COMPARISON OF ESTIMATORS FOR PROBABILITY OF DEATH USED IN ACTUARIAL SCIENCE*

KOLOS ÁGOSTON[1]

In this paper a brief summary of the estimators for probability of death is given. Three estimators will be investigated, two of them are parametric ones. To derive a parametric estimator a distribution has to be assumed. Unfortunately the distributions occurring in real life differ from these assumed ones. It will be investigated how can these estimators be applied if we assume other distributions. The bias and efficiency of the estimators are analysed by using Monte Carlo simulation.

The maximum likelihood estimator is the most common in the actuarial practice due to its appealing point estimation properties. The Kaplan–Meier estimation is a better choice, if the purpose is to give a better confidence interval.

KEYWORDS: Actuarial sciences; Estimators; Simulation.

The estimation of probabilities of death (or other failures, for instance disability) is crucial in the life insurance industry. The most often used probability[2] is that a person aged exactly *x*, will die in a year, denoted by q_x. The main goal of this paper is to give a good estimator for this probability. There is a simple estimator for ratios, and this estimator can be used for estimating the probability of death as well: *d/n*, where *d* is the number of deaths, and *n* is the number of insured lives (where everyone is exactly *x* year old at the beginning). This estimator has good properties: in the frame of a binomial model it is a maximum likelihood estimator (consistent, asymptotically unbiased, and asymptotically most efficient), and it is unbiased in small samples too.

Though this estimator has very attractive properties, its application has special requirements as well. The most important one says that there is only one reason for exit, and this is death (or failure in general). Let us suppose that there are *n* elements (persons) at the beginning, *d* of them die within a year, and the others survive. It means that there is no lapse or censoring, as statisticians say. Furthermore, let us assume that a person – say Mr. Smith – buys an insurance policy, but after half a year he withdraws his policy. From

* The author is indebted to Professor *Erzsébet Kovács*, Professor *Angus Macdonald*, *László Papp* and *Krisztina Szőke* for their help.

[1] Assistant of the Department of Operations Research, Budapest University of Economic Sciences and Public Administration.

[2] These probabilities can be found in the life tables. The premium calculation is based on these probabilities as well.

that time the company has no information whether Mr. Smith has died or survived the year. Given that we do not know this, the ratio estimator cannot be applied any more.

There are many estimators that can handle this deficiency, but there is not an universal one. The paper will compare the properties of the three most widely used estimators. These are the Kaplan–Meier estimator, which is a non-parametric one, the actuarial estimator, and the maximum likelihood estimator.

1. PROPERTIES OF THE THREE COMMONLY USED ESTIMATORS

In this section the derivation and the properties of the three mentioned estimators will be given.

The *Kaplan–Meier estimator* can be derived as follows. Let us divide the one-year interval into sub-intervals, so that only one event occurs in any of these sub-intervals: a failure or a censor event. If within the sub-interval a censor event occurs, one can say that the estimated probability (of failure) is 0 for this interval. If there is only a failure within the sub-interval, the ratio estimator can be applied. The estimated probability for the whole year is one minus the product of the estimated survival probabilities (equals one minus the estimated probability of failure).

It should be noted that the observed lifetime is the difference between the starting time and the time of exit. The cause of exit can either be death or censoring. Let us arrange the observed lifetimes in standing order, and let t_1 be the smallest and t_n is the largest observed lifetime; d_i is a dummy variable which takes the value 1 if the i^{th} person died, otherwise it is 0. Using these variables, the Kaplan–Meier estimator is:

$$\hat{q} = 1 - \prod_{i=1}^{n} \left(\frac{n-i}{n-i+1} \right)^{d_i}$$

If d_n is 1 (i.e. that the last person died) then the estimator gives 1 for the probability of death. In this case it is said that estimation of the probability of death is not defined.

In a small numerical example three persons buy a specified insurance policy. The first one withdraws his contract at the one third of the year. The second dies at the two thirds of the year, while the third survives the year. The estimated probability for the first third year is 0, for the second third year is 1/2 and for the third year is also 0. So the estimated probability for the whole year is: $1-\left(\frac{2}{3}\right)^0\left(\frac{1}{2}\right)^1\left(\frac{0}{1}\right)^0 = \frac{1}{2}$.

The *actuarial estimator* is a parametric one. Here it has to be assumed a rule of how the probability of death in a fraction of year is related to the probability of death in the whole year. According to the common actuarial notation ${}_sq_{x+t}$ means the probability of failure during the interval $(x+t, x+t+s)$ for a person who is alive at the time $x+t$ ($0 \le t,s \le 1$, $t+s \le 1$). It is also common that if the length of the interval is a whole year and this year begins on the investigated person's birthday, we do not need to write the left index: $q_x={}_1q_x$. Using these notations the assumed relation between the probabilities is:

$${}_{1-t}q_{x+t}=(1-t)q_x,$$

which is called the *Balducci* assumption. If the probability of death is 0.003 in a year, the probability of death in the interval (x+2/3, x+1) for a person who is alive at the time x+2/3 is: 1/3·0.003=0.001.

The idea behind the actuarial estimator is to adjust the estimator *d*/*n*. The number of deaths is given, so *n* has to be modified. It is better if one says that *n* means *n* years instead of *n* persons. So if somebody withdraws his contract, he does not stand in risk for a whole year, only for a fraction of year. If the person in the previous example withdraws his contract at one third of the year, he stands in risk for one third year. If somebody dies, he stands in risk for the whole year. The explanation for this is that in the case of estimator *d*/*n*, if somebody dies, he stands in risk for the whole year as well. Then the actuarial estimator is as follows: the number of deaths divided by 'standing in risk', which is called 'initial exposed to risk' in the actuarial literature. If no censoring event occurs, the actuarial estimator is equal to the common ratio estimator *d/n*.

Using this estimator for the previous example, the initial exposed to risk is 1/3+1+1=7/3, the number of deaths is 1, so the estimated probability is 1/(7/3)=0.43.

The next question is that how is this connected to Balducci hypothesis. Let us consider a small example. At the beginning there are l+w members, w of them withdraw their contract after one third year. The expected number of deaths is $(l+w)\cdot q_x$, but the insurance company does not know about the deaths of the those who withdrew after the withdrawal. We can calculate the expected number of these deaths by using the Balducci hypothesis: $w\cdot 2/3q_x$.

The expected number of deaths known by the insurance company equals the expected number of deaths in a year minus the expected number of deaths the company does not know about:

$$d=(l+w)\cdot q_x-w\cdot(2/3)q_x=l\cdot q_x+(1/3)\cdot w\cdot q_x \qquad /1/$$

If equation /1/ is arranged for q_x, we get the actuarial estimator:

$$\hat{q}_x=\frac{d}{l+\frac{1}{3}w}.$$

To derive the actuarial estimator we used the Balducci assumption. It can be proved that if the Balducci hypothesis is true then the actuarial estimator is asymptotically unbiased. (It was believed for quite a long time that the actuarial estimator is a moment estimator, but it was proved 20 years ago that this is not true.)

The *maximum likelihood estimator* is a very popular one too. Its main idea comes from realizing that however the probability of death is a good measure, it has a disadvantage as well, namely it is related to an interval. Sometimes it is better to use a measure for a point instead. We can get this measure as a limit: $\lim\limits_{t\to 0}({}_tq_x/t)$, which is denoted by μ_x and called as the *force of mortality*. The probability of death can be calculated by using the force of mortality:

$${}_tq_x=1-\exp\left(-\int_x^{x+t}\mu_\tau d\tau\right).$$

The assumption of constant force of mortality means that $\mu_x = \mu$ in the investigated interval. In other words:

$${}_sq_{x+t}=1-(1-q_x)^s.$$

If we assume the constant force of mortality, we get a maximum likelihood estimator:

$$1-\exp(-d/E_x^c),$$

where E_x^c is called 'central exposed to risk' in the actuarial literature. The 'central exposed to risk' is the sum of observed times. It can easily be calculated: if somebody withdraws his contract, the calculation is the same as in the case of the initial exposed to risk, but if somebody dies, he stands in risk only until the time of death.

Calculating the estimated probability of death for the previous example:

– Central exposed to risk: (1/3+2/3+1)=2.
– Estimated probability: 1–exp(–1/2)=1–0.61=0.39.

As it was mentioned, in the case of maximum likelihood estimator, we assume a constant force of mortality. Therefore, if the constant force of mortality is appropriate then this estimator is asymptotically unbiased and asymptotically efficient.

Up to this point we wanted to derive an estimator which we can handle easily, this is the reason for using these assumptions instead of believing that the data follow these assumptions. In real life the probability of death is increasing with age[3] so it is a good hypothesis that the force of mortality also increases during the one-year time interval. Unfortunately, the force of mortality does not increase either in the case of Balducci assumption or in the case of constant force of mortality. It is necessary to investigate the behaviour of these estimators under more realistic assumptions such as uniform distribution of deaths[4] or a Gompertz mortality law.[5]

2. MONTE CARLO SIMULATIONS

In order to compare the properties of the former described estimators Monte Carlo simulation are used. Three scenarios for the probability of failure (1%, 5%, 30%), and three scenarios for the sample size (5, 30 and 1 000) are chosen. The uniform distribution of death does not have any parameter, so the situation in the case of these assumption is simple.

[3] This statement holds for all ages greater than 6 in Hungary, and holds for ages greater than 25 in western countries as well.

[4] Uniform distribution of death means that the (expected) number of deaths is the same for all intervals whose lengths are equal. Let us suppose that the probability of death is 0.1 and the sample size is 100 (for instance). It means that 10 deaths will occur (expectedly), 5 of them in the first half-year, 5 of them in the second half-year. The probability of death in the first half-year is: 5/100, and in the second half-year 5/95 (because 5 persons died in the first half-year, so there are only 95 persons alive at the beginning of the second half-year). We can see that the probability of death increases during the year.

[5] Gompertz described an expression for the force of mortality: $\mu_x = Bc^x$. $B > 0$, and B is close to 0, $c > 1$ and c is close to 1. In this case the force of mortality increases so the probability of death increases as well.

The Gompertz mortality law has two parameters but one of them (B) is irrelevant for our analysis.[6] The other parameter was fixed as $c = 1.1$.[7]

Two numbers are simulated for each person. One for failure time[8] and one for censoring time, and results are stored. We will investigate the described assumptions for failure time, but we will use the constant force of mortality for censoring time (to reduce the number of scenarios). The probability of censoring (the probability that a censoring event occurs within a year) is 10 percent (for the same reason). A censoring time and a survival time for each member of the sample will be simulated. So we can calculate simulated estimates. We will repeat it 100 000 times, so 100 000 simulated values for each estimates will be got.

For each estimator the mean and standard deviation will be presented and the estimators will be compared by means of the mean square error (MSE).

It has been mentioned that parametric estimators perform better because their variance is smaller. For this reason we start the simulation by comparing the Kaplan–Meier estimator and the actuarial estimator when the Balducci hypothesis holds for the failure times, then we will compare Kaplan–Meier estimator with the maximum-likelihood estimator when constant force of mortality holds for the failure times. The results can be seen in Tables 1,2,3 and 4.[9]

Table 1

Probability of death

(estimated by actuarial versus Kaplan–Meier estimators)

Sample size	True parameter	Actuarial estimator			Kaplan–Meier estimator		
		Mean	Standard deviation	MSE	Mean	Standard deviation	MSE
5	0.01	.00984106	.04687197	.0021970068	.00989700	.04761310	.0022670183
	0.05	.04951430	.10266833	.0105410215	.04998000	.10514971	.0110564611
	0.30	**.29686623**	.21397758	.0457962241	.29901983	.22096043	.0488244723
30	0.01	.01003072	.01929079	.0003721355	.01004699	.01938276	.0003756935
	0.05	.05005749	.04202622	.0017662061	.05012668	.04229819	.0017891531
	0.30	.29949622	.08721694	.0076070486	.29985433	.08878433	.0078826779
1000	0.01	.00999492	.00332315	.0000110434	.00999542	.00333146	.0000110986
	0.05	.05001927	.00727703	.0000529556	.05001958	.00730579	.0000533750
	0.30	.30005796	.01506360	.0002269155	.30006824	.01529334	.0002338910

Note: Numbers in bold mean that the mean differs from the theoretical parameter value at a 5 percent significance level.

[6] We have to know how ${}_tq_x$ is related to q_x. In case of Gompertz mortality law ${}_tp_x = \exp\left(-\int_x^{x+t} Bc^{\tau} d\tau\right) = \exp\left(-(B/\ln(c))c^x\left(c^t-1\right)\right)$. So $\ln({}_tp_x)/\ln(p_x) = \left(c^t-1\right)/(c-1) \Rightarrow {}_tp_x = \exp\left(\ln(p_x)\left(c^t-1\right)/(c-1)\right)$.

[7] This value is appropriate for the Hungarian life tables.

[8] In the simulation first a random probability is simulated (*rnd*). If this number is greater than the probability of failure then the person survives the interval, so the failure time is 1. Else a t value is sought so ${}_tq_x$ equals the simulated number (probability). This t will be the simulated failure time.

[9] When the last person died we defined the value of Kaplan-Meier estimator as 1. This event does not occur when the sample size is 30 or 1000. It occurs a few times when the sample size is 5 but the probability of death is small. It occurs in 0.5 percent of the cases when the sample size is 5 and the probability of death is 30 percent.

Table 2

t and p-values for testing means in Table 1

Sample size	Probability of death	Maximum likelihood estimator		Kaplan–Meier estimator	
		t	*p*	*t*	*p*
5	0.01	-1.072	0.284	-0.684	0.494
	0.05	-1.496	0.135	0.060	0.952
	0.30	-4.631	0.000	-1.403	0.161
30	0.01	0.504	0.614	0.767	0.443
	0.05	0.433	0.665	0.947	0.344
	0.30	-1.827	0.068	-0.519	0.604
1000	0.01	-0.484	0.628	-0.434	0.664
	0.05	0.837	0.403	0.847	0.397
	0.30	1.217	0.224	1.411	0.158

Table 3

Probability of death

(estimated by maximum likelihood versus Kaplan–Meier estimators)

Sample size	True parameter	Maximum likelihood estimator			Kaplan–Meier estimator		
		Mean	Standard deviation	MSE	Mean	Standard deviation	MSE
5	00.01	.00998304	.04590550	.0021073153	.01003533	.04634266	.0021476433
	0.05	.04978460	.09999569	.0099991834	.05005150	.10099568	.0102001306
	0.30	**.29751627**	.20653649	.0426634866	.30079983	.21285520	.0453079778
30	0.01	.01008313	.01874080	.0003512246	.01008447	.01875697	.0003518311
	0.05	**.05025774**	.04094795	.0016768008	**.05026694**	.04101747	.0016825041
	0.30	**.29936770**	.08510391	.0072430761	.29965704	.08605093	.0074048795
1000	0.01	.00999778	.00322633	.0000104092	.00999807	.00322852	.0000104234
	0.05	.05002049	.00704600	.0000496465	.05002204	.00705419	.0000497620
	0.30	.30003392	.01479451	.0002188786	.30004775	.01492556	.0002227747

Note: Numbers in bold mean that the mean differs from the theoretical parameter value at a 5 percent significance level.

Table 4

t and p-values for testing means in Table 3

Sample size	Probability of death	Maximum likelihood estimator		Kaplan–Meier estimator	
		t	*p*	*t*	*p*
5	0.01	–0.117	0.907	0.241	0.810
	0.05	-0.681	0.496	0.161	0.872
	0.30	-3.803	0.000	1.188	0.235
30	0.01	1.403	0.161	1.424	0.154
	0.05	1.990	0.047	2.058	0.040
	0.30	-2.349	0.019	-1.260	0.208
1000	0.01	-0.217	0.828	-0.189	0.850
	0.05	0.920	0.358	0.988	0.323
	0.30	0.725	0.468	1.012	0.312

Some conclusions can be made from Table 1 and Table 3. According to the theory there is a bias for each estimator (if the sample size is finite). In case of actuarial and maximum likelihood estimator we can verify this result for small samples and large probabilities. One cannot reveal this bias (with *t*-test) in case of Kaplan-Meier estimator. According to the theory all three estimators are asymptotically unbiased. It can be seen that the simulated estimates are getting closer to their true value as the sample size increases. Not only the bias decreases with the sample size, but also the *t*-value (in absolute terms) decreases (see Table 2 and Table 4), although this statement holds for large probabilities only.

The MSE for the parametric estimators are smaller than those for the Kaplan-Meier estimator. For small samples the Kaplan-Meier estimator is more precise, but as the sample size increases the bias becomes smaller in case of parametric estimators. However the standard deviation is smaller for the parametric estimators and it will result in that the mean square error is smaller for these estimators in any (of these) cases. In both cases the differences in efficiency are tiny.

After investigating the properties of the estimators, when the appropriate distribution was assumed, we can analyse how they behave when other distributions fitting better to real life practice are assumed. The results of this sensitivity analysis can be seen in Table 5 and Table 7, while Table 6 and Table 8 contain *t* and *p* values.

According to the results shown in Table 5 and Table 7, the Kaplan–Meier estimator can be considered as an unbiased estimator, while actuarial estimator and maximum likelihood estimator are biased. This bias becomes significant when both the probability of failure and the sample size are large. The bias is decreasing with the sample size (in case of actuarial and maximum likelihood estimators), but the *t*-statistics is increasing (in absolute terms), so we can conclude that there are significant biases in case of these estimators.

The standard deviations are smaller in case of actuarial and maximum likelihood estimators. The MSE for the actuarial and maximum likelihood are smaller than that for the Kaplan–Meier. However the parametric estimators are not unbiased any more, they are preferred to the non-parametric estimator (with respect to MSE), but the difference is tiny again. When the probability of death is small, the actuarial estimator is more efficient than the maximum likelihood estimator.

Table 5

Probability of death
(estimated by Monte Carlo simulations – uniform distribution of deaths)

Sample size	True parameter	Actuarial estimator			Maximum likelihood estimator			Kaplan–Meier estimator		
		Mean	Standard deviation	MSE	Mean	Standard deviation	MSE	Mean	Standard deviation	MSE
5	0.01	.01004973	.04603766	.0021194683	.01006415	.04615370	.0021301681	.01007950	.04635700	.0021489781
	0.05	.04941703	.09926364	.0098536101	**.04937001**	.09923540	.0098480609	.04969042	.10034615	.0100694465
	0.30	**.29758266**	.20853141	.0434911915	**.29278514**	.20245981	.0410420288	.30011650	.21253080	.0451693556
30	0.01	.01007813	.01872772	.0003507335	.01008056	.01873355	.0003509522	.01009136	.01876892	.0003522808
	0.05	.05016675	.04085613	.0016692513	.05017393	.04086155	.0016696964	.05023810	.04098634	.0016799370
	0.30	**.29839330**	.08518692	.0072593935	**.29619100**	.08362243	.0070072192	.30012069	.08619878	.0074302436
1000	0.01	.00999218	.00323165	.0000104436	.00999311	.00323222	.0000104473	.00999366	.00323391	.0000104582
	0.05	**.04993466**	.00706568	.0000199281	**.04994595**	.00706792	.0000499584	.04997909	.00708180	.0000501523
	0.30	**.29846589**	.01476271	.0002202910	**.29663580**	.01451429	.0002219826	.30004328	.01492269	.0002226885

Note: Numbers in bold mean that the mean differs from the theoretical parameter value at a 5 percent significance level.

Table 6

t and p-values for testing means in Table 5

Sample size	Probability of death	Actuarial estimator		Maximum likelihood estimator		Kaplan–Meier estimator	
		t	*p*	*t*	*p*	*t*	*p*
5	0.01	0.342	0.732	0.446	0.656	0.542	0.588
	0.05	-1.857	0.063	-2.008	0.045	-0.976	0.329
	0.30	-3.666	0.000	-11.269	0.000	0.173	0.863
30	0.01	1.319	0.187	1.360	0.174	1.539	0.124
	0.05	1.291	0.197	1.346	0.178	1.837	0.066
	0.30	-5.964	0.000	-14.404	0.000	0.443	0.658
1000	0.01	-0.765	0.444	-0.674	0.500	-0.620	0.535
	0.05	-2.924	0.003	-2.418	0.016	-0.934	0.350
	0.30	-32.861	0.000	-73.291	0.000	0.917	0.359

Table 7

Probability of death

(estimated by Monte Carlo simulations – Gompertz mortality law)

Sample size	True parameter	Actuarial estimator			Maximum likelihood estimator			Kaplan–Meier estimator		
		Mean	Standard deviation	MSE	Mean	Standard deviation	MSE	Mean	Standard deviation	MSE
5	0.01	.00993604	.04561098	.0020803653	.00994494	.04571968	.0020902923	.00998842	.04607912	.0021232851
	0.05	.04977052	.09976533	.0099531747	.04968104	.09965294	.0099308098	.05002325	.10075283	.0101511326
	0.30	**.29766343**	.20906922	.0437154001	**.29520474**	.20511519	.0420952368	.29964883	.21270307	.0452427209
30	0.01	.00996515	.01861666	.0003465813	.00997956	.01865877	.0003466912	.00997956	.01865877	.0003481502
	0.05	.05000431	.04072089	.0016581909	.04999852	.04070856	.0016571870	.05009891	.04087626	.0016708781
	0.30	**.29891286**	.08491974	.0072125444	**.29871793**	.08441351	.0071272851	.30007567	.08581999	.0073650771
1000	0.01	.00998985	.00322569	.0000104052	.00998997	.00322570	.0000104052	.00999882	.00323103	.0000104396
	0.05	**.04993295**	.00704448	.0000496292	**.04993493**	.00704485	.0000496342	.04999675	.00706198	.0000498715
	0.30	**.29898562**	.01478491	.0002196227	**.29905454**	.01472266	.0002176507	.29998348	.01492052	.0002226223

Note. Numbers in bold mean that the mean differs from the theoretical parameter value at a 5 percent significance level.

Table 8

t and p-values for testing means in Table 5

Sample size	Probability of death	Actuarial	Estimator	Maximum likelihood	Estimator	Kaplan–Meier	Estimator
		t	*p*	*t*	*p*	*t*	*p*
5	0.01	-0.443	0.658	-0.381	0.703	-0.079	0.937
	0.05	-0.727	0.467	-1.012	0.312	0.073	0.942
	0.30	-3.534	0.000	-7.393	0.000	-0.522	0.602
30	0.01	-0.592	0.554	-0.582	0.561	-0.346	0.729
	0.05	0.033	0.974	-0.012	0.990	0.765	0.444
	0.30	-4.048	0.000	-4.803	0.000	0.279	0.780
1000	0.01	-0.995	0.320	-0.984	0.325	-0.116	0.908
	0.05	-3.010	0.003	-2.921	0.003	-0.145	0.885
	0.30	-21.696	0.000	-20.307	0.000	-0.350	0.726

In real life practice, the exact variances of the estimators are unknown, so they have to be estimated as well using the following variance estimators:

Kaplan–Meier: $(1-\hat{q})^2\left(\sum_{i=1}^{n}((n-i)(n-i+1))^{-1}d_i\right)$

Actuarial : $\dfrac{\hat{q}(1-\hat{q})}{E_x}$

Maximum likelihood: $(1-\hat{q})^2\left(d/(E_x^c)^2\right)$

In case of actuarial estimator the estimated variance is appropriate when there is no censoring event (binomial model). In practice it is common to say that this estimation also holds when there are censoring events as well. The estimated variance of maximum likelihood estimator is derived by using the Cramer–Rao lower bound. Estimated standard errors can be seen in Table 9.

Table 9

Estimated standard errors

Sample size	Probability of death	Actuarial estimator	Maximum likelihood estimator	Kaplan–Meier estimator*
		Uniform distribution of deaths		
5	0.01	.04063340	.04062139	.04076844
	0.05	.08851087	.08822361	.08898437
	0.30	.18657172	.18334180	.18753764
30	0.01	.01839034	.01839380	.01842987
	0.05	.04017906	.04017366	.04028098
	0.30	.08406086	.08312117	.08474512
1000	0.01	.00322649	.00322677	.00322877
	0.05	.00706318	.00706393	.00707581
	0.30	.01480634	.01465886	.01492114
		Gompertz mortality law		
5	0.01	.04041854	.04037235	.04064237
	0.05	.08870958	.08835855	.08916791
	0.30	.18640674	.18383433	.18713953
30	0.01	.01828577	.01828596	.01832763
	0.05	.04012449	.04010929	.04024556
	0.30	.08410439	.08355575	.08466530
1000	0.01	.00322613	.00322615	.00323082
	0.05	.00706322	.00706269	.00707853
	0.30	.01481115	.01473583	.01490333

* If the last person died, the expression of estimated variance is not correct (we have to divide by 0). In this case we set the variance 0. As it was mentioned earlier, it did not occur when the sample size is 30 or greater.

In Table 9 we can see that the variance is under-estimated for small samples. This bias decreases as the sample size increases, but it does not disappear.

Our further purpose is to give an interval estimation for the probability of death. So for each case a 95 percent probability interval will be calculated for the estimated prob ability, i.e. we have 100 000 confidence intervals for each scenario. Then we count the number of cases when the confidence interval does not contain the estimated parameter. If the estimation process is appropriate, the number of these cases has to be very close to 5000, that is their share must be close to 5 percent. Table 10 presents the results.

Table 10

Share of cases when the confidence interval does not contain the actual parameter

Sample size	Probability of death	Actuarial estimator	Maximum likelihood estimator	Kaplan–Meier estimator
		percent		
		Uniform distribution of deaths		
5	0.01	95.32	95.33	95.33
	0.05	78.59	78.60	78.60
	0.30	21.56	21.60	21.90
30	0.01	74.99	74.99	74.99
	0.05	23.30	23.31	23.30
	0.30	6.13	6.58	6.21
1000	0.01	9.63	9.62	8.72
	0.05	5.48	5.46	5.47
	0.30	5.17	5.61	5.01
		Gompertz mortality law		
5	0.01	95.35	95.36	95.36
	0.05	78.49	78.50	78.50
	0.30	21.63	21.92	21.95
30	0.01	75.21	75.21	75.21
	0.05	23.21	23.21	23.21
	0.30	6.08	6.62	6.12
1000	0.01	9.55	9.55	8.56
	0.05	5.36	5.37	5.38
	0.30	5.11	5.18	5.09

We can see that the results are very poor, especially for small sample sizes. This is an important reason why these estimators cannot be used for small samples. According to Table 10, maximum likelihood estimator achieves the worst results. If the probability of death is small (which is the most relevant case in insurance problems) and the sample size is large enough, the performance of Kaplan–Meier estimator is the best. In case of Kaplan–Meier estimator the confidence interval is wider (since the estimated variance is larger), but this wider interval adheres more to the facts.

3. CONCLUSION

All these three estimators have almost the same properties. The parametric estimators are more robust, i.e. they perform well when the inappropriate distributions are assumed. They are preferred to Kaplan–Meier estimator with respect to MSE, but the difference in efficiency is rather small. As we have seen, the confidence interval for Kaplan–Meier estimator is wider, but this wider interval covers the true value more frequently. The differences in frequency are small again. If the sample size tends to infinity, the MSE for Kaplan–Meier estimator tends to 0 (since it is a consistent estimator), but the MSE for parametric estimators keeps to a positive value, since there is a (significant) bias in this case. Based upon these results the use the Kaplan–Meier estimator can be suggested in life insurance statistics.

REFERENCES

ÁGOSTON, K. – KOVÁCS, E. (2000): *Halandósági modellek*. Budapesti Közgazdaságtudományi és Államigazgatási Egyetem, Operációkutatási Tanszék. Budapest. Aktuárius jegyzetek 3.

BATTEN, R.W. (1978): *Mortality table construction*. Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

BENJAMIN, B. – POLLARD, J. H. (1992): *The analysis of mortality and other actuarial statistics.* Butterworth-Heinemann Ltd. Linacre House, Jordan Hill, Oxford.

DORRINGTON, R. E. – SLAWSKI, J. K. (1993): A defence of the conventional actuarial approach to the estimation of the exposed-to-risk. *Scandinavian Actuarial Journal*, p. 107–113.

HOEM, J. M. (1984): A flaw in actuarial exposed-to-risk theory. *Scandinavian Actuarial Journal*, p. 187–194.

KAPLAN, E. L. – MEIER, P. (1958): Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, Vol. 53. p. 457–481.

Lee, E. T. (1992) : *Statistical methods for survival data analysis.* John Wiley & Sons, Inc. New York, Chichester, Brisbane, Toronto, Singapore.

MACDONALD, A. S. (1996): An actuarial survey of statistical models for decrement and transition data. I: Multiple state, Poisson and binomial models. *British Actuarial Journal*, No. 2. p. 129–155.; II: Competing risks, non-parametric and regression models. *British Actuarial Journal*, No. 2. p. 429–448.; III: Counting process models. *British Actuarial Journal*, No. 2. p. 703–726.

ISSN 0039 0690

Published by the Central Statistical Office
Editor in charge: dr. Tamás Mellár
Executive editor: dr. László Hunyadi
Printed by the Akadémiai Nyomda
3340 – Martonvásár, 2001
Director in charge: Lajos Reisenleitner

Managing editor: Orsolya Dobokayné Szabó
Editors: dr. Attila Domokos, Andrea Polyák, Mariann Szűcsné Bruckner, Mária Visi Lakatos
Technical editors: Éva Bálinthné Bartha, Ágnes Simonné Káli

Editorial office: Budapest II., Keleti Károly utca 5–7. Postal address: P.O.B. 51. Budapest, 1525.
Phone: (361)-487-4343, Telefax: (361)-487-4344,
Internet: www.ksh.hu/statszml E-mail: statszemle@ksh.gov.hu
Publishing office: Central Statistical Office, Budapest II., Keleti Károly utca 5–7.
Postal address: P.O.B. 51. Budapest, 1525. Phone: (361)-345-6000
The publication can be purchased at the Statistical Special Bookshop:
Budapest II. Keleti Károly utca 10. Phone: (361)-212-4348