Comparing population distributions with censored data in area of complaints

Angelina Rajda-Tasior[1]

Abstract

Issues relating to the analysis of survival can be used to test the products life or time associated with the complaints. In many cases they concern research based on data censored because of the unavailability of some individual data at certain time. Situations such as a lack of information about some periods of time, or suspected that they come from different population enforce special treatment of all data obtained from the sample, what can cause that modification of the classical methods will be necessary.

The paper presents the non-parametric test to compare two distributions with censored data. This situation occurs when the researcher does not have full information about elements in the sample. Attention was paid to the economic benefits arising from the use of that solution. The characteristics of the proposed method were tested by using a computer simulation.

Keywords: *censored data, complaints analysis, permutation test*

JEL Classification: C41

1. Introduction

Comparing the two populations is very important from the practical point of view. It allows to compare two different objects for example: products, services, industrial and also economic processes. The paper presents non-parametric analyzes for comparing two groups of objects. First part of study refers to permutation test in the verification of the hypothesis of the equality for distributions in the groups. Permutation tests were applied irrespective of distribution of tested variables even when the sample size is small. For data with censored observations was used two non-parametric analysis such us Kaplan–Meier and Nelson–Aalen. For testing difference in survival functions among groups applied test with ρ-parameter based on weight log rank. All of tests refer to quality data in area of complaints (quantity or time of realization).

2. Survival analysis

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. Survival analysis

[1] University of Economics in Katowice, Department of Statistics, 1 Maja 50, 40-287 Katowice, Poland, e-mail: angelina.rajda@gmail.com.

were primarily developed in the medical and biological sciences, but they are also widely used in the social, finance and economic sciences, as well as in engineering. Survival analysis is also called as duration analysis, transition analysis, failure time analysis and time-to-event analysis. For example: time from born to death object, time from get illness to death object, time from operation to death object, time from start working company to exit from the market, time from being unemployed to find the job etc. It is also possible to use that analysis in area of quality control. For example: time from produce to failure object, time from buying to failure object, life time of complaints etc.

Time failure distribution can be described by cumulative distribution function F(t), density function $f(t)$, survival function $S(t)$, hazard function $\lambda(t)$ and cumulative hazard function $A(t)$ (Rossa, 2003, 2005; Stanisz, 2007).

Let T be a random variable with probability density function $f(t)$ and cumulative distribution function:

$$F(t) = P(T \leq t), \tag{1}$$

giving the probability that the event has occurred by duration t. It will often be convenient to work with the complement of the cumulative distribution function, the survival function:

$$S(t) = P(T > t) = 1 - F(t), \tag{2}$$

which gives the probability that the event of interest has not occurred by duration t. Hazard function is defined as follows:

$$\lambda(t) = \frac{F(t+h) - F(t)}{hS(t)} = \frac{f(t)}{S(t)}. \tag{3}$$

For $t > 0$ value of survival function S(t) determines probability survival in a period [0, t], for $t > 0$ value of hazard function λ(t) determines risk failure after time t, only when object will survival until specified time t. Cumulative hazard function is defined as follows:

$$\Lambda(t) = \int_0^t \lambda(u)du. \tag{4}$$

3. Censored observations

Observations are called censored when the information about their survival time is incomplete. The most commonly encountered form is right censoring. Censoring is an important issue in survival analysis, representing a particular type of missing data. Censored observations may occur in a number of different areas of research. In economics study can research on the “survival” times of new businesses or the “survival” times of products such as

automobiles. In quality control research, it is common practice to study the "survival" of parts under stress (failure time analysis). Statistical methods for dealing with censored data have a long history in the field of survival analysis and life testing (Kalbfleisch and Prentice, 1980; Miller, 1981; Nelson, 1982; Lee and Wang, 2003; Hosmer et al., 2008; Therneau and Grambsch, 2010). The type of censored data are not discussed in that paper but (Cohen, 1991) who stands several types of censored data as follows: observed, truncated, censored; left, right, interval censored; Type I, Type II, randomly censored; single and multiple censoring values.

4. Non-parametric tests to compare two groups

Parametric methods assume that the underlying distribution of the survival times follows certain known probability distributions. Popular ones include the exponential, Weibull, and lognormal distributions. The description of the distribution of the survival times and the change in their distribution as a function of predictors is of interest. Model parameters in these settings are usually estimated using an appropriate modification of maximum likelihood.

A nonparametric estimator of the survival function, the Kaplan Meier method is widely used to estimate and graph survival probabilities as a function of time. It can be used to obtain univariate descriptive statistics for survival data, including the median survival time, and compare the survival experience for two or more groups of subjects. To test for overall differences between estimated survival curves of two or more groups of subjects several tests are available, including the log-rank test. This can be motivated as a type of chi-square test, a widely used test in practice, and in reality is a method for comparing the Kaplan–Meier curves estimated for each group of subjects.

Most nonparametric tests for verifications hypothesis for equality of survival functions are modification of test using ranks for example sum ranks Wilxona test or Kruskala–Wallis test (Baszczyńska and Domański, 1998).

Let n_i, for $i = 1, 2, \ldots, k$ will be a number of objects from i-th population.

Let $(X_{i1}, I_{i1}), \ldots, (X_{in_i}, I_{in_i})$ for $i = 1, 2, \ldots, k$ will be independent sample with right censored data randomly selected from i-th population, where:

$$X_{ij} = \min(T_{ij}, C_{ij}), \tag{5}$$

$$I_{ij} = I_{\{T_{ij} \le C_{ij}\}} = \begin{cases} 1 & for \quad T_{ij} \le C_{ij} \\ 0 & for \quad T_{ij} > C_{ij} \end{cases}, \tag{6}$$

and T_{ij} means failure time *j*-th object in *i*-th group, C_{ij} is a right censoring time of object. Hypothesis where $S_i(k)$ means survival function for *i*-th population can be enrolled as following:

$$H_0 : S_1(t) = S_2(t) = \ldots = S_k(t) \text{ for each } t \in R,$$

and also in another form:

$$H_0 : \Lambda_1(t) = \Lambda_2(t) = \ldots = \Lambda_k(t) = \Lambda$$

where $\Lambda_i(t) = -\ln S_i(t)$ is cumulative hazard function of all objects in *i*-th group.

The test statistic of weight log-rank test is based on Nelson–Aalen estimator given in formula (7). It is a non-parametric estimator of cumulative hazard function (Domański et al., 2014):

$$\hat{\Lambda}(t) = \int_0^t \frac{dN(u)}{Y(u)} = \sum_{j:I_j=1, X_j \le t} \frac{1}{n_j} \tag{7}$$

where

$$Y(t) = \sum_{i=1}^{n} Y_i(t),\ N(t) = \sum_{i=1}^{n} N_i(t),\ n_j = Y(X_j).$$

The statistic of weight log-rank test is given by following formula:

$$V_i = \int_0^{\infty} W_i(u) d\left[\hat{\Lambda}_i(u) - \hat{\Lambda}^i(u)\right] \tag{8}$$

where

$$\hat{\Lambda}_i(t) = \int_0^t \frac{dN_i(u)}{Y_i(u)},\ \hat{\Lambda}^i(t) = \int_0^t I_i(u) \frac{dN(u)}{Y(u)},\ W_i = KY_i.$$

5. Permutation test

Permutation tests were introduced by R.A. Fisher and E.J.G. Pitman in 1930's (Berry at al., 2014). In permutation tests the observed value of the test statistic is compared with the empirical distribution of this statistics under the null hypothesis. Lehmann shows that permutation tests are generally asymptotically as good as the best parametric ones (Lehmann, 2009). The main application of these tests is a two-sample problem (Efron and Tibshirani, 1993). There are following steps in dealing with permutation tests (Good, 2005):

1. Identify the null hypothesis and the alternative hypothesis.
2. Choose a test statistic *T*.
3. Compute the value T_0 of the test statistic for the sample data.

4. Determine by the series of permutations the frequency distribution of the test statistic under the null hypothesis (T_1, T_2, …, T_N, where $N \geq 1000$).
5. Make a decision using this empirical distribution as a guide.

The decision concerning a verified hypothesis is made on the basis of ASL (*Achieved Significance Level*) value (Efron and Tibshirani, 1994) has the following form:

$$ASL = P\left(T \geq T_0\right), \tag{9}$$

for which estimation is obtained by the following formula:

$$ASL \approx \frac{card\{i : T_i \geq T_0)}{N}. \tag{10}$$

This notation applies, where the H_0 rejection area is right-sided. In the case of the left-sided rejection area in above notation inequality should be changed. If the value of *ASL* is < α, then H_0 will be rejected, otherwise H_0 hypothesis cannot be rejected.

Permutation tests could be used in practice because of flexibility of the test statistic and minimal assumptions (Butar and Park, 2008). Permutation tests can be applied to all kind of data and to values of normal or non-normal density. A typical problem which can be considered with permutation test is comparing two populations (Kończak, 2012). Let S_1 and S_2 are two samples of the sizes n_1 and n_2. These hypotheses have following form:

$$H_0 : F(x) = G(x),$$

$$H_1 : F(x) \neq G(x).$$

A typical test statistic for comparing means in two populations has the following form:

$$T = \bar{X}_1 - \bar{X}_2. \tag{11}$$

The high values or the small (negative) values are against the hypothesis H_0. The *T* statistic deals with univariate variable.

6. Research

Two empirical data sets with censored observations were analyzed. Both of them refer to area of quality control. Specifically, it relates to a life time of objects during the time *t* and also to the life time of existing complaints. The data sets contain censored observations means that there are several missing values in the samples. One comparing distribution is based on permutation test and the other one on the Kaplan–Meier and the Nealson–Aalen tests, which are non-parametric methods. All calculations were conducted in R program. This program is

well suited for statistical analysis and for permutation testing as well. Significance level $\alpha = 0.05$ in all performed tests was assumed.

6.1. Failure time

Idea of application the proposed method presents following example. Data concerns to the life time of the objects. The life time of item was calculated as number of days from the date when object was sold to the date when object stopped working and complaint was made. Some of the dates are missing. Which means that survival time of object is unknown. From that reason imputation technique based on mean replacement was applied. Than for comparing two samples permutation test were performed. Objects are classified in two groups but they are belonging to the same class of furniture items.

Let assume that the observations of the monitored process could be written as a sets of sequence: $\boldsymbol{X}=\{x_1, x_2 \ldots, x_n\}$ and $\boldsymbol{Y}=\{y_1, y_2^* \ldots, y_n\}$ which represents the times of survival products for two groups. Variable y_i^* represents censored observation.

Suppose $X_1, X_2 \ldots, X_n$ are independent, identically distributed random variables with $F(s)$ right sided cumulative distribution function. Similarly, let $Y_1, Y_2 \ldots, Y_n$ be independent, identically distributed random variables with common right cumulative distribution function $G(s)$. The null hypothesis H_0: $F(X) = G(Y)$ that the X and Y random variables have the same distributions will be tested using test based on comparing means (Efron, 1967).

Null hypothesis can be tested using by permutation test. The value of test statistics T_0 on the basis of (13) for the sample data has been calculated, then N = 1,000 permutations of variables X and Y were performed and values of statistics T_i (i =1, 2, …, N) were determined. The decision concerning a verified hypothesis is made on the basis of ASL formula (12). *Achieved Significance Level* value was computed. ASL > α than H_0 cannot be rejected. What means that there is not enough evidence available to suggest that the null hypothesis is false at the 95% significance level. What means there is not enough evidence available to suggest that the distributions are identical.

6.2. Survival time

This example refers to survival analysis, specifically, it relates to how long complaints were realized. As was mentioned survival analysis could be use in area of quality control. The analysis concerns the time from product was complained to the time when product was repaired or replaced by a new. Table 1 shows part of data which were use in that example. Let t be a number of periods that object was considered by producer. It means that first object was

examines for 14 days. Next the variable the *event*, let by symbol *C* which is type of censored data. When that variable *C* is 1, it means that complain was examined, what is positive situation. In classical survival analysis it could be named as failure. When value is 0 it means that there is no information about object. That object is exactly called as censored. Another categorical variable is *g* which grouping object in two categories 1 or 2, but they are in the same class of furniture.

c	StartDateComplaint	StopDateComplaint	Time [t]	Event [failure]	Object	Group [g]
4	2013-04-24	2013-05-08	14	1	IDA	2
2	2013-02-05	2013-02-20	15	1	IDA	2
7	2013-07-23	2013-08-22	30	1	IDA	2
6	2013-06-05	2013-06-11	6	1	IDA	2
5	2013-05-20	2013-05-28	8	1	ADA	1
10	2013-10-21		0	0	IDA	2
1	2013-01-08	2013-02-05	28	1	IDA	2
1	2013-01-08		0	0	IDA	2
1	2013-02-01	2013-02-08	7	1	IDA	2
2	2013-02-06	2013-04-17	70	1	IDA	2
9	2013-09-25	2013-10-11	16	1	ADA	1
7	2013-07-24	2013-09-20	58	1	IDA	2
8	2013-08-23	2013-09-13	21	1	IDA	2
6	2013-06-10	2013-07-25	45	1	IDA	2
7	2013-06-27	2013-07-18	21	1	IDA	2

Table 1. The view of data set with censored observations.

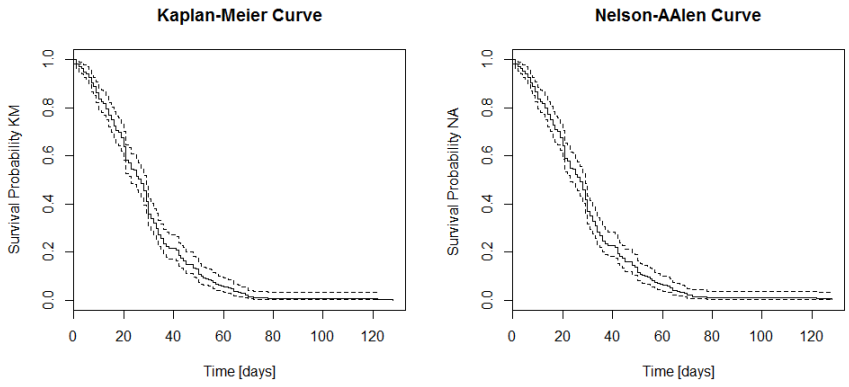

Fig. 1. Kaplan–Meier and Nealson–Aalen survival function for complaints life time.

Survival functions given as a result by Kaplan–Meier and Nealson–Aalen analysis together for both groups of objects are shown on picture 1. Survival function are going down over the time for both analysis. They are both non-parametric analysis but the second one is based on cumulative hazard function, but it is another way to specify survival function for that case.

For comparing two survival functions were applied test from a family of tests parameterized by parameter ρ. Test is based on weights on each events (called in survival analysis as a failure) of $S(t)^{\rho}$, where S is the Kaplan-Meier estimate of survival. When ρ = 0 this is the log-rank test (Mantel–Haenszel test), and when ρ = 1 it is equivalent to the Peto & Peto modification of the Gehan–Wilcoxon test (Harrington and Fleming, 1982). Results after that analysis consist in the Table 2.

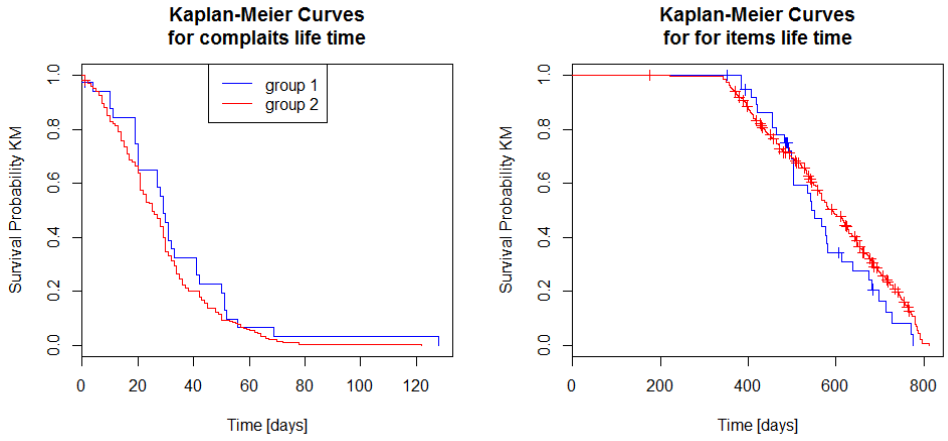

Fig. 2. Kaplan-Meier survival functions of life time items and complaints.

ρ	p-value	X^2
0	0.195	1.7
1	0.24	1.4

Table 2. Results of Mantel–Haenszel and Peto & Peto test for life complaints.

In both tests p-value is bigger than significant level (α = 0.05) therefore hypothesis H_0 cannot be rejected. What means that there is not enough evidence available to suggest that the null hypothesis is false at the 95% significance level.

Conclusion

From the practical point of view is very important to compare the two populations. It allows to compare two any objects. Specially, comparing of survival time or failure time can be used in area of quality control. The paper presents two non-parametric methods of dealing with data sets with censored values. As an example of verification of the hypothesis about equality of population distributions was applied a permutation tests based on expected value. Because of many limitation, for example no tables of critical values of test statistics used permutation test as an alternative on parametric test. In research for testing difference in survival functions among groups used two non-parametric estimators such as the Kaplan–Meier and Nelson–Aalen, and also test with ρ-parameter based on weight log rank was applied. All of that analyzes is not rejecting the null hypothesis. It means that there is not enough evidence available to suggest that the null hypothesis is false at the 0.95 significance level in both analyzed cases.

References

A Chronicle of Permutation Statistical Methods. (2014). Cham: Springer International Publishing AG.

Baszczyńska A., Domański, C. (1998). Nonparametric Inference in the Case of Random Censoring. In *Proccedings of 24th Macromodels' 97. Transition to Market System. Modeling and Forcasting of Economic and Social Consequences*. Łódź, 127-140.

Butar, F. B., & Park, J. W. (2008). Permutation tests for comparing two populations. *Journal of Mathematical Science & Mathematics Education V3*, (2), 19-30.

Cohen, A. (1991). Truncated and Censored Samples. *Statistics: A Series of Textbooks and Monographs.*

Domański, C., Pekasiewicz D., Baszczyńska A., Witaszczyk A. (2014). *Testy statystyczne w procesie podejmowania decyzji*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Efron, B. (1967). The two-sample problem with censored data. In *Proccedings of 5th Berkeley Symposium*. California, 831-853.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap*. New York: Chapman & Hall.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. New York: Springer.

Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika, 69*(3), 553-566.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. New York: Wiley.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: John Wiley & Sons.

Kończak, G. (2012). *Wprowadzenie do symulacji komputerowych*. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego.

Lee, E. T., & Wang, J. W. (2003). Statistical Methods for Survival Data Analysis. *Wiley Series in Probability and Statistics.*

Lehmann, E. L. (2009). Parametric versus nonparametrics: Two alternative methodologies. *Journal of Nonparametric Statistics, 21*(4), 397-405.

Miller, R. G., Gong, G., & Muñoz, A. (1981). *Survival analysis*. New York: Wiley.

Nelson, W. (1982). *Applied life data analysis*. New York: Wiley.

Oakes, D., & Feng, C. (2010). Combining stratified and unstratified log-rank tests in paired survival data. *Statist. Med. Statistics in Medicine, 29*(16), 1735-1745.

Oja, H., & Randles, R. H. (2004). Multivariate Nonparametric Tests. *Statistical Science Statist. Sci., 19*(4), 598-605.

Othus, M., & Li, Y. (2010). A Gaussian Copula Model for Multivariate Survival Data. *Statistics in Biosciences Stat Biosci, 2*(2), 154-179.

Rajda-Tasior A. (2014). Inference about product reliability by the analysis of complaints as a strategy for manufacturing process optimization. In *Proceedings of 32rd International Conference Mathematical Methods in Economics*. Olomouc: Palacký University, 843-848.

Rossa, A. (2003). *Niestandardowe metody estymacji rozkładów czasu trwania zjawisk w aspekcie ich zastosowań w ekonomii i ubezpieczeniach*. Łódź: Wydaw. Uniwersytetu Łódzkiego.

Rossa, A. (2005). *Metody estymacji rozkładów czasu trwania zjawisk dla danych cenzurowanych oraz ich zastosowania*. Łódź: Wydaw. Uniwersytetu Łódzkiego.

Stanisz A. (2007). *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycy. Tom 3. Analizy wielowymiarowe*. Kraków: Statsoft Polska.

Therneau, T., & Grambsch, P. (2010). *Modeling survival data: Extending the cox model extending the Cox model*. New York, NY: Springer-Verlag New York.

Xue, L., Wang, L., & Qu, A. (2009). Incorporating Correlation for Multivariate Failure Time Data When Cluster Size Is Large. *Biometrics, 66*(2), 393-404.