Available online at http://www.ssstj.sci.ssru.ac.th

Suan Sunandha Science and Technology Journal
©2016 Faculty of Science and Technology, Suan Sunandha Rajabhat University

Minimum K-S estimator using PH-transform technique

Somchit Boonthiem, Supawan Khotama, Supap Sakha, Watcharin Klongdee*

Risk and Insurance Research Group, Department of Mathematics, Faculty of Science, Khon Kaen University
123 Moo 16 Mittapap Rd., Nai-Muang, Muang District, Khon Kaen 40002, Thailand

Corresponding author E-mail: *kwatch@kku.ac.th

Abstract: In this paper, we propose an improvement of the minimum Kolmogorov-Smirnov (K-S) estimator using proportional hazards transform (PH-transform) technique. The data of experiment is 47 fire accidents data of an insurance company in Thailand. This experiment has two operations, the first operation, we minimize K-S statistic value using grid search technique for nine distributions; Rayleigh distribution, gamma distribution, Pareto distribution, log-logistic distribution, logistic distribution, normal distribution, Weibull distribution, log-normal distribution, and exponential distribution and the second operation, we improve K-S statistic using PH-transform. The result appears that PH-transform technique can improve the minimum K-S estimator. The algorithms give better the minimum K-S estimator for seven distributions; Rayleigh distribution, logistic distribution, gamma distribution, Pareto distribution, log-logistic distribution, normal distribution, Weibull distribution, log-normal distribution, and exponential distribution while the minimum K-S estimators of normal distribution and logistic distribution are unchanged.

Keywords: Minimum K-S estimator, PH-transform technique, Grid search technique.

1. Introduction

Majority data analysis methods depend upon the assumption that data were sampled from a normal distribution or at least from a distribution which is sufficiently close to a normal distribution that so called parametric test. If the conditions for the parametric are not met, non-parametric tests are useful in this situations. The Kolmogorov-Smirnov (K-S) statistic is a well-known nonparametric statistic test used to solve the goodness of fit (GOF) between a hypothesized distribution function and an empirical distribution function. The K-S statistic test is the same as Chi-Square test but K-S statistic consider the evaluating of maximum distant between the empirical cumulative distribution function and the theoretical cumulative distribution function. The proportional hazards transform (PH-transform) has been proposed to calculate the risk adjusted premium by (Wang, S., 1995). Furthermore, PH-transform can be used to quantify risk process, risk parameter, and risk dependency.

In this paper, we are interested in minimum K-S estimator using PH-transform technique to improve the parameter estimation. Related research such as an algorithm for computing the parameter estimates in a univariate probability model for a continuous random variable that minimizes the K–S statistic presented and implemented by (Weber, M.D., et al., 2006). The algorithm uses an evolutionary optimization technique to solve for the estimates. Several simulation experiments demonstrate the effectiveness of this approach. The tool is modified by extending it in order to use the Kaplan-Meier estimate of the cumulative distribution function for distribution function (CDF) for right-censored data. The algorithm computes Minimum K-S estimators for several different continuous univariate distributions, uses an evolutionary optimization algorithm, and recommends the distribution and parameter estimates that best minimize the K-S statistic (Wieczorek, J., 2009).

In the next section, we introduce materials and methods, and interpret definition of K-S statistic and PH-transform including distributions that using in this experiment and maximum likelihood estimator (MLE). In section 3, we describe the process of experiment and the results. In section 4, we conclude and discuss the future applications of the outlined method here.

2. Materials and Methods

2.1 Kolmogorov-Smirnov statistic

The Kolmogorov-Smirnov test is based on the following mathematical definition.

Definition 1. Let $x_1, x_2, \dots, x_N$ such that $x_1 < x_2 < \cdots < x_N$ be a sample of N independent and identically distributed observations of a real-valued one-dimensional random variable X that has parameter vector θ. The cumulative distribution function (CDF) of X is denoted by $F_X(x;\theta)$. The Kolmogorov-Smirnov statistic (K-S statistic) of $F_X(x;\theta)$ is given by

$$D(\theta) = \max_{1\le i\le N} \left| F_X(x_i;\theta) - \frac{i}{N} \right|.$$

2.2 Grid Search

Grid search is a traditional algorithm of performing hyper-parameter optimization which is a simple exhaustive searching through a manually specified subset of the hyper-parameter space of learning algorithm.

In this paper, we use grid search for minimizing the K-S statistic in order to obtain a set of parameters. Then, we improve the parameters by using PH-transform. The algorithm of grid search has 5 steps as shown in section 3.

2.3 PH-transform

In casualty insurance, a risk is a non-negative random loss X defined by its cumulative distribution function,

$$F_X(t) = \Pr(X \le t)$$

or survivor function,

$$S_X(t) = 1 - F_X(t).$$

The proportional hazards transform is based on the following mathematical definition.

Definition 2. Give any random variable X with survivor function $S_X(t)$, the equation

$$S_Y(t) = [S_X(t)]^{\frac{1}{\rho}}, \rho > 0.$$

Define another random variable Y with survivor function $S_Y(t)$. The mapping $S_X \longmapsto S_Y$ called the proportional hazards transform (PH-transform).

In this paper, we use PH-transform to change cumulative distribution function form for improving the better K-S statistics.

2.3 Distributions are used in experiment

In this section, we present some properties of experiment as location, scale and shape parameters of the Rayleigh, logistic, Gamma, Pareto, log-logistic, normal, Weibull, log-normal, and exponential distributions as seen in (Sinsomboonthong, S. 2015).

2.3.1 Rayleigh distribution

The probability density function of the Rayleigh distribution has a scale parameter σ. The probability density function is given by

$$f_X(x;\sigma) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \qquad x \ge 0. \tag{1}$$

The cumulative distribution function is in the form

$$F_X(x;\sigma) = 1 - e^{-\frac{x^2}{2\sigma^2}}, \qquad x \ge 0. \tag{2}$$

2.3.2 Logistic distribution

The logistic distribution has a positive real scale parameter σ. The probability density function is given by

$$f_X(x;\ \sigma) = \frac{1}{\sigma} \frac{e^{-\left(\frac{x}{\sigma}\right)}}{\left(1 + e^{-\left(\frac{x}{\sigma}\right)}\right)^2}, x \in \mathbb{R}. \tag{3}$$

The cumulative distribution function is in the form

$$F_X(x;\ \sigma) = \frac{1}{1 + e^{-\left(\frac{x}{\sigma}\right)}}, x \in \mathbb{R}. \tag{4}$$

2.3.3 Gamma distribution

The Gamma distribution has a positive shape parameter α and a positive scale parameter β. The probability density function is given by

$$f_X(x;\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, x \ge 0, \tag{5}$$

where $\Gamma(\alpha)$ is the Gamma function evaluated at α. The cumulative distribution function is in the form

$$F_X(x;\alpha,\beta)=\frac{\gamma\left(\alpha,\frac{x}{\beta}\right)}{\Gamma(\alpha)},x\geq 0, \qquad (6)$$

where $\gamma\left(\alpha,\frac{x}{\beta}\right)$ is the lower incomplete Gamma function.

2.3.4 Pareto distribution

The Pareto distribution has a positive shape parameter α and a positive scale parameter β. The probability density function is given by

$$f_X(x;\alpha,\beta)=\frac{\alpha\beta^{\alpha}}{x^{\alpha+1}},x\geq\beta. \qquad (7)$$

The cumulative distribution function is in the form

$$F_X(x;\alpha,\beta)=1-\left(\frac{\beta}{x}\right)^{\alpha},x\geq\beta. \qquad (8)$$

2.3.5 Log-logistic distribution

The log-logistic distribution has a real location parameter μ and a positive real scale parameter σ. The probability density function is given by

$$f_X(x;\mu,\sigma)=\frac{1}{\sigma}\frac{1}{x}\frac{e^{z}}{(1+e^{z})^{2}},x\geq 0. \qquad (9)$$

where $z=\frac{\ln(x)-\mu}{\sigma}$. The cumulative distribution function is in the form

$$F_X(x;\mu,\sigma)=\left(1+\left(\frac{x}{\mu}\right)^{-\sigma}\right)^{-1},x\geq 0. \qquad (10)$$

2.3.6 Normal distribution

The normal distribution or, as it is often called, the Gaussian distribution is the most important distribution in statistics which has a real location parameter μ and a positive real scale parameter σ. The distribution is given by

$$f_X(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}},\ x\in\mathbb{R}. \qquad (11)$$

The cumulative distribution function is in the form

$$F_X(x;\mu,\sigma)=\frac{1}{2}\left(1+\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right),x\in\mathbb{R} \qquad (12)$$

where $\mathrm{erf}(t)=\frac{2}{\sqrt{\pi}}\int_0^t e^{-x^2}dx$.

2.3.7 Weibull distribution

The Weibull distribution has a positive real shape parameter α and a positive real scale parameter β. The distribution is given by

$$f_X(x;\alpha,\beta)=\frac{\alpha x^{\alpha-1}}{\beta^{\alpha}}e^{-\left(\frac{x}{\beta}\right)^{\alpha}},x\geq 0. \qquad (13)$$

The cumulative distribution function is in the form

$$F_X(x;\alpha,\beta)=1-e^{-\left(\frac{x}{\beta}\right)^{\alpha}},x\geq 0. \qquad (14)$$

2.3.8 Log-normal distribution

The log-normal distribution has a real location parameter μ and a positive real scale parameter σ. The probability density function is given by

$$f_X(x;\mu,\sigma)=\frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^{2}}{2\sigma^{2}}},x>0. \qquad (15)$$

The cumulative distribution function is in the form

$$F_X(x;\mu,\sigma)=\frac{1}{2}+\frac{1}{2}\mathrm{erf}\left[\frac{\ln x-\mu}{\sqrt{2}\sigma}\right]. \qquad (16)$$

2.3.9 Exponential distribution

The exponential distribution has a scale parameter β. The probability density function is given by

$$f_X(x;\beta)=\frac{1}{\beta}e^{-\frac{x}{\beta}},x\geq 0. \qquad (17)$$

The cumulative distribution function is in the form

$$F_X(x;\beta)=1-e^{-\frac{x}{\beta}},x\geq 0. \qquad (18)$$

2.4 Maximum Likelihood Estimator

The Maximum Likelihood Estimator (MLE) is a method for estimating a parameter based on a random sample. The basic idea is to choose a value for the parameter that maximizes the probability of the observations that actually occurred in the random sample. Although this approach is most fruitful in the context of continuous population random variables with parameters that take values along a continuum of real numbers, it also works for discrete random variables and parameters taking discrete values. In this paper we use the MLE to estimate the initial parameter for some distributions as following definition as seen in (Leonard, A.A., and Mark, M.M., 2015).

Definition 3. Let X be a population random variable with a parameter $\theta.$ Assuming X is continuous, the density function as $f_X(x;\theta)$ in order to emphasize the dependency on θ as well as $x.$ The likelihood function is given by

$$L(\theta;x_1,x_2,\dots,x_n)=\prod_{i=1}^{n} f_X\,(x_i;\theta) \qquad (19)$$

taking logarithm in (19), $\ln L(x;\theta),$ is called log-likelihood function that solved by taking derivative of $\ln L(x;\theta)$ with respect to the parameter $\theta,$ and setting it equal to zero in order to solve for $\theta.$ The value of θ that obtained from maximizing the function $\ln L(x;\theta)$ is exactly that the same value of $\theta,$ is called maximum likelihood estimator.

3. Results and Discussions

This experiment uses some of fire data of an insurance company in Thailand. The purpose of this paper is to improve minimum K-S estimator using PH-transform technique and we have two operations of experiment. The first operation, we minimize K-S statistic value using grid search technique for nine distributions. The data consist of the claim size as shown in Table 1.

We used MLE to compute the initial parameters. After that, we calculate the K-S statistic as shown in Table 2.
Next, minimum K-S estimator estimates parameters that minimizes the K-S statistic. Minimum K-S estimator is a numerical optimization method that moves from the initial parameter to a better solution. The Minimum K-S estimator using grid search algorithm can describe as following:

Step 1: Compute the K-S statistic from the initial parameters a, b or d where a is scale parameter, b is location parameter and d is shape parameter.

Step 2: Randomly change the parameter value. We can do this by choosing constant r_1, r_2 or r_3 (real number) and let $a' = a + r_1, b' = b + r_2,$ and $d' = d + r_3.$

Step 3: Compute the K-S statistic value with a, b or $d.$

Step 4: Compare the K-S statistic value which were obtained by step 1 and 3. If the K-S statistic value of the step 3 is greater than step 1, then repeat step 2 with set $a = \acute{a}, b = \acute{b}$ and $d = \acute{d}.$ Otherwise, we choose a new constant r_1, r_2 or r_3 then go on to step 2.

Table 1. The claim size of fire insurance in Thailand (million)

Times	Claim size	Times	Claim size	Times	Claim size
1	35.5	**17**	26.7	**33**	69.9
2	26.4	**18**	32.4	**34**	122.7
3	64.9	**19**	76.5	**35**	158.9
4	127.3	**20**	33.2	**36**	33.1
5	57.7	**21**	25.7	**37**	60
6	21.8	**22**	60.2	**38**	104.3
7	67.3	**23**	132.2	**39**	29.2
8	48.5	**24**	20.9	**40**	63.1
9	23.6	**25**	65.8	**41**	90
10	22.3	**26**	55.3	**42**	27.2
11	84.6	**27**	33	**43**	22.4
12	21.4	**28**	22.1	**44**	27.5
13	51.5	**29**	24.2	**45**	57.2
14	20.7	**30**	44.4	**46**	34.2
15	40.1	**31**	20.4	**47**	53.2
16	29.3	**32**	30.8		

Table 2. The K-S statistic by computing from initial parameter

Distribution	Parameters			K-S statistic value
	Scale	Location	Shape	
Rayleigh	43.02991	-	-	0.22218
Logistic	18.25640	-	-	0.87754
Gamma	16.89217	-	3.02242	0.16515
Pareto	20.40000	-	1.34604	0.11675
Log-logistic	0.34005	3.72054	-	0.12438
Normal	33.11345	51.05532	-	0.19137
Weibull	1.69222	-	57.75607	0.15542
Log normal	0.57106	3.75845	-	0.14337
Exponential	51.05532	-	-	0.30811

Step 5: Repeat step 2 to step 4 until the K-S statistic value remains constant. The results are shown in the following Table 3.

In Table 2 - 3**,** we found that K-S statistic of distributions in Table 3 is better than K-S statistic of distributions in Table 2, except the exponential distribution. We use 200 situations of simulation for our criteria in Table 3.

Table 3. Minimum K-S estimator

Distribution	Parameters			K-S statistic value
	Scale	Location	Shape	
Rayleigh	36.05355	-	-	0.12705
Logistic	36.99975	-	-	0.44620
Gamma	11.35555	-	3.85453	0.10397
Pareto	20.78183	-	1.24468	0.08111
Log-logistic	0.37615	3.72054	-	0.11197
Normal	20.14758	41.34690	-	0.12797
Weibull	1.62774	-	57.75607	0.14661
Log normal	0.68813	3.75845	-	0.11888
Exponential	25.77485	-	-	0.52554

In our second experiment, we consider minimum K-S estimator with PH-transform technique for the same distributions as we used in the first experiment. The PH-transform technique is in the form

$$F_Y(t) = 1 - S_Y(t) = 1 - \left(S_X(t)\right)^c.$$

The K-S statistic of $F_Y(t;\theta)$ is given by

$$D(\theta) = \max_{1\le i\le N}\left|F_Y(t_i;\theta) - \frac{i}{N}\right|.$$

PH-transform technique algorithm is described as following:

Step 1: Compute the K-S statistic with PH-transform technique for the parameters $c = 1, a, b$ or d (a is scale parameter, b is location parameter and d is shape parameter) is obtained from Table 3.

Step 2: Randomly change the parameter value. We can do this by choosing constant r (real number) and let $\acute{c} = c + r$.

Step 3: Compute the K-S statistic value with a, b or d of Table 3 and $\acute{c}$.

Step 4: Compare the K-S statistic value which were obtained from step 1 and 3.

If the K-S statistic value of the step 3 is greater than step 1, then repeat step 2 with set $c = \acute{c}$, a, b or d of Table 3. Otherwise, we choose a new constant r then go on to step 2.

Step 5: Repeat step 2 to step 4 until the K-S statistic value remains constant.

The methodology of the minimum K-S estimator as shown in Figure 1.

In Table 4, we obtained that minimum K-S estimator with PH-transform technique in Table 4 is better than K-S statistic of distributions in Table 3, except logistic distribution and normal distribution.

Table 4. List of parameters by minimum K-S estimator with PH-transform technique for nine distributions

Distribution	parameters				K-S statistic value
	Scale	Location	Shape	c	
Rayleigh	36.05355	-	-	0.00245	0.12698
Logistic	36.99975	-	-	0.00014	0.44620
Gamma	11.35555	-	3.85453	0.00077	0.10392
Pareto	20.78183	-	1.24468	0.00008	0.08110
Log-logistic	0.37615	3.72054	-	0.00130	0.11193
Normal	20.14758	41.34690	-	0.00001	0.12797
Weibull	1.62774	-	57.75607	0.00046	0.14659
Log normal	0.68813	3.75845	-	0.00022	0.11887
Exponential	25.77485	-	-	0.72173	0.30005

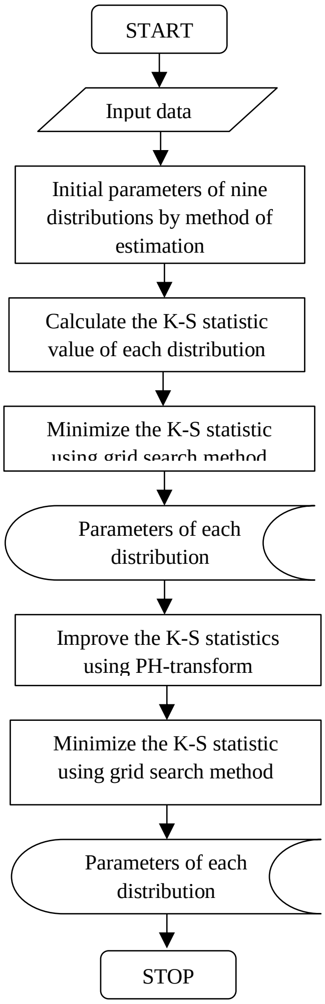
START
Input data
Initial parameters of nine distributions by method of estimation
Calculate the K-S statistic value of each distribution
Minimize the K-S statistic using grid search method
Parameters of each distribution
Improve the K-S statistics using PH-transform
Minimize the K-S statistic using grid search method
Parameters of each distribution
STOP

Figure 1. Methodology flow chart for minimum K-S estimator

4. Conclusions

In implementing the minimum K-S estimator, PH-transform technique can improve minimum K-S estimator with two operations of experiments. The algorithms give better the minimum K-S estimator

for seven distributions; Rayleigh distribution, logistic distribution, gamma distribution, Pareto distribution, log-logistic distribution, normal distribution, Weibull distribution, log-normal distribution, and exponential distribution while the minimum K-S estimators of normal distribution and logistic distribution are unchanged. For future research, minimum K-S estimator can be applied with other algorithm such as E-M algorithm, Markov Chain Monte Carlo (MCMC) algorithm, etc. Its applications are used in actuarial science instance, we can use our tool to estimate parameter vector of the claim severities for Weibull distribution as in (Khotama, S., et al., 2015), other biostatistics, and biomedical research.

5. Acknowledgement

This research was supported by Khon Kaen University, and Sakon Nakhon Rajabhat University, Thailand.

6. References

Khotama, S., Thongjunthug, T., Sangaroon, K., and Klongdee, W., (2015). On Approximating the Minimum Initial Capital of Fire Insurance with the Finite-time Ruin Probability using a Simulation Approach. KKU Research Journal, Vol.20, No.3, 267-271.

Leonard, A.A., and Mark, M.M., (2015). Probability and Statistics with applications (2nd ed.). America, MA: ACTEX Plublications, Inc.

Sinsomboonthong, S.(2015). Statistical Distributions (10th ed.). Bangkok, MA: *Chamchuree products c*ompany, Ltd.

Wang, S.(1995) Insurance pricing and increased limits ratemaking by proportional hazards transforms. Insurance: Mathematics and Economics, Vol. 17, 43-54.

Weber, M.D., Leemis, L.M., and Kincaid, R.K. (2006). Minimum Kolmogorov-Smirnov test statistic parameter estimates. Journal of *Statistical Computation and Simulation*, Vol.76, No.3, 195-206. doi: 10.1080/00949650412331321098.

Wieczorek, J. (2009). Finite Sample Properties of Minimum Kolmogorov-Smirnov Estimator and Maximum Likelihood Estimator for Right-Censored Data: Portland State University (Master Thesis). Available from ReseachGate.