Use of the Kaplan-Meier Estimator in Actuarial Science

Patrícia Teplanová[1] and Michal Závodný[2]

[1] University of Economics in Bratislava, Faculty of Economic Informatics,
Dolnozemská cesta 1, Bratislava 5, 852 35
Slovak Republic
patricia.teplanova@euba.sk
[2] University of Economics in Bratislava, Faculty of Economic Informatics,
Dolnozemská cesta 1, Bratislava 5, 852 35
Slovak Republic
michal.zavodny@euba.sk

https://doi.org/10.53465/EDAMBA.2022.9788022550420.481-492

Abstract. There are many reasons why an insured person lapses (cancels) his policy. Lapse risk is one of the main risks, which is also defined in Directive Solvency II. Lapse analysis can be performed by various statistical methods. In this paper, we illustrate the possibility of using survival analysis to calculate the lapse ratio. Survival analysis does not have to be used only in medical research but nowadays finds application also in economics e.g., actuarial science. We focus on the Kaplan-Meier estimator, the most used method of modeling survival times. In practice, not every insurance policy has to lapse, so survival times from these policies should be right-censored. The Kaplan-Meier method allows to include these censored observations in the model. We use R programming language to calculate Kaplan-Meier estimation for survival times and to plot survival functions. Since the Kaplan-Meier model is univariate model, we focus on the impact of sex on insurance policy lapses.

Keywords: Survival analysis, Kaplan-Meier estimator, R programming language.

JEL classification: *C14, G22*

1 Lapse analysis in Insurance industry

Survival analysis is one of the oldest statistical methods. It was originally created in the medical field, where the time from the beginning of the patient's treatment to his death was monitored. Later, this method found application in other areas, such as economics, demography, or insurance. Survival analysis can be characterized as a set of statistical methods and procedures for examining data, where the primary random dependent variable is the time of occurrence of a previously known event. Therefore, in survival analysis we examine the length of time that elapses from the beginning of the event to its occurrence. This time can be defined as days, weeks, months, or years.

As we mentioned, one of the most frequently observed events that we can analyze is the death of an individual, i.e. time from birth to death. Survival analysis is the universal method of examining data, with which we can analyze various types of upcoming positive or negative events (illness, bankruptcy, liquidation of insurance claims, cancellation of insurance, and others).

The beginnings of survival analysis date back to the 17th century. This method was first used by Jan de Witt in 1671 in insurance to calculate the value of life annuities. In the next two centuries, scientists tried to explain the course of life respectively population mortality. At the beginning of the 20th century, the actuarial method of survival analysis was created. Paul Eugene Böhmer contributed to its creation. His work represents the revolution in the concept of survival analysis. However, his estimate of mortality rates remained forgotten for years, and in half a century, it was revised by Kaplan and Meier. Kaplan and Meier's contribution is considered one of the most important in the whole modern period of survival analysis. [4]

The insured person may terminate the life insurance policy for many reasons, such as high premiums, low financial returns, distrust of the insurance company, or he does not need to cover health risks anymore. If the client decides that he does not want to continue with the contract, this contract is canceled. The client will stop paying the premium, and if it has been agreed in advance in the contract conditions, the insurance company will pay the redemption value. A high cancellation rate, especially at the beginning of insurance, can affect the profitability of an insurance company. Therefore, a penalty charge in the life insurance sector in the first years of insurance has been introduced, which is gradually reduced during the insurance period (in the first years of insurance, for some products, this charge may be 100% of the insurance value). However, in some cases, it is advantageous to lapse the insurance policy.

Lapses are part of the insurer's risks that are not fully controllable. Therefore, the insurer should analyze and handle this risk. In practice, we encounter the calculation of the percentage of cancellation (lapse ratio), especially in the forecast of cash flows and profitability of products. In general, an insurer should know and quantify all its business risks. It is good to know after which period contracts have the greatest tendency to lapse and to predict the number of contracts in the portfolio for future periods. The loss caused by the lapse ratio is difficult to quantify. However, one option is to compare the cash flow with a 0% lapse ratio and with the calculated lapse rate.

Monitoring the lapse rate in the insurance company is essential in the calculation of the Solvency Capital Requirement according to the legislation – Directive Solvency II. The Solvency Capital Requirement (SCR) represents the required total value of the own funds of a European insurance or reinsurance company. The SCR must take into account all quantified risks to which these companies are exposed. It is the minimum amount of capital needed to cover potential losses that may occur during one year with a probability of 99,5%.

The SCR is calculated according to Solvency II as the sum of the basic capital requirement, the capital requirement for operational risk, and the adjustment for the ability to absorb losses of technical provisions and deferred tax liabilities. The basic

SCR is usually calculated once a year and covers at least the following groups of risks: non-life underwriting risk, life underwriting risk, health underwriting risk, market risk, credit risk, and operational risk. [3]

Article 105 of the Solvency II Directive defines the lapse risk as "the risk of loss, or of adverse change in the value of insurance liabilities, resulting from changes in the level or volatility of the rates of policy lapses, terminations, renewals, and surrenders". The lapse risk belongs to the sub-module of the life underwriting risk module, which is determined on the basis of a standard formula as [2]:

$$SCR_{life} = \sqrt{\sum_{i,j} Corr_{i,j} \times SCR_i \times SCR_j} \tag{1}$$

The individual combinations i and j represent combinations of the following submodules [2]:

- mortality risk,
- longevity risk,
- morbidity risk,
- life-expense risk,
- revision risk,
- lapse risk,
- and life-catastrophe risk.

One of the most common problems in data processing in survival analysis is censoring. Not for all subjects who enter the observation the event needs to occur during the research period. However, it would be wrong to exclude them from the analysis, as we would get skewed results. Therefore, the concept of censoring has been introduced. Thus, in these subjects, we observe not a survival time but a censored survival time. The censoring time determines time from the beginning of the observations to the last known mention of the subject. We also observe the censoring period for subjects who were excluded from the research for some reason.

It is important to distinguish truncation from censoring. When truncating data, we analyze only those subjects in which the monitored event occurred in the given interval $(t_L; t_R)$. We distinguish left-truncated data, in which we determine the time t_L and the time $t_R = \infty$. Right truncated data, where $t_L = 0$ and determine the time t_R. Or left and right truncated data (we determine both time t_L and time t_R). The person who overcame the event outside the time interval is removed from the research.

We define variable δ_i as a censoring indicator, which takes value 0 if the event occurred during observation or 1 if we censor the survival time.

We discern three types of censoring: right, left and interval censoring.

Let C_i as a random non-negative variable representing the censored time for the corresponding i-th observation and T_i a random variable representing its survival time. Then we say that the time T_i is right-censored if $C_i < T_i$. On the other hand, if $C_i > T_i$, then the time T_i expresses the time of occurrence of the event in the i-th subject. Then the survival time of the i-th subject is defined by the variable $U_i = min(T_i, C_i)$ and the censoring indicator δ_i ($\delta_i = 1 \leftrightarrow U_i = T_i \vee \delta_i = 0 \leftrightarrow U_i = C_i$).

The reasons for using right censoring are:

- no event occurred during the observation before the end of the observed period,

- the observed subject voluntarily withdraws from the research or is excluded from it,
- during the observation other event than the one monitored occurs and the subject can no longer be monitored (e.g. death if we examine the effectiveness of the treatment) – this reason for censoring is also related to competing risks.

Let C_i as a random non-negative variable representing the censored time for the corresponding i-th observation and T_i a random variable representing its survival time. Then we say that the time T_i is left-censored if $C_i > T_i$. On the other hand, if $C_i < T_i$, then the time T_i expresses the time of occurrence of the event in the i-th subject. Then the survival time of the i-th subject is defined by the variable $U_i = max(T_i, C_i)$ and the censoring indicator δ_i ($\delta_i = 1 \leftrightarrow U_i = T_i \vee \delta_i = 0 \leftrightarrow U_i = C_i$).

Left censoring is used, for example in research where the recruitment of subjects takes a longer time, during which subjects are not monitored, and monitoring does not begin after the recruitment period has elapsed.

Let C_i as a random non-negative variable representing the censored time for the corresponding i-th observation in which the event did not occur and D_i is a discrete random variable representing the time when the investigated event first occurred. Then, if we denote T_i as the interval-censored survival time, T_i is in the interval $C_i < T_i \leq D_i$. An example of interval censoring is the analysis of virus infectivity within a population. A person who was negative at time C, tested positive at time D. Thus, the actual time of virus infection is in the range of the interval $(C, D]$.

It is important that the censored times C_i are independent of the survival times T_i.

2 Nonparametric methods for estimating the survival function

Survival analysis is used, among other things, especially in the field of medicine and epidemiology, where our point of interest is human life or health. This fact is difficult to describe by any given probability distribution. For this reason, there is a need to invent methods for calculating the survival probability that does not require any assumptions about the distribution of the random variable survival time T – nonparametric models.

In this chapter, we will define two methods for calculating nonparametric estimates of the survival function. First, we define the empirical survival function and then the best-known nonparametric method – the Kaplan-Meier estimate. In addition to the mentioned methods, there are other models for calculating the survival probability, such as the Nelson-Aalen estimate, the Breslow estimate or the Ephron estimate [7].

2.1 Empirical survival function

We define the basic survival function as the probability that a person of age x will live to age $x + t$ (survive another t years) [8]:

$$S_x(t) = P(T_x \geq t) = 1 - F_x(t) = 1 - P(T_x < t) \tag{2}$$

We assume a set of n observations, with no observation of survival time censored. Then the survival function (equation (2)) can be estimated using the empirical survival function $\hat{S}(t)$. Equivalently, we can estimate the empirical function of the survival distribution $\hat{F}(t)$.

Based on the equation (2), we express the empirical survival function as a complement to the empiric survival distribution function. Thus, as the probability that the observed subject will live to time t (its survival time will be greater than or equal to t) [1]:

$$\hat{S}(t) = 1 - \hat{F}(t) = \frac{number\ of\ observations\ with\ survival\ time\ T \geq t}{n} \tag{3}$$

Estimation of the survival function by the empirical function is the simplest estimate, but it cannot be used if some data from the analysis are censored. For this reason, we do not encounter this estimate in practice. It serves only to simplify the calculations or to illustrate the basic knowledge of the issue of survival probability.

In the empirical survival function, we assume that its value is constant between two occurrences of the investigated events. Based on these facts, we say that the survival function is a step-by-step non-increasing survival time function T.

2.2 Kaplan-Meier estimate

The Kaplan-Meier estimate is the most widely used nonparametric estimate of the survival function. The authors presented this estimate in 1958 in their article “Nonparametric Estimation from Incomplete Observations”. As the name implies, this method can be applied to a data set, which also contains censored observations. The Kaplan-Meier estimate is a limited case of the mortality table method. [5]

We will create time intervals for the calculation while each interval will include only one time of occurrence of the event – death and death will always occur at the beginning of the interval. Furthermore, several persons may be subject to the event under investigation at the same time, and thus created intervals may not include only one death.

Suppose the survival times at which death occurred, i.e. $t_1, t_2, \ldots, t_k$. Subsequently, we arrange these times from the shortest to the longest, so $t_1 < t_2 < \ldots < t_k$. Each of these times represents the beginning of a time interval. However, our dataset may also include censored survival times $t_{c1}, t_{c2}, \ldots, t_{cm}$. We have two options, either we will not consider censored times as the beginning of a new interval and will be part of the interval between two deaths, or we will create additional intervals at the beginning of which the survival time will be censored.

Let t_0 as the beginning of the research and t_1 as the time of the first death, then the first interval will be in the range $[t_0; t_1)$. The next interval will be from the first time of the death to the second, i.e. $[t_1; t_2)$, etc. Suppose that just before a certain time t_j are alive n_j persons, where $j = 1,2,\ldots k$. We further define d_j as the number of deaths at the time t_j. Then the probability of death in the short time interval $\left[t_{j-\delta}; t_j\right]$, where δ

represents a short time unit, can be estimated as d_j/n_j. And the corresponding probability of survival at time t_j:

$$\widehat{p_j} = 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j} \tag{4}$$

where

$$n_j = n_{j-1} - d_{j-1} - c_{j-1} \tag{5}$$

c_j – number of censored observations.

The Kaplan-Meier estimate of the survival function is based on the product of the conditioned probabilities that a person will survive time t_j ($t_j \leq t$) provided that he has lived to that time [6]:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j} \tag{6}$$

In the next chapter of the article, we apply theoretical knowledge to real data concerning the lapse of insurance contracts in an unnamed universal insurance company. In practice, we distinguish between two types of lapses, lapse without payment of redemption value and with payment of redemption value (full encashment). For the purposes of this article, we will not distinguish between these two types, but we will present overall lapses of life insurance contracts. It should be noted that the lapse analysis is not only applicable to insurance but also in the bank sector – the time during which the client remains in the institution.

3 Lapse analysis using Kaplan-Meier estimator

For lapse analysis, we use Kaplan-Meier nonparametric estimator described in chapter below. Our dataset consists of 2 451 life insurance policies, where the main insured risk is death. These policies were sold between 2000 – 2015. We observed 2 345 lapses and the remaining observations were right-censored. Our analysis was performed in R programming language with package "survival" [10].

Firstly, we split survival times into intervals with the same survival time for one or more observations. Then, we estimated survival probabilities with the Kaplan-Meier method. Table 1 shows survival probability for each year of policy duration. We also add 95% confidential intervals for survival probabilities. We focus on the first 10 years of policy durations.

Table 1. Kaplan-Meier estimator of survival probability

Year	*Number of policies at risk*	*Number of lapses*	*Survival probability*	*Lower 95% CI*	*Upper 95% CI*
1	1230	1221	50,18%	0,4824	0,522

2	789	441	32,19%	0,3039	0,3409
3	670	119	27,34%	0,2563	0,2916
4	570	130	22,03%	0,2045	0,2374
5	368	172	15,01%	0,1366	0,165
6	303	65	12,36%	0,1113	0,1374
7	264	39	10,77%	0,0961	0,1207
8	238	33	9,87%	0,0876	0,1113
9	190	15	8,75%	0,077	0,0995
10	135	36	5,18%	0,0434	0,062

Source: own processing

There is a significant probability (49,82 %) that the policy of this death insurance will lapse in the first year of its duration. This could be caused by the benefit of this insurance product – policyholders may lapse their policy in the first year without giving a reason. After 10 years, there are only 5,18 % of living policies.

On figure 1, we can see that the median survival time is 365 days (intersect of purple and pink lines). The survival function is a step function with step size $\frac{number\ of\ lapses}{number\ of\ policies\ in\ risk}$ at a given time t, when a lapse occurs. At the beginning ($t = 0$) probability of survival is equal to 1 (100 %), then survival probability decreases over time and at the end of the research survival probability is zero. The survival function is rapidly decreasing in the first year and after that the decrease is more linear. Dashed line on the figure represents confidential intervals (same as in table 1). We can also see some ticks on the survival function curve, this is caused by the censoring of observations.

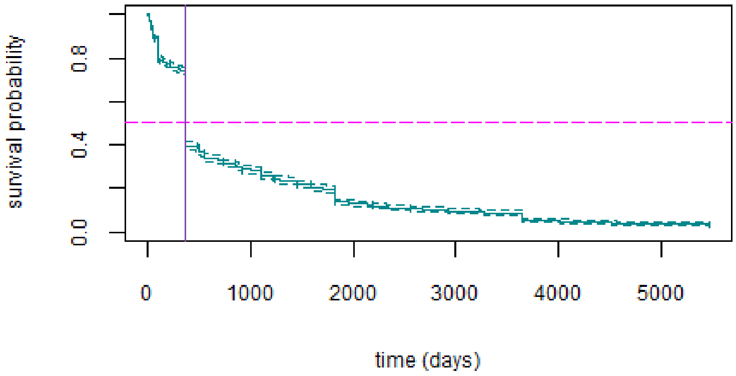

Fig. 2. Kaplan-Meier survival function. *Source: own processing*

The distribution function of time to lapse describes the cumulative probability of lapse for a policy. It is a complementary to the survival function (see formula 2). Since survival function is decreasing, the distribution function is increasing (figure 2).

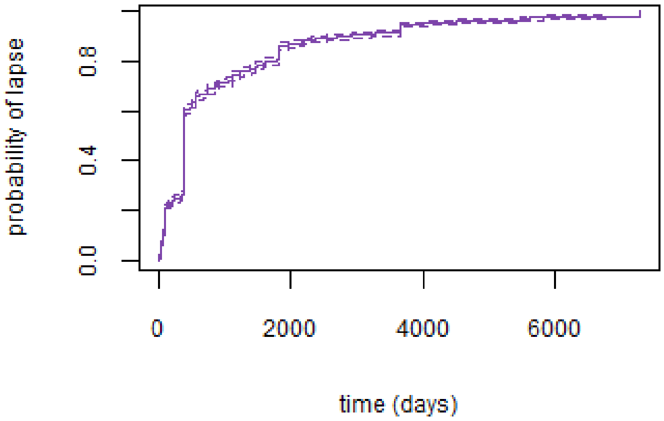

Fig. 2. Cumulative distribution function of time to lapse. *Source: own processing*

The lapse ratio (yearly lapse ratio) used in cash-flow analysis of an insurance company to calculate technical reserve is illustrated in table 2 for each year of policy duration. It means that every year number of policies decreases by lapse ratio. Reserve is recalculated every year with a corresponding number of policies.

Table 2. Yearly lapse ratio

Year	Distribution function	Lapse ratio	Number of policies
0			100 000
1	49,82%	49,82%	50 180
2	67,81%	17,99%	32 190
3	72,66%	4,85%	27 340
4	77,97%	5,31%	22 030
5	84,99%	7,02%	15 010
6	87,64%	2,65%	12 360
7	89,23%	1,59%	10 770
8	90,13%	0,90%	9 870
9	91,25%	1,12%	8 750
10	94,82%	3,57%	5 180

Source: own processing

Suppose that our new portfolio of same death insurance coverage has at the beginning 100 000 policies and no other insurance policies are sold (hypothetical portfolio), in the table 2 we can see decreasing evolution of number of policies for each year (with assumption that no policies are mature, and no insured persons die in a period of 10 years).

Lapse analysis is very important because insurance company does not have to hold the reserve for all policies at time $t = 0$, but only for an appropriate number of policies. It is also important for calculation of SCR described in chapter 1.

In the next steps of our analysis, we focused on the influence of sex on insurance lapses. We can see in table 3 (columns 2 and 4) that men are more likely to lapse their death insurance policies than women (lower survival probability). The main difference is in years 2 – 10.

The cumulative hazard function represents the overall risk of event occurrence from the beginning to the given time t. [1] The hazard function could be in some cases greater than 1, this depends on the selected time unit. [6] A higher hazard function means a lower probability of survival.

Cumulative hazard function is calculated as [1]:

$$H(t) = \int_0^t h(u)du \tag{7}$$

where $h(u)$ represents hazard function or as:

$$H(t) = -\log(S(t)) \tag{8}$$

Table 3. Kaplan-Meier estimator of survival probability for each sex

Year	*Men - Survival probability*	*Men - cumulative hazard*	*Women - Survival probability*	*Women - cumulative hazard*
1	50,00%	0,6183	50,45%	0,6201
2	31,77%	1,0446	32,80%	1,0224
3	26,13%	1,2383	29,09%	1,142
4	21,11%	1,4478	23,37%	1,3574
5	13,34%	1,856	17,45%	1,6318
6	10,52%	2,0851	15,05%	1,7777
7	8,94%	2,2464	13,44%	1,8894
8	8,04%	2,3514	11,42%	2,0493
9	7,53%	2,4164	10,52%	2,1304
10	5,60%	2,6808	8,95%	2,2855

Source: own processing

On the top of Figure 3 we can see the survival function for men (green curve) and survival function for women (pink curve). On the bottom, we can see the number of policies at risk for each sex. Survival functions for each sex are almost the same in the first year but slightly different from the beginning of the second year to the end of the research (see also table 2).

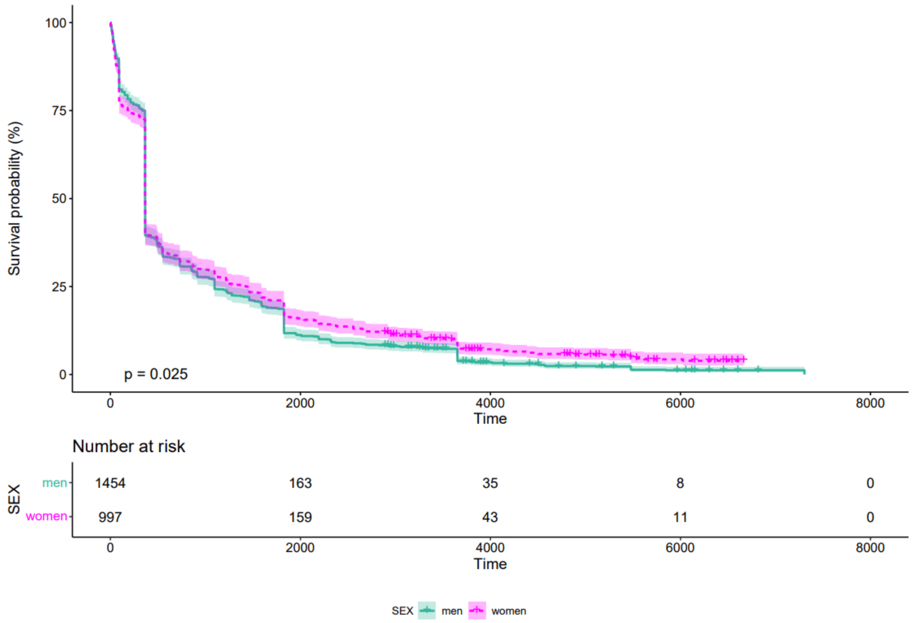

Fig. 3. Kaplan-Meier survival function for each sex. *Source: own processing*

For comparison of two or more Kaplan-Meier survival functions we used Log-rank test statistics which was compared with Chi-square test with one degree of freedom (number of compered survival functions – 1) [7]:

$$Log - rank\ test\ statistic = \frac{(O_M - E_M)^2}{E_M} + \frac{(O_W - E_W)^2}{E_W} \qquad (9)$$

where $O_{M/W}$ means observed survival time and $E_{M/W}$ means expected survival time.

We defined two statistical hypotheses:

$$H_0: S_M(t) = S_W(t) \qquad (10)$$

$$H_1: S_M(t) \neq S_W(t) \qquad (11)$$

Based on test statistics and p – value, we reject zero hypothesis, so the survival function for men is significantly different from the survival function for women.

Conclusion

Kaplan-Meier estimator is a useful statistical method to analyze survival times not only in medical research but also in the finance sector, especially in actuarial science. Kaplan-Meier method allows to work with censored observations and use the information about censored times. This is an advantage in contrast with other nonparametric methods. Since this model has easily interpretable results, this method can be simply explained to the public who does not have such knowledge in actuarial science.

In this article, we illustrate the use of survival analysis in the life insurance industry specifically in lapse analysis. We modeled the lifetime of insurance policies using the Kaplan-Meier estimator on a real dataset of death insurance policy from a universal insurance company. The main finding in model output is that 50 % of policies lapsed during the first year caused by policy benefit, which is the possibility of terminating the contract in the first year of insurance coverage without any reason. In our research, we found out that sex has a significant impact on policy lapses after the first year of insurance policy duration (men have a higher lapse ratio than women). This analysis could be extended by other survival analysis methods e.g., Cox semi-parametric regression model.

Acknowledgement

This article is a partial output of the research project of VEGA no. 1/0166/20 entitled „Determination of capital requirement to cover selected catastrophic risks in life and non-life insurance“.

References

1. Collet, D.: Modelling Survival Data in Medical Research (third edition). Published by Taylor & Francis Group, LLC, Bristol (2015).
2. Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II).
3. Gondova, A.: Regulácia kapitálových požiadaviek v poisťovníctve. NBS – BIATEC No. 9. Bratislava (2015). https://www.nbs.sk/_img/Documents/_PUBLIK_ NBS_FSR /Biatec/Rok2015/09- 2015/biatec_09_2015_04Gondova.pdf, last accessed 2022/06/01.
4. Haberman, S.: Landmarks in the history of actuarial science (up to 1919). Actuarial Research Paper No. 84. Faculty of Actuarial Science & Insurance, City University London, London (1996).
5. Kaplan, E. L., Meier, P.: Nonparametric Estimation from Incomplete Observations. Published by American Statistical Association. Journal of the American Statistical Association, Vol. 53, No. 282, pp. 457-481, USA (1958).
6. Klenbaum, D. G., Klein, M.: Survival Analysis: A Self-Learning Text. 3rd Edition, Springer, New York (2012).
7. Moore, D. F.: Applied Survival Analysis Using R. Published by Springer International Publishing, Switzerland (2016).

8. Šoltesová, T.: Aktuárske modely v životnom poistení. Vydavateľstvo Letra Edu, Bratislava (2019).
9. Teplanová, P.: Analýza prežitia a jej využitie v životnom poistení [Diplomová práca]. Ekonomická univerzita v Bratislave, Bratislava (2021).
10. Therneau, T. M.: A Package for Survival Analysis in S. version 2.38. (2015). URL:https://CRAN.R-project.org/package=survival, last accessed 2022/06/01.