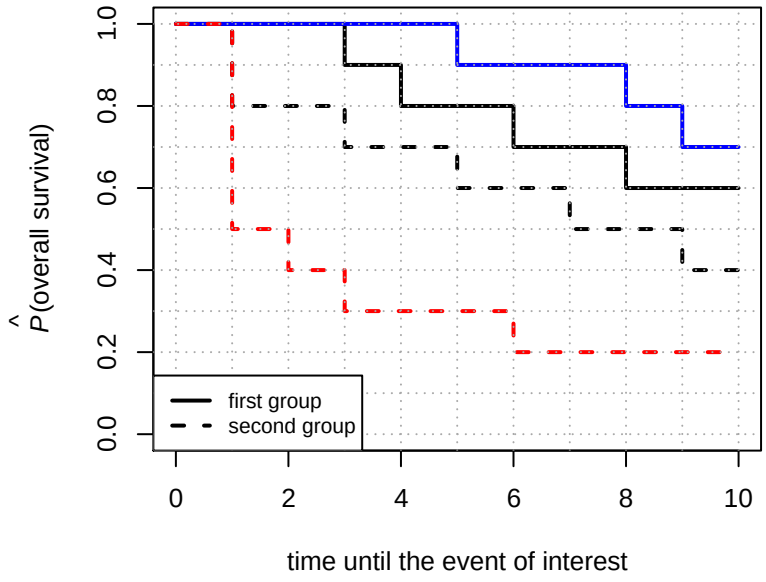
## 2 An assumption-free alternative to the log-rank test

We propose an assumption-free alternative to the log-rank test based on a discrete combinatorial calculation of possible states where one would obtain data at least as extreme as the observed data, which corresponds to the original definition of the *p*-value. All the possible states could be considered as orthogonal paths in the two-dimensional chart, including the (non-crossing)

survival curves. By calculating (or estimating) the numbers of the paths at least at extreme as the plotted two curves, we get a point estimate of the *p*-value as a proportion of paths contradicting the same way or even more to the observed survival curves.

Again, let us suppose there are two groups of individuals (marked by indices 1, and 2, respectively) and $k \in \mathbb{N}$ distinct event times. Let the $\{t_1, t_2, \cdots, t_k\}$ be an ordered tuple of the event times. At each event time $t_j$, we can compute the number of individuals experienced the event for both groups, similarly to the contingency tables as Tab 1. By doing this, consequently, once we get the proportions of subjects at risk, $\frac{r_{1,j}}{r_j}$, and $\frac{r_{2,j}}{r_j}$, respectively, for each event time $t_j$, we could plot the time-event survival curves similarly to Fig. 1.

**Fig. 2: Two time-event survival curves in a survival plot (bold lines) and examples of monotonic orthogonal paths above (blue dashed line) and below (red dashed line) the original survival curves**




Source: authors' own research

For simplicity, let us assume that the survival curves do not cross themselves. By adding a grid into the Fig. 1, we get Fig. 2, which straightforwardly suggests to calculating (or estimating) a number of monotonic paths starting at the proportion of subjects at risk $\frac{r_{1,0}}{r_0} = \frac{r_{2,0}}{r_0} = 1$ and ending – after $k$ event times – at the proportion of subjects at risk $\geq \frac{r_{1,k}}{r_k}$ (one of such paths is the blue line in Fig. 2) or $\leq \frac{r_{2,k}}{r_k}$ (one of such paths is the red line in Fig. 2).

Let $N_{k,u.v}$ stands for the number of all monotonic orthogonal paths (respecting the grid, i. e. all segments of such a path are parallel to horizontal or vertical lines of the grid and its edges are aligned to grid points) starting at the proportion 1 (left upper corner of the Fig. 2) and ending after $k$ event times at the proportion of subjects at risk $\frac{u}{v}$ (a point with coordinates $\left[k, \frac{u}{v}\right]$ in Fig. 2). Eventually, let $N^{+}_{i,k,u.v}$ (or $N^{-}_{i,k,u.v}$) be a number of all orthogonal paths starting at the proportion 1, going above (or below) the $i$–th survival curve or tangentially meeting it (without crossing it) and ending at the proportion of subjects at risk $\geq \frac{u}{v}$ (or $\leq \frac{u}{v}$) after $k$ event times. The numbers $N^{+}_{i,k,u.v}$ and $N^{-}_{i,k,u.v}$ could be computed exhaustively in a combinatorial way or estimated by numerical simulations.

Let us test a null hypothesis $H_0$ that the survival curves are not significantly different. The tricky part is that, since we do not need any assumptions for this testing, we do not require modeling a null distribution. The *p*-value is the probability of obtaining data at least as extreme as the data currently observed, assuming that the null hypothesis is correct. To be more specific, let the *p*-value mark as $p$, then

$$p = p\text{-value}$$

$$p = P(\text{obtaining data at least as extreme as the observed data}|H_0)$$

$$p = P\left(\frac{N^{+}_{1,k,r_{1,k}.r_k} \cdot N^{-}_{2,k,r_{2,k}.r_k}}{\left(\sum_{j=0}^{r_k} N_{k,j.r_k}\right)^2 - N_{\text{crossing curves}}}\right), \tag{4}$$

where $N_{\text{crossing curves}}$ is a number of pairs of such survival curves in the survival plot that cross themselves. Again, the number $N_{\text{crossing curves}}$ can be calculated either using a combinatorial approach, or be numerically simulated.

## 3 A simulation study

We compared the robustness of the log-rank test and the proposed assumption-free method by simulating pairs of random non-crossing curves that are not significantly different and calculating the first type errors, supposing that a more robust method should have less value of the first type error. Firstly, we generated pairs of survival curves such that both curves in the pair follow one generated negatively exponential survival function following the form

$$s(t) = \sigma\left(e^{-\frac{10+\varepsilon}{10000}t}\right) \tag{5}$$

where $\varepsilon$ is a random noise term and follows a standard normal distribution, i. e. $\varepsilon \sim N(0,1)$ and $\sigma(\cdot)$ is a function rounding its argument to the nearest integer. Applying the formula (5), we

generated $n = 1000$ pairs of significantly not different survival curves and compared them using both the log rank test, and the above proposed method. By counting up numbers of cases where *p*-value was lower than or equal to $0.05$, we got the point estimates of the first type error frequencies, as demonstrated in table Tab. 2.

**Tab. 2: Point estimates of the first type error rates for the log-rank test and the proposed method**

| | method | |
|---|---|---|
| | the log-rank test | the proposed method |
| # of simulated cases in total | 1000 | 1000 |
| # of cases when p-value $\leq 0.05$ | 56 | 12 |
| point estimate of the first type error rate | 0.056 | 0.012 |

Source: authors' own research

The point estimate of the first type error rate for the log-rank test is about 0.056, which is controlled by the common setting of the alpha level equaled to 0.050. On the other hand, the point estimate of the first type error rate for the method introduced above is about 0.012, therefore lower than the one for the log-rank test. The proposed method seems to be more robust than the log-rank test.

## Conclusion

We have discussed some of the issues that may affect the robustness (and statistical power) of the log-rank test. The log-rank test is considered to be nonparametric, but still is limited by a quality of the used data (following the principle "garbage in, garbage out"); whenever the censoring might influence the event of interest, or the censoring rates are not balanced across the groups of individuals, the robustness of the statistical power of the log-rank test may be affected. Similar to other common tests of statistical inference, the output of the log-rank test may be biased by both the first or second type errors.

The introduced method uses an original definition of the *p*-value as a conditional probability of getting data at least as extreme as the observed data, assuming the null hypothesis is correct. By combinatorial or exhaustive calculation of orthogonal paths in the grid of survival plot, we can get a ratio of the number of all pairs of the paths corresponding to the survival curves opposing the null hypothesis and the number of all non-crossing pairs of possible paths.

Based on the simulation of the first type error rates, the proposed method proved to be of higher robustness than the log-rank test, which is in accordance with our prior expectations. The assumption-free version of the log-rank test seems to be a valid alternative for the comparison of two time-event curves. Besides, the method and theory behind it could also be a topic for a new R package development.

## Acknowledgment

This research was supported by the grant no. 45/2020 provided by Internal grant agency of University of Economics, Prague.

## References

Heinze, G., Gnant, M., & Schemper, M. (2003). Exact Log-Rank Tests for Unequal Follow-Up. *Biometrics, 59*(4), 1151–1157. https://doi.org/10.1111/j.0006-341x.2003.00132.x

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association, 53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Kong, F. (1997). Robust covariate-adjusted logrank tests. *Biometrika, 84*(4), 847–862. https://doi.org/10.1093/biomet/84.4.847

Li, C. (2018). Doubly robust weighted log-rank tests and Renyi-type tests under non-random treatment assignment and dependent censoring. *Statistical Methods in Medical Research, 28*(9), 2649–2664. https://doi.org/10.1177/0962280218785926

Li, H., Han, D., Hou, Y., Chen, H., & Chen, Z. (2015). Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods. *PLOS ONE, 10*(1), e0116774. https://doi.org/10.1371/journal.pone.0116774

Mantel N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports, 50*(3), 163–170.

Mehta, C. R., Patel, N. R., & Gray, R. (1985). Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2 × 2 Contingency Tables. *Journal of the American Statistical Association, 80*(392), 969. https://doi.org/10.2307/2288562

Peto, R., & Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A (General), 135*(2), 185. https://doi.org/10.2307/2344317

R Core Team. R. (2018) *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Song, R., Kosorok, M. R., & Cai, J. (2007). Robust Covariate-Adjusted Log-Rank Statistics and Corresponding Sample Size Formula for Recurrent Events Data. *Biometrics*, *64*(3), 741–750. https://doi.org/10.1111/j.1541-0420.2007.00948.x

Therneau T (2020). *A Package for Survival Analysis in R*. R package version 3.1-12, https://CRAN.R-project.org/package=survival.

Thomas, D. G. (1975). Exact and asymptotic methods for the combination of 2 × 2 tables. Computers and Biomedical Research, 8(5), 423–446. https://doi.org/10.1016/0010-4809(75)90048-8

Yang, S., & Prentice, R. (2009). Improved Logrank-Type Tests for Survival Data Using Adaptive Weights. *Biometrics*, *66*(1), 30–38. https://doi.org/10.1111/j.1541-0420.2009.01243.x

## Contact

Lubomír Štěpánek
Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
lubomir.stepanek@vse.cz

Filip Habarta, Ivana Malá, Luboš Marek
Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
{filip.habarta, malai, marek}@vse.cz