Lognormal Distribution in Health Insurance: Interval Estimation Methods for the Average Cost and Correction of Truncated Values

Marli Amorim[2*], Joana Fernandes[1, 2], Teresa Alpuim[1]

[1] Faculty of Sciences at the University of Lisbon, Portugal, Lisbon.
[2] Multicare - Health Insurer, Portugal, Lisbon.

* Corresponding author. Tel.: 918 258 985; email: marli.amorim.ferreira@multicare.pt
Manuscript submitted March 1, 2018; accepted December 12, 2018.
doi: 10.17706/ijapm.2019.9.2.101-110

Abstract: Among the diseases with high treatment costs, oncologic and cardiovascular diseases are nowadays the main causes of death in Portugal as in many other high income per capita countries. This situation is a challenge for health insurance companies and financial institutions. Although there is currently some affordable health insurance in case of severe diseases, such as cardiovascular disorders or cancer, the available capital is easily depleted. The study of costs associated with many serious diseases shows that very often lognormal distribution fits well to the costs distribution. Therefore, a simulation study was made to compare different interval estimation methods of the average cost of a lognormal distribution. This work compares bootstrap parametric and non-parametric methodologies with Cox and large sample normal based methods and the results were applied to a breast cancer Portuguese dataset. Furthermore, for the lognormal distribution, a correction to the truncated values of costs following capital depletion is proposed, and the impact of this correction illustrated via its application to a heart failure dataset.

Key words: Bootstrap, confidence interval, costs distribution, heart and cancer diseases, lognormal distribution.

1. Introduction

In Portugal, as in many other countries, longevity is increasing and the proportion of young people decreasing. Under these circumstances, it is expected that the incidence of oncological and cardiovascular diseases will increase significantly in the near future. Simultaneously, recent developments in medical treatment have increased the rate of survival for patients suffering from these diseases. These two factors lead to high prevalence and inevitably to an increase in associated treatment costs. This situation concerns the funding institutions, as in the case of health insurance companies. Although there is currently some affordable health insurance for such serious diseases as cardiovascular or cancer, the available capital is often insufficient [1], [2]. This is a cause of concern for insurers who want adequate products to serve the clients' needs. This paper presents the results of research to study the costs associated with these diseases, but the long history of truncated values, because of capital depletion, was also a relevant problem. The results show that lognormal distribution fits well to the cost distribution for some of these serious diseases [3]. Assuming that costs follow this probability distribution, the objectives of this study are twofold: to compare some confidence interval estimation methods to the cost's mean value; and in case there exists

depletion in data to propose an estimation method of truncated values.

2. Methodologies

2.1. Interval Estimation Comparison

A simulation study was made to compare different interval estimation methods of the average cost when it follows a lognormal distribution. The methods under consideration were bootstrap parametric and non-parametric methodologies, namely, normal, studentized, basic and percentile [4]–[7], Cox [8], [9] and large sample normal based methods. Considering X as a random variable with lognormal distribution, $Y = ln(X)$ is normally distributed with mean μ and variance σ^2. Therefore, the mean and variance of X, m and v, are functions of μ and σ^2 given by:

$$m = e^{\mu+0.5\sigma^2} \text{ and } v = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

A brief description of the considered interval estimation methods for m follows.

- *Naïve*: *Naïve* confidence interval consists in the exponential transformation of normal based confidence interval limits, applied to data after logarithmic transformation. This approach produces the interval

$$\left(\exp\left(\bar{Y} \mp t_{1-\frac{\alpha}{2};n-1}\frac{S_Y}{\sqrt{n}}\right)\right),$$

where $t_{1-\frac{\alpha}{2};n-1}$ is the $1-\frac{\alpha}{2}$ quantile of a Student's t distribution with $n-1$ degrees of freedom.

- **Cox and Cox-t**: Cox confidence interval uses the maximum likelihood estimators for μ and σ^2 to estimate $\ln(m)$ and its variance. More specifically, the maximum likelihood estimator of $\ln(m)$ is given by $\bar{Y} + S_Y^2/2$, with variance

$$Var\left(\bar{Y} + \frac{S_Y^2}{2}\right) = \sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_Y^4}{2(n-1)}}.$$

As this estimator has asymptotic normal distribution, and replacing σ_Y^2 and σ_Y^4 by its consistent maximum likelihood estimators, S_Y^2 and S_Y^4, respectively, an asymptotic confidence interval for $\ln(m)$ may be easily constructed. Applying the exponential transformation to the upper and lower bound of that interval, the confidence interval for $\ln(m)$ is given by

$$\left(\exp\left(\bar{Y} + \frac{S_Y^2}{2} \mp q_{1-\frac{\alpha}{2}}\sqrt{\frac{S_Y^2}{n} + \frac{S_Y^4}{2(n-1)}}\right)\right),$$

where $q_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of a standard normal distribution. When the sample size is small, some authors recommend to use the quantile of a Student's t distribution with $n-1$ degrees of freedom (Cox-t method).

- **Large**: Large sample confidence interval consists in the usual large sample normal based interval, that is,

$$\left(\bar{X} \mp z_{1-\frac{\alpha}{2}}\frac{S_x}{\sqrt{n}}\right),$$

with $z_{1-\frac{\alpha}{2}}$ the $1-\frac{\alpha}{2}$ quantile of a standard normal distribution.

Several bootstrap methodologies were also considered, namely: Parametric methods (bootP), in which samples from a pre-defined distribution are generated, with parameters previously estimated from the initial sample; and non-Parametric methods (bootNP), in which the sample is constructed using a resampling process of the initial sample.

Therefore, the bootstrap confidence intervals in study were:

- **boot(P/NP)Normal:** The normal bootstrap confidence interval is based in the large sample interval, but has a bias correction associated. The estimates for the bias and standard deviation are calculated from the bootstrap samples. This confidence interval is given by:

$$\left(\bar{X} - b_B \mp z_{1-\frac{\alpha}{2}}\widehat{se}_B\right),$$

with B the number of bootstrap samples and $z_{1-\frac{\alpha}{2}}$ the $1-\frac{\alpha}{2}$ normal quantile. Also, b_B represents the bias of bootstrap estimates to sample estimates and $\widehat{se}_B$ is the bootstrap standard deviation and they are given, respectively, by:

$$b_B = \frac{\sum_{b=1}^{B}\left(\bar{X}^{(b)} - \bar{X}\right)}{B}$$

And

$$\widehat{\text{se}}_B = \sqrt{\frac{\sum_{b=1}^{B}\left(\bar{X}^{(b)} - \bar{X}^{(.)}\right)^2}{B-1}},$$

where $\bar{X}^{(.)} = \frac{\sum_{b=1}^{B}\bar{X}^{(b)}}{B}$.

- **boot(P/NP)Studentized:** The studentized bootstrap confidence interval is based in the large sample interval as well, but the pivotal quantity $Z = (\bar{X} - m)/se_{\bar{X}}$ is replaced by $Z^* = \frac{\left(\bar{X}^{(b)} - \bar{X}\right)}{se^{(b)}}, b = 1, .., B$. This confidence interval is given by

$$\left(\bar{X} - z^*_{1-\frac{\alpha}{2}}\widehat{se}_B, \bar{X} - z^*_{\frac{\alpha}{2}}\widehat{se}_B\right),$$

with B the number of bootstrap samples, $\widehat{se}_B$ the bootstrap standard deviation, and z^*_τ the τ quantile of bootstrap distribution, Z^*.

- **boot(P/NP)Basic:** The Basic bootstrap method is based in the fact that

$$P\left(q_{\frac{\alpha}{2}} \leq \bar{X}^{(b)} \leq q_{1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

and, consequently,

$$P\left(q_{\frac{\alpha}{2}} - \bar{X} \leq \bar{X}^{(b)} - \bar{X} \leq q_{1-\frac{\alpha}{2}} - \bar{X}\right) = 1 - \alpha.$$

Assuming that $\left(\bar{X}^{(b)} - \bar{X}\right)$ is a good approximation for $(\bar{X} - m)$, the confidence interval can be written

as:

$$\left(2\bar{X} - q^*_{1-\frac{\alpha}{2}}, 2\bar{X} - q^*_{\frac{\alpha}{2}}\right),$$

with B the number of bootstrap samples and q^*_τ the τ quantile of bootstrap distribution of $\bar{X}$.

- **boot(P/NP)Percentile:** In contrast with the previous methods, the percentile bootstrap interval is not based in pivotal quantities. This interval is constructed under the assumption that $\bar{X}^{(b)} \sim N(\bar{X}, \widehat{se}^2)$ and the confidence interval $\bar{X} \mp \widehat{se} z_{1-\frac{\alpha}{2}}$ is approximated by:

$$\left(q^*_{B+1;\frac{\alpha}{2}}, q^*_{B+1;1-\frac{\alpha}{2}}\right)$$

with B being the number of bootstrap samples and q^*_τ the τ quantile of bootstrap distribution of $\bar{X}$.

For different values of the parameters μ (mean value of the normal distribution associated) and σ (standard deviation of the normal distribution associated) and the sample size *n*, a set of 1 000 sample replicates was generated in order to understand what are the most adequate methods of interval estimation. To evaluate the performance of each method for each of these scenarios, 1 000 balanced confidence intervals at the 0,95 confidence level were constructed, using all the methods in study. The methods producing a coverage level, that is, a proportion of intervals including the true value of *m* smaller than 0,8, were eliminated. Then for each of the other methods with a good coverage level (greater than 0,8), the average interval amplitude was computed and compared with the coverage level. That is, the precision of each method was measured through the ratio [10]

$$\frac{Average\ amplitude}{Coverage\ level}$$

and the methods producing smaller ratios were selected. This indicator was used in order to exclude the methods that produced high coverage levels but amplitudes excessively large.

2.1. Truncated Costs Correction

As mentioned before, many elements in the available samples of costs are truncated due to the depletion of capital. Thus, these values may be treated as observations from a lognormal population truncated at a constant c. The estimation of the true values of these observations can be made through the use of the conditional expected value.

The conditional expected moments of a lognormal distribution can be evaluated from the incomplete higher order moments [11], which are given by:

$$\int_0^c x^k f_X(x)\,dx = E(X^k)\Phi\left(\frac{\ln(c) - \mu - k\sigma^2}{\sigma}\right), \tag{1}$$

where $f_X(x)$ is the density function of the random variable X (defined in section 2.1).

From here, it is easy to see that the k^{th} order moment of the truncated lognormal random variable is

$$E(X^k|X \geq c) = \frac{1}{P(X \geq c)}\int_0^{+\infty} x^k f_X(x)\,dx = \frac{1 - \Phi\left(\frac{\ln(c) - \mu - k\sigma^2}{\sigma}\right)}{1 - \Phi\left(\frac{\ln(c) - \mu}{\sigma}\right)}E(X^k). \tag{2}$$

Considering $k = 1$, the conditional expected value is given by:

$$E(X|X \geq c) = \frac{e^{\mu+\frac{\sigma^2}{2}}}{1-\Phi\left(\frac{\ln(c)-\mu}{\sigma}\right)}\left[1-\Phi\left(\frac{\ln(c)-(\mu+\sigma^2)}{\sigma}\right)\right]. \quad (3)$$

The truncated observations in the sample were replaced by an estimated value of its true value given by the conditional expected value calculated with the help of formula 3 and where μ and σ^2 are replaced by its maximum likelihood estimates. The prediction mean square error of these estimates is exactly the conditional variance of the truncated observations. Taking $k = 2$ in formula 2 the conditional second moment is obtained and, consequently, the conditional variance is

$$Var(X|X \geq c) = \frac{\Phi\left(-\frac{\ln(c)-(\mu+2\sigma^2)}{\sigma}\right)}{\Phi\left(-\frac{\ln(c)-\mu}{\sigma}\right)}E(X^2) - [E(X|X \geq c)]^2, \quad (4)$$

where

$$E(X^2) = e^{2\mu+2\sigma^2}, \quad (5)$$

with c is the positive truncation point.

3. The Data

To illustrate the application of the several interval estimation methodologies, a sample of 513 breast cancer patients was used. From these, only 4 were men, and all of them were diagnosed between the years 2005 and 2012, with ages ranging from 27 to 87 years old.

The correction of truncated costs with the use of the conditional expected value of a lognormal distribution was applied to the inpatient costs, in the year of diagnosis, of 558 heart failure patients, of which 72 had unlimited capital and 16 exceeded contracted capital. Both sets of data were kindly provided by the health insurance company Multicare.

The treatment costs of various oncological and cardiovascular diseases fit well to the lognormal distribution. Shapiro-Wilks and Lilliefors adjustment tests were applied to the hypothesis of lognormal distribution for the treatment cost, and they lead to non-rejection for at least some of the usual levels of significance (breast cancer: p-values 0,013 and 0,115; heart failure: p-values of 0,411 and 0,389). These results agree with the QQ-plot presented in Fig. 1b) and 2b.

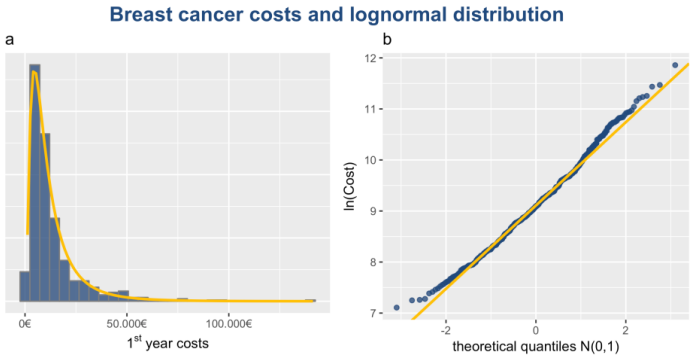

Fig. 1. a) Histogram of treatment cost and density curve of lognormal distribution. b) QQ-plot of the logarithm of cost.

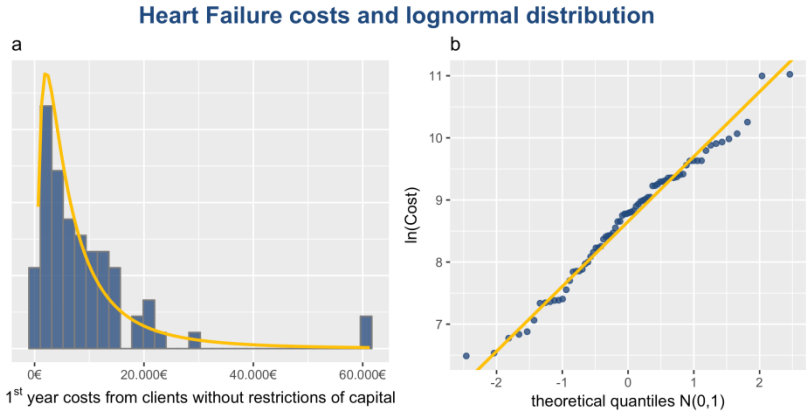

Fig. 2. a) Histogram of treatment cost and density curve of lognormal distribution. b) QQ-plot of the logarithm of cost.

4. Results

4.1. Interval Estimation Comparison

The simulation study concluded that the best methodologies for the interval estimation of the mean of a lognormal distribution are the non-parametric percentile bootstrap, if the standard deviation is small, and the parametric percentile bootstrap, when the standard deviation is large. For large samples, namely larger than 50, the best method is the Cox confidence interval. These results are summarized in Fig. 3.

		σ						
		0,5	0,75	1	1,25	1,5	1,75	2
n	5	bootPPercentile	bootPPercentile	bootPPercentile	bootPPercentile	bootPPercentile	bootPPercentile	bootPPercentile
	15	bootNPPercentile	bootNPPercentile	bootNPPercentile	bootPPercentile	bootPPercentile	bootPPercentile	bootPPercentile
	30	bootNPPercentile	bootNPPercentile	bootNPPercentile	bootNPPercentile	bootPPercentile	bootPPercentile	bootPPercentile
	50	bootNPPercentile	Cox	Cox	Cox	Cox	Cox	Cox
	100	bootNPPercentile	Cox	Cox	Cox	Cox	Cox	Cox
	250	bootNPPercentile	Cox	Cox	Cox	Cox	Cox	Cox
	500	bootNPPercentile	Cox	Cox	Cox	Cox	Cox	Cox

Fig. 3. Summary of the bilateral confidence intervals chosen for estimating the mean of a lognormal variable.

When the standard deviation is small most of the methods produce good results. However, when the samples are small, or the standard deviation is large, parametric methods produce better results than exclusively sample based methods. For large samples Cox methods are the best and do not differ significantly from each other.

Some expected but interesting points worth noting are that, in contrast with other methods, Naïve method does not improve its performance with large samples. Rather this method is worse the larger the sample size or the standard deviation. The reason is that the Naïve method is a confidence interval for e^{μ} and not for $m = e^{\mu+\frac{\sigma^2}{2}}$. As for Large confidence interval, it is interesting that although it is rarely the best option, its results are usually not very different than the results of the selected intervals.

Application to Dataset of first year breast cancer costs:

As referred in the data description section, the lognormal distribution fits well to the costs of most types of cancer. When considering only one type of cancer, however, it is possible to identify different patterns of costs, associated with the severity of the disease. The medical community uses frequently the word "stage" to refer to the different levels of the disease progression. This variable has a great impact on costs since an advanced stage is usually associated with a more aggressive treatment and therefore more expensive.

Table 1 presents the results for logarithms of costs for breast cancer, dividing the patients into subsamples according the stage of the disease. The table shows the estimated standard deviation and the interval estimation method which produced the best result.

Table 1. Dimension and Logarithms of Costs Standard Deviation of Each Stage Sample

	Stage				
	In Situ	**I**	**II**	**III**	**IV**
σ	0,62	0,77	0,72	0,72	0,96
n	44	205	162	81	21
CI method	bootNPPercentile	Cox	Cox	Cox	bootNPPercentile

Using stage information to split data results into more homogeneous groups, the adjustment tests show improvements in most cases.

Fig. 4 shows the results for the confidence intervals methods selected for each stage.

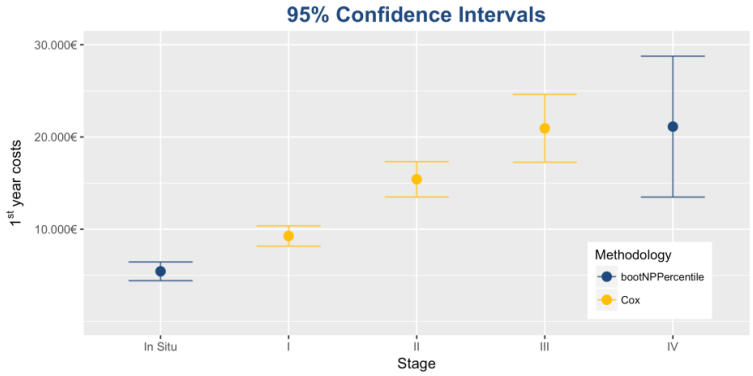

Fig. 4. 95% Confidence interval for average cost of treatment of breast cancer in first year after diagnosis, per stage.

An increase in the average costs is usual until stage III, but not always in stage IV. In fact, in case of advanced disease, many patients died before finishing the treatment and some chose not even to start the treatment. For breast cancer first-year costs, the calculated confidence intervals follow this expected behavior. For stage IV, the confidence interval has a larger amplitude because of smaller sample dimension $(n = 21)$ and larger standard deviation $(\sigma = 1)$.

4.2. Truncated Cost Correction

The contracted capitals by coverage may be one of two types: limited or unlimited. When capitals are limited, insurance companies pay, per year, claims whose total amount does not exceeds the contracted capital so that a surplus has to be borne by the insured person. If the capitals are unlimited, companies cover any amount per year of contract. In the case of limited capital, whenever the value of the claims exceeds the contracted capital, the insurer will have access only to right truncated data, since the individual likely will continue his treatment, despite having to pay the excess costs without the insurer's co-participation.

It should be noted that the consequence of using truncated values in the calculation of the average cost of a given diagnosis is that such calculation will be underestimated, since some of these observations are less than the actual costs. Similarly, the removal of these observations from the sample results in an even more underestimated average cost of the diagnosis.

Therefore, estimating the actual cost of patients who exceeded the capital is an extremely important

matter for adequate pricing in health insurance [12]-[14]. Once the company sells policies with unlimited capital, these were used to estimate the truncated amounts of the limited capital products.

Assuming that the costs distribution of the unlimited capital does not differ significantly from that of the limited capital, it is possible to estimate the parameters of the lognormal distribution, necessary for the calculation of the conditional expected value, based on the unlimited sample, in order to calculate an estimate for the truncated costs in the sample of the limited capitals. [15]

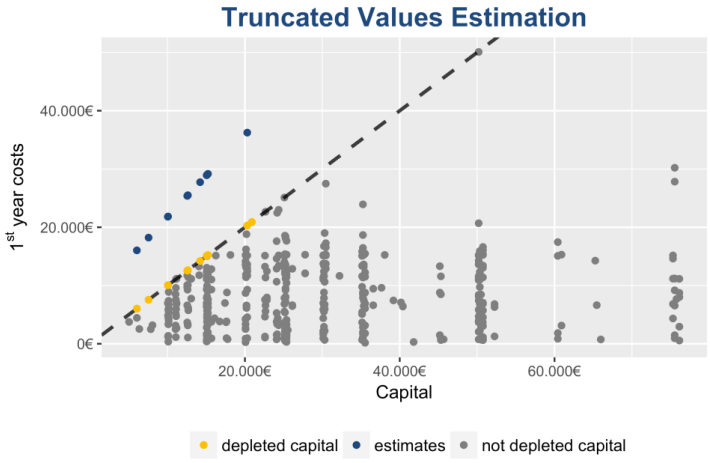

Fig. 5. Correction of truncated costs by depletion of capital, using the conditional expected value from a lognormal distributed variable.

Fig. 5 illustrates the result of the above-described estimation procedure, applied to the inpatient observations occurred in the year of diagnosis in clients with heart failure. Three types of observations may be observed: customer costs that did not reach the contracted capital (gray points); costs of customers whose capital was insufficient (yellow points); and estimates of real costs (blue dots).

Note that the estimation process was based on formula 3, where X is a random variable that represents the cost which has lognormal distribution, with mean m and variance v, and c represents the contracted capital.

Correcting the truncated values before calculating the average costs results in an increase of 276,24€ in the final result. These numbers illustrate the underestimation problem caused by the use of truncated observations. Also, using solely the values associated with unlimited capital tends to overestimate the mean value, in this case, a difference of 518,72€. The precision of the estimate of the mean value is also underestimated in the case of non-corrected observations, once the capital limits force the observations to be artificially small. After correction, the estimated standard-deviation of the average of costs increases from 295,41€ to 461,53€. Similarly, the use of the sample including only unlimited capitals leads to a poor estimate, because the sample size is too small (1 252,05€).

5. Conclusion

From this study it was possible to select the methodologies producing more accurate confidence intervals for the mean of a lognormal distribution, according to the population variance and the sample size. These confidence intervals are necessary to evaluate the tariff of health insurance products and simultaneously may be applied to identify potential abusive or fraudulent usage from providers. As for many insurance contracts, the coverage capitals are limited, thus it is important to provide corrections to the truncated values observed in these cases. The proposed correction showed a good performance when

applied to the data set used for the study. This correction is of particular importance when the goal is the pricing of products with higher capital than the usual.

Acknowledgment

The authors would like to thank FCT - Foundation for Science and Technology (Portugal) and health insurer Multicare for funding the Industrial PhD scholarships SFRH/BDE/51974/2012 and SFRH/BDE/52123/2013, which allowed the development of this work. The authors also want to express their gratitude to CMAF CIO - Center for Mathematics, Fundamental Applications and Operations Research for partial funding and support of this research work.

References

[1] Brooks, E., *et al*. (2010). Health insurance and cardiovascular disease risk factors. *The American Journal of Medicine*, *123(8)*, 741-747.

[2] Calcagno, J., *et al*. (2006). Cardiovascular disease and health care system impact on functionality and productivity in Argentina: A secondary analysis. *Value in Health Regional Issues,* 11, 35-41.

[3] Zuanetti, D., Diniz C., & Leite, J. (2006) A lognormal model for insurance claims data. *REVSTAT-Statistical Journal*, *4(2)*, 131-142.

[4] Bradley, E., & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* London: Chapman & Hall.

[5] Chernick, M., & LaBudde, R. (2014). *An Introduction to Bootstrap Methods with Applications to R*. New Jersey: John Wiley & Sons.

[6] Davison, A., & Hinkley, D. (1997) *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

[7] Olsson, U. (2005). Confidence intervals for the mean of a lognormal distribution. *Journal of Statistics Education, 13(1).*

[8] Chami, P., *et al*. (2007). On efficient confidence intervals for the lognormal mean. *Journal of Applied Sciences*, *7*, 1790–1794.

[9] Zhou, X., & Gao, S. (1997). Confidence intervals for lognormal mean. *Statistics in Medicine, 16*, 783-790.

[10] Ferreira, M. (2017). Avaliação do Risco Oncológico- Avaliação da Viabilidade de um Seguro/Cobertura Específica. Doctoral dissertation, University of Lisbon, Portugal.

[11] Aitchison, J., & Brown J. (1963). *The Lognormal Distribution with Special Reference to Its Uses in Economics.* Cambridge: Cambridge University Press.

[12] Kaas, R., *et al*. (2008). *Modern Actuarial Risk Theory - Using R.* Heidelberg: Springer.

[13] Pitacco, E. (2014). *Health Insurance. Basic Actuarial Models*. EAA Series. Springer.

[14] Vanduffel, S., *et al*. (2008). Optimal approximations for risk measures of sums of lognormals based on conditional expectations. *Journal of Computational and Applied Mathematics, 221*, 202-218.

[15] Ginos, B. F. (2009). Parameter estimation for the lognormal distribution. Master dissertation, Brigham Young University, Provo.

Marli Amorim was born in Portugal, in 1988. She received the master in statistics in Faculty of Sciences, Lisbon University in 2011 and is nowadays finishing her PhD program in the same institution. Her work have been in the context of an industrial PhD scholarship with the national Health Insurer Multicare, within the topic ''the actuarial risk of oncologic disease for health insurers''. Until 2016 she worked as invited assistant teacher in Lisbon University. Nowadays she is working, as actuary, in the health insurer Multicare

Joana Fernandes took a first degree in applied mathematics, in 2009, and a master in statistics, in 2012, both at the University of Lisbon. Nowadays she is finishing her PhD program in the same institution. She received an Industrial PhD scholarship, co-funded by health insurer Multicare and FCT, to develop research within the subject ``Cardiovascular risk assessment in health insurance business''. Until 2016 she worked as teaching assistant at University of Lisbon. Nowadays she is working, as actuary, in the health insurer Multicare.

Teresa Alpuim took a first degree in mathematics, in 1981, and a master in statistics and operations research, in 1985, both at the University of Lisbon. She did her Ph.D. in 1989, in probability and statistics, in the same University. She is a full professor at University of Lisbon and has research work in several areas of statistics, namely, statistics of extreme values, time series, spatial statistics, Kalman filter, linear models, and its applications, mainly, to environmental sciences and insurance and actuarial sciences.