Journal of Quantitative Economics, Vol. 8 No. (2), July 2010

# LENGTH OF STAY – A DATA ANALYTIC APPROACH

**D. V. S. SASTRY**[1]
**R. K. SINHA**[2]

**Abstract**

*The length of stay (LOS) at a hospital of a patient contains important information in health statistics irrespective of the fact whether the patient is insured or not. A longer stay may result in resource restrictions with regard to the availability of beds at hospitals. Some of the determinants of length of stay are the demographic characteristics and hospital characteristics. This paper attempts to model the distribution of and try to explain the variation in LOS through age, gender and type of disease for the Indian data. The data set used is the insured data submitted by third party administrators.*

**Keywords**: Health statistics, ICD codes, length of stay, standard deviation, Skewness.

**JEL Classifications**: C10, C21, I11

## Introduction

Length of stay at a hospital (the difference between the date of discharge and date of admission into a hospital) has some important features as a measure. It is simple to calculate and easy to understand by everyone. This simple measure has implications for different stakeholders in the ecosystem of medical care. For a patient, the more the stay in a hospital, the more expensive it could be. The stay in a hospital not only covers the hospitalization charges, medicines and other associated costs like investigations etc., but also room and nursing charges. If the patient is covered by an insurance scheme, the expenditure is borne by the insurer (government or private insurance).

For a hospital, shorter stays imply release of capacity in terms of beds so that more patients can be treated. Though longer stays will increase revenues, hospitals try to discharge or move the patients quickly to other hospitals so as to release capacity. Hospitals need to balance these diversified objectives while laying down their management policies.

An insurer, under the terms of contract with the policyholder has to bear the hospital expenses, which are losses to them. As such, the distribution of claim amounts paid to the policyholders will guide the insurer in fixing the future premiums. The premium should cover the

[1] Director General, Research & Development Department, Insurance Regulatory & Development Authority, Hyderabad, 3rd Floor, Parishram Bhawan, Hyderabad – 500 004. E-mail: dvssastry@hotmail.com

[2] Deputy Director, Research & Development Department, Insurance Regulatory & Development Authority, Hyderabad, 3rd Floor, Parishram Bhawan, Hyderabad – 500 004. E-mail: rksinha@irda.gov.in

The views expressed in this paper are of the authors and do not necessarily represent the organization they belong to.

expected loss and other fixed costs associated with the insurance business. In this connection, the probability distribution of length of stay becomes important.

Some diseases require longer treatment and thereby longer stay at the hospital. As such, the distribution of LOS exhibits long right tail. Observing this, many authors pointed out that the distribution of LOS is right skewed. [Atienza (2005), Lim and Tongkumchum (2009) etc.]. Because of this property, lognormal, Weibull and gamma distributions are choices for LOS data. Empirical work in this regard suggested that Weibull and lognormal distributions are appropriate [Marazzi et al 1998]. It may be recalled that the choice of a distribution from a set of competing distributions depends on statistical tests like Anderson – Darling, $x^2$ and Kolmogorov – Smirnov etc. Besides, prior knowledge of the variable and the data set also help in the choice of a distribution. For example, some authors have deleted LOS of less than one day, which as recorded zero leads to computational problems. This helped them in obtaining familiar distribution forms for LOS data.

Hellervik and Rodgers (2006) argued that competition for resources between different groups of patients is a significant factor affecting the distribution of the LOS. Competition can take place between hospitals, between departments and between different types of patients within the same department. For English hospitals data, the distribution of LOS was well described by Power Law. MacLean and Richman (2001) argued that resources can be excessive for a small subset of users and at the same time could be minimal or even none for the others.

Atienza (2005) proposed a mixture of gamma and lognormal distributions instead of using either of them so as to get better results. Lim and Tongkumchum (2009) handled the skewness in LOS for patients who died in the hospitals by two methods: (a) logistic regression with LOS of 7 days or more taken as the outcome and (b) linear regression on natural logarithms of LOS after adding an appropriate constant (0.5) to cope with LOS equal to zero.

It is well known that the measure of skewness is very sensitive to the presence of outliers in the data. An outlier in the right tail can unduly increase the skewness co-efficient making it difficult to interpret [Brys et al (2003)]. Unusual long stay in a hospital that varies in length as well as over time hampers the statistical analysis of LOS. Therefore, fitting a distribution based on the descriptive statistics and drawing conclusions may sometimes lead to improper identification of a distribution. An appropriate probability distribution helps the actuary in fine tuning the premium calculations and also helps to estimate the survival probabilities, which in LOS context refers to further stay in a hospital.

The second important use of LOS data is to ascertain the factors that explain variations in LOS. Martin and Smith (1996) showed that demographic characteristics of patients and some hospital characteristics are two important determinants of LOS. The important demographic characteristics are age, gender, type of disease etc. Hospital size, location of the region, the type of hospital etc. are some of the hospital characteristics which have impact on LOS.

Studies on LOS have used either hospital data available for case studies or data released by the governments. For example, NHS releases data from hospital records, and in US, the data are from Medicaid etc. or from hospitals. These data also records the reasons for discharge from the hospital, like death, shift to another hospital etc. However, studies on LOS are not available based on insurance records. This study focuses on LOS as available from Indian insurance claim records. The paper is organized as follows: The Section 1 gives a detailed account of data used, and Section 2 deals with exploratory analysis and summary statistics.

Section 3 deals with the differences in LOS due to gender, age etc. Section 4 tries to estimate a regression function for explaining the variation in LOS. Conclusions and future work is presented in section 5.

## 1. Data Description

The Insurance Regulatory and Development Authority (IRDA) regulates insurance companies operating in India. Insurance business in India was recently opened up to private participation including joint ventures. Lot of importance has been attributed to the collection of statistics on insurance business across the companies, which was hitherto not available for analytical studies. In this background, IRDA *inter alia* collects transaction level data from the insurance companies and third party administrators (TPAs) on the health insurance portfolio. The transaction level data are aggregated at different levels. The latest summary statistics for 2008-09 as aggregated from around 2.08 million records were hosted on www.irdaindia.org in May 2010. IRDA has also hosted transaction level data of 1,00,000 records for 2008-09. These transaction level data relate to details available from claim records; however, premium details are not available.

Each record contains many fields, including demographic details of the insured, type of disease for which a claim was made etc. It becomes difficult to physically verify each and every field for its accuracy as well as for transcription errors. Besides, quality of data (i.e. error free data) varies across insurers. Despite cleansing the data for errors, many discrepancies still crept in to the database. At the aggregate level analysis, the impact of such errors could probably be less; but at the micro level the conclusions may be distorted.

The data covers different hospitals, disease categories, gender, age of the claimants etc. Illnesses have been grouped using ICD 10 codes into 17 broad categories. Sufficient care was taken to verify whether the dependent and independent variables are error-free.

For the present study, records which have errors either in any one of the above variables namely date, age, gender and proper disease code have been excluded. Further, few records, which lacked reliability or precision (for example, perinatal cases of other than infants), have been deleted. Records with LOS of more than 60 days have not been considered. Thus, from a file of 100000 records, 35595 records were selected, which met the above requirements.

Length of stay is a continuous variable; however, it is recorded as a discrete variable in integers. Length of stay of '0' day means either an out patient treatment or does not require hospitalization. Some treatments may not require hospitalization; for example, cataract surgery etc. As observed in many other studies, the Indian data also showed large number of observations with LOS equal to '0' day. In the data set used for the present study, around 15 per cent of observations are with '0' day and another 22 per cent showed, LOS of 1 day. The number of observations for LOS of 2 days was around 18 per cent. *Prima facie* it can be expected that the first and second quartile will be around 1.5 and 2.5. As such, the histogram will show peaks at zero to two days and slowly tapers showing a large tail as there are 90 patients whose stay was for more than sixty days. As observed earlier, this type of long tail distorts the conclusions. As such, the data needs to be truncated for obtaining robust estimates. We consider two data sets: one with all data points and another truncated data set. For trimming the data, we followed the methodology of Kulinskaya et al (2005). The trimming criteria adopted was 1.5 inter quartile range from the upper quartile. In order to make the discrete variable continuous, following Lim

and Tongkumchum (2009), every data point is added with 0.5. This will also take care of the log transformation for LOS of '0' days.

## 2. Data Analysis

The descriptive statistics of LOS is given in Table 1.

**Table 1. Descriptive Statistics of LOS**

| | |
|---|---|
| Sample Size | 35595 |
| Mean | 3.8574 |
| Standard Deviation | 4.8892 |
| $Q_1$ | 1.5000 |
| Median ($Q_2$) | 2.5000 |
| $Q_3$ | 4.5000 |
| Skewness | 5.2359 |
| Kurtosis | 40.947 |

It shows that the LOS data of Indian hospitals is right skewed and leptokurtic. The Inter-quartile range and the outliers are shown in Box plot (Fig 1). The presence of large number of extreme values in the LOS data set can be seen from the plot.

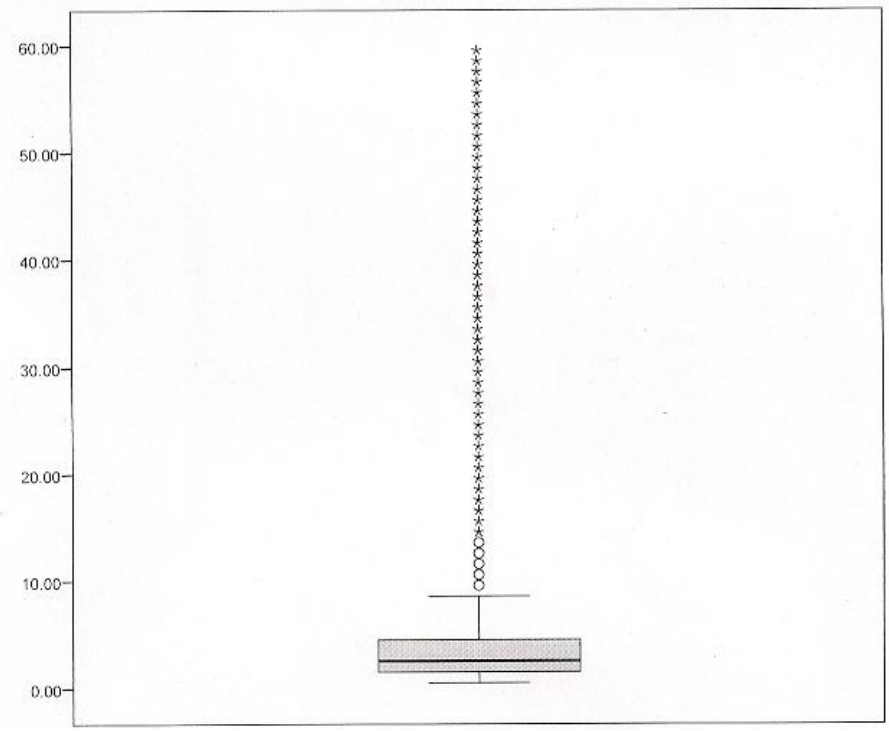


**Figure 1. Box Plot of LOS**

The suitability of an appropriate model can *prima facie* be known from the histogram of the data.

The histogram of LOS for the entire data set is presented in Fig 2a.

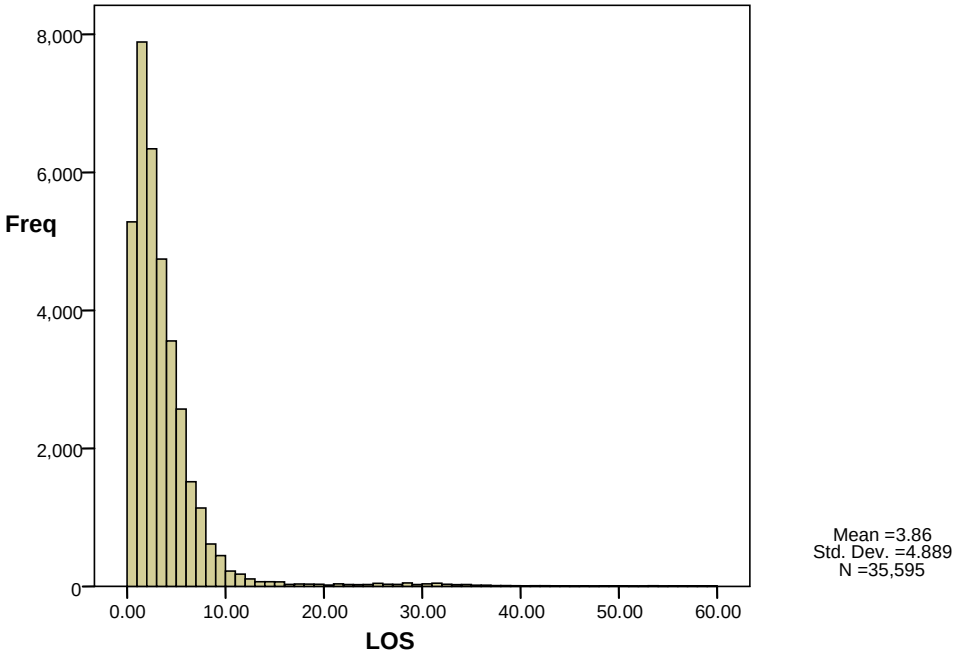


**Figure 2: Histogram of LOS for entire range**

As the long tail beyond 10 days is not clearly visible, the histogram for less than 10 days and more than 10 days are presented separately in Fig 3 and 4 respectively. The behaviour of the right tail of LOS can be seen in Fig 4.

The histogram of a quantity with Power Law distribution appears as a straight line when plotted on a logarithmic scale. Further, in order to see whether LOS behaves as Power Law, the graph in logarithmic scale is presented in Fig 5, from which the nature of right tail can be clearly seen. The large fluctuations in the right tail could be due to the fact that the number of sample points in these bins is smaller compared to the total sample size. The fat tail therefore points out that power law does not hold good for our data set.

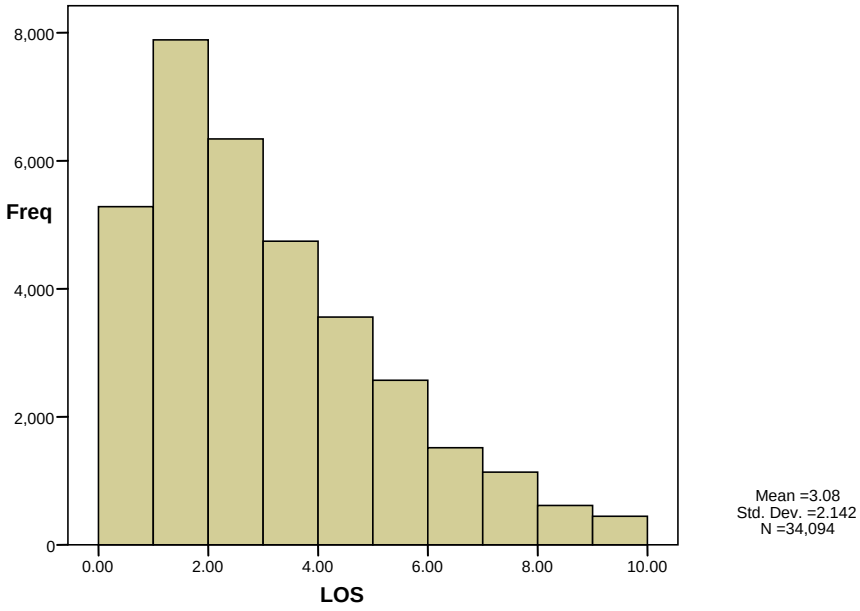


**Figure 3. Histogram of LOS for upto 10 days**

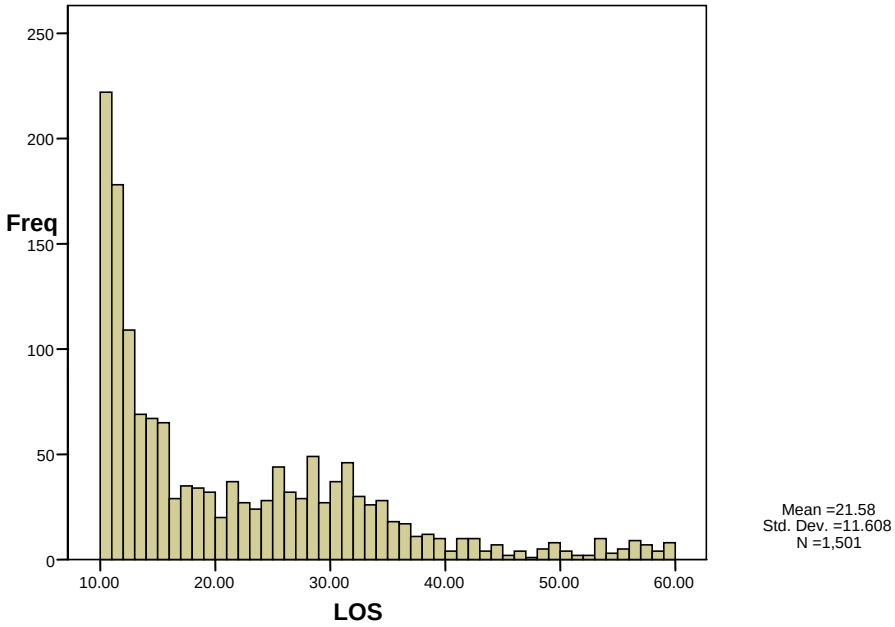


**Figure 4. Histogram of LOS for more than 10 days**

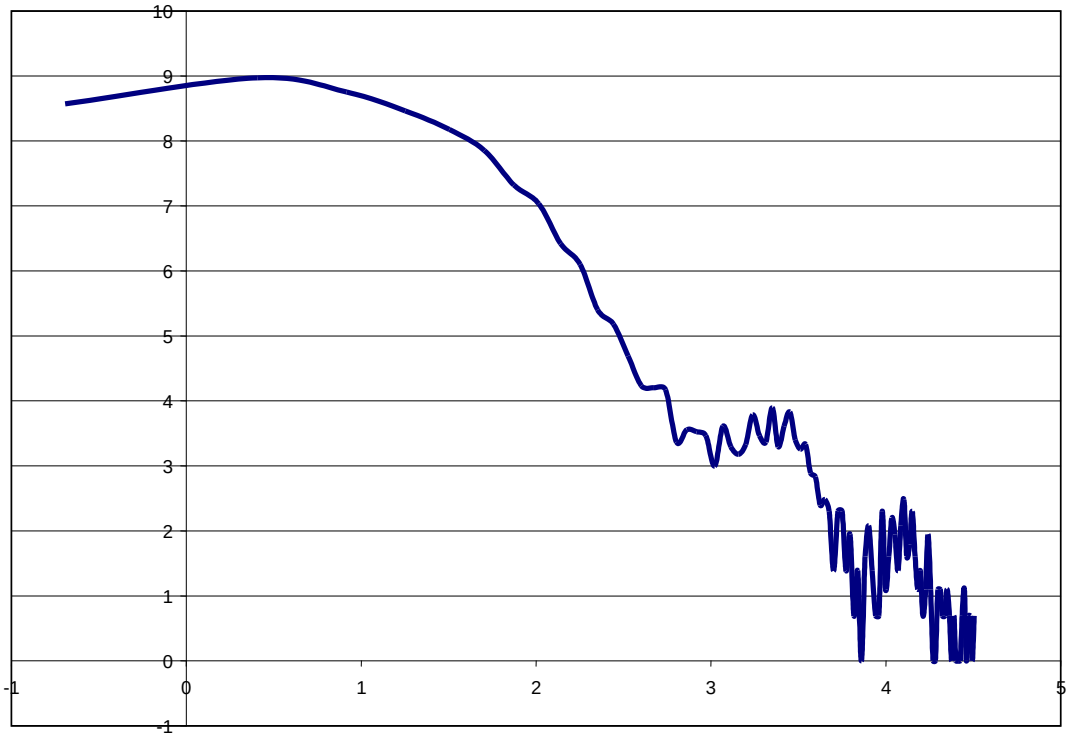


**Figure 5. Plot of relative frequency of LOS on Logarithmic scale**

The appropriateness of lognormal, gamma and weibull distributions for length of stay have been shown by many data sets across hospitals and time periods. This apriori information has been used to the current data set also. EasyFit software fits many distributions and ranks the distributions according to test statistics like Kolmogorov - Smirnov, Anderson – Darling and $x^2$ test. The appropriateness of the distribution was also judged through Q-Q plot, with a preference towards suitability at around main body of the data. Based on the above criterion, the 2-parameter lognormal distribution was chosen which has a density:

$$f(x) = (1/x\sigma\sqrt{2\pi})Exp[-1/2(\ln x - \mu/\sigma)^2 \qquad 0 \le x < \infty$$

The distribution function is of the form

$$f(x) = \Phi(\ln x - \mu/\sigma)$$

where, $\Phi$ is the Laplace Integral

The estimated parameters are $\mu$ = 0.92394 and $\sigma$ = 0.91972. The descriptive statistics of the model vis-à-vis the observed data set are given below:

**Table 2. Comparison of Model Statistics with Actual**

| *Statistics* | *Lognormal* | *Observed* |
|---|---|---|
| Mean | 3.8454 | 3.8574 |
| SD | 4.4348 | 4.8892 |
| Skewness | 4.9937 | 5.2359 |
| Kurtosis | 65.062 | 40.947 |
| First Quartile | 1.3547 | 1.5000 |
| Second Quartile | 2.5192 | 2.5000 |
| Third Quartile | 4.6846 | 4.5000 |

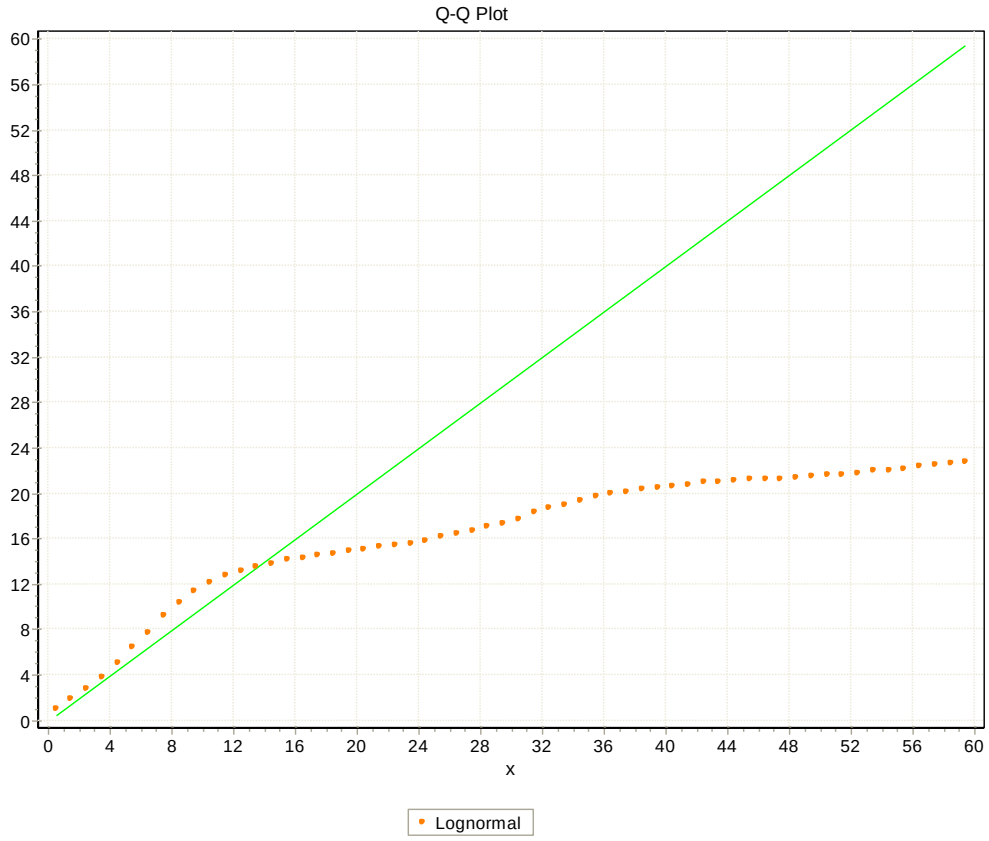


**Figure 6. Q-Q plot of complete data set**

From the Q-Q plot, linearity may be observed upto LOS of 10 days, confirming earlier trimming criteria. In fact, $Q_3 + 1.5*(Q_3 - Q_1)$ of the dataset works out to 9.5 days. It covers 96 per cent of the data upto the trimming point.

However, the long tail cannot be ignored for practical reasons, as these events can happen with small probability but the financial affect of these are heavy for insurance applications. These extreme values need to be properly modeled. The classical extreme value theory (also known as Block Maxima Approach) is based on three asymptotic extreme value distributions identified by Fisher and Tippett (1928). Jenkinson (1955) combined these three distributions into a single mathematical form with cumulative distribution function (CDF).

$$F(X: k,\sigma,\mu)) = e^{-(1+kz)-1/k}, \qquad k \neq 0$$

$$= e^{-e^{-z}}, \qquad k = 0$$

Where, $z = x-\mu/\sigma$ and $k$, $\sigma$ and $\mu$ are shape, scale and location parameters respectively. 'X' is the maximum of an *epoch*. While $\sigma > 0$, $k$ and $\mu$ can take any real value. The range of the GEV distribution depends on k: given by

$$1+k\frac{(x-\mu)}{\sigma} > 0 \qquad \text{for } k \neq 0$$

$$-\infty < x < +\infty \qquad \text{for } k = 0$$

The probability density function of GEV is:

$$f(x) = \begin{cases} \frac{1}{\sigma}\exp\ -(1+kz)^{-1/k}\ (1+kz)^{-1-1/k} & k \neq 0 \\ \frac{1}{\sigma}\exp(-z-\exp(-z)) & k = 0 \end{cases}$$

Here, 3 cases arise as below:

**Case I:** When k = 0, it becomes the Type I GEV, known as Gumbel distribution. The Gumbel distribution is unbounded (defined on the entire real axis) with probability density function:

$$f(x) = \frac{1}{\sigma}\exp(-z-\exp(-z))$$

Where $z=(x-\mu)/\sigma$, $\mu$ is the location parameter, and $\sigma$ is the distribution scale ($\sigma>0$). The shape of the Gumbel model does not depend on the distribution parameters.

**Case II:** When k > 0, it becomes the Type II GEV, known as Frechet distribution with probability density function:

$$f(x) = \frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1}\exp\left(-\left(\frac{\beta}{x}\right)^{\alpha}\right)$$

Where $\alpha$ is the shape parameter ($\alpha>0$), and $\beta$ is the scale parameter ($\beta>0$). This distribution is bounded on the lower side ($x>0$) and has a heavy upper tail.

**Case III:** When $k < 0$, it becomes the Type III GEV, known as Weibull distribution. The two-parameter Weibull distribution has the probability density function:

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$$

The Weibull distribution is defined for x>0 and both parameters shape (α) and scale (β) are positive. Thus, the block maxima approach of extreme value theory points out that the distribution beyond a threshold could be from a family of distributions (Frechet, Weibull or Gumbel).

There is a second approach, namely the Peak-Over-Threshold (POT), which defines distribution of excesses over a threshold u as:

$F_u(y) = P[X-u \leq y | X > u]$, for $0 \leq y < x-u$.

The distribution of excesses represents the probability that the variable (X) exceeds the threshold u by at most an amount y (y=x-u), given the information that X exceeds the threshold. In terms of the underlying F

$$F_u(y) = \frac{F(y+u) - F(u)}{1 - F(u)}$$

The underlying distribution function may have an infinite right endpoint, i.e. it allows the possibility of an arbitrarily very large loss with very small probability. It describes that the limiting distribution of excess over threshold flows the Generalized Pareto Distribution (GPD).

The descriptive statistics of both the data sets (below and above 10 days of LOS) is given in Table 3.

**Table 3. Descriptive Statistics of LOS (below and above 10 days)**

| *Data Set* | *LOS <10 days* | *LOS >10 days* |
|---|---|---|
| Sample Size | 34094 | 1501 |
| Mean | 3.0771 | 21.5819 |
| Standard Deviation | 2.1420 | 11.6082 |
| $Q_1$ | 1.5000 | 11.5000 |
| Median ($Q_2$) | 2.5000 | 17.5000 |
| $Q_3$ | 4.5000 | 28.5000 |
| Skewness | 0.8950 | 1.1770 |
| Kurtosis | 3.232 | 3.908 |

The box plot of the tail (LOS of more than 10 days) is presented in Fig 7.

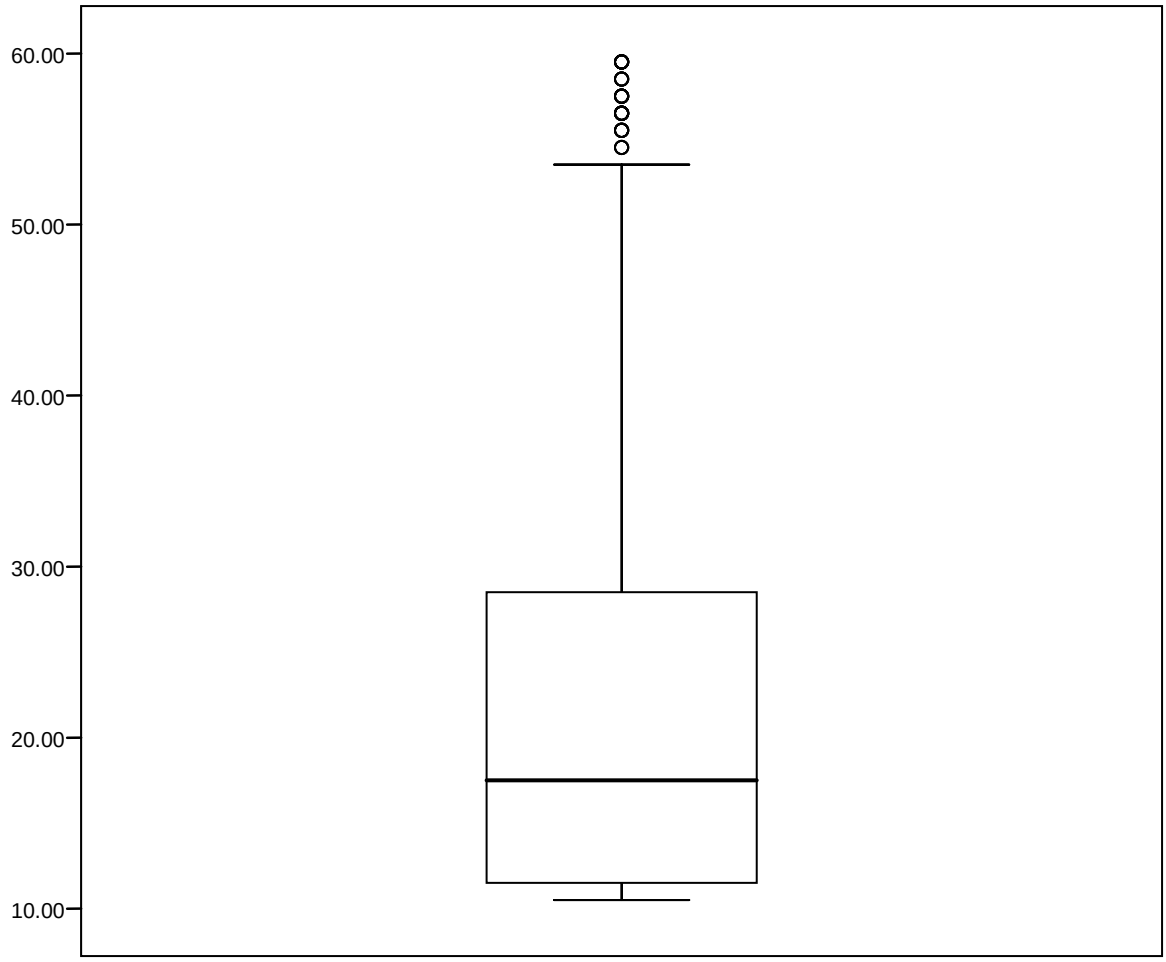


**Figure 7. Box plot of LOS for more than 10 days**

The EasyFit has shown that Weibull distribution with 2-parameters fits well for LOS of over 10 days and the estimated parameters are $\alpha = 0.88741$ and $\beta = 10.755$. The QQ plot (Fig 8) confirms the suitability of the distribution for the tail.

For example, the probability that the LOS would be 37 days, given that it exceeds 10 days is 0.007328 resulting in an expected frequency of 11. The actual data has also the same frequency for this.

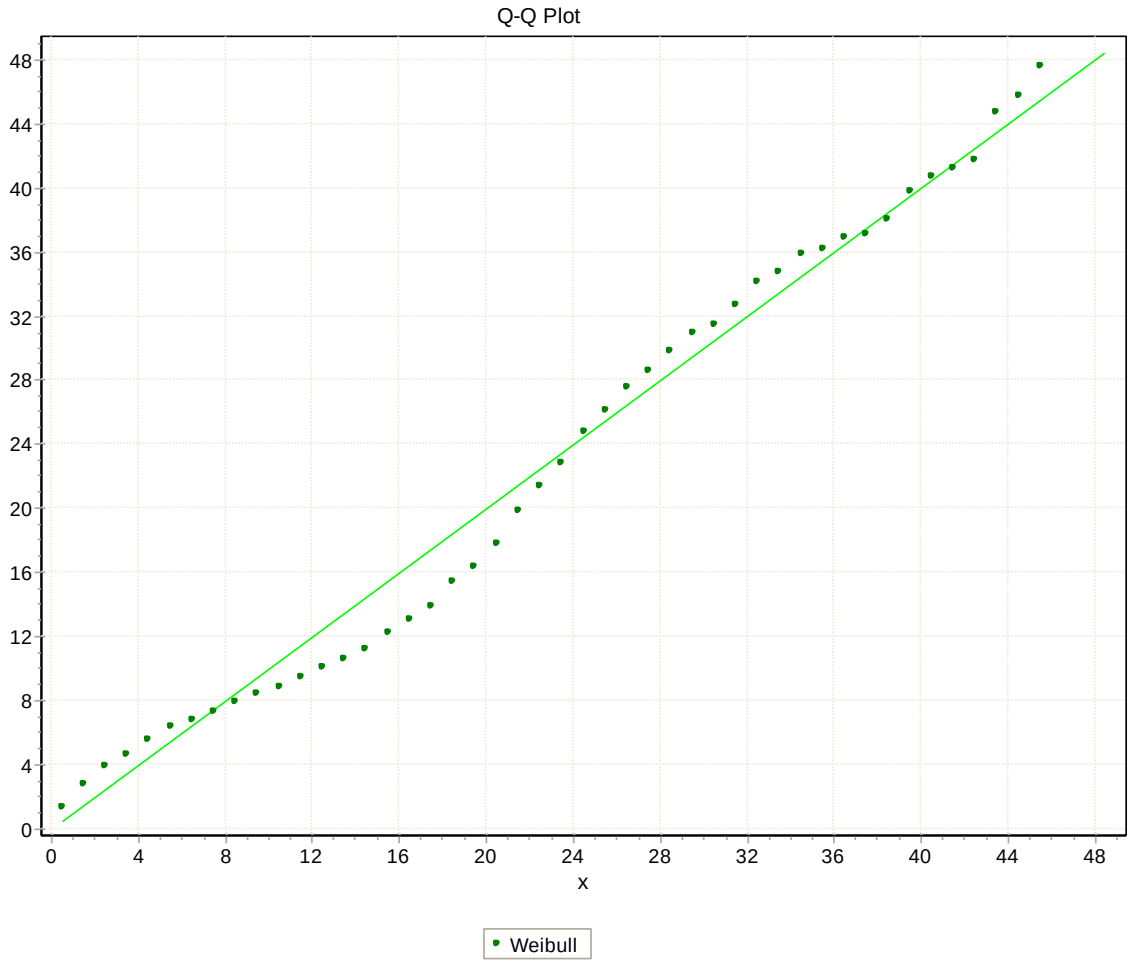


**Figure 8. QQ plot of LOS (exceedances over 10 days)**

From the above analysis, it can be concluded that for the data set used, lognormal distribution fits properly for LOS up to 60 days. However, as it has a long right tail, the appropriate distribution for LOS of more than 10 days is Weibull. Using this distribution, the insurer can properly arrive at a premium, which can take care of extreme events also. Having fitted the distribution of LOS, it needs to be seen which factors affect the variations in LOS. If the determining factors are known, regression techniques can be used.

However, exploratory data techniques could also show whether differences exist across factors. In the following, the LOS across genders is seen for ascertaining differences due to gender. Similar questions are also answered with respect to age, diseases, etc. These can be checked from regression techniques subsequently.

## 3. Difference in LOS among Age groups

LOS according to age groups is presented in the following table.

**Table 4. Percentage distribution of length of stay by age-groups**

| LOS (Days) | Age - groups | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 00 to 09 | 10 to 19 | 20 to 29 | 30 to 39 | 40 to 49 | 50 to 59 | 60 to 69 | 70 to 79 | 80 to 89 | 90 + | Total |
| 00 to 01 | 5.10 | 9.22 | 6.55 | 8.77 | 13.04 | 22.59 | 30.38 | 32.64 | 23.83 | 14.29 | 14.84 |
| 01 to 02 | 22.36 | 22.66 | 19.47 | 20.49 | 23.45 | 23.75 | 24.87 | 22.25 | 16.80 | 14.29 | 22.16 |
| 02 to 03 | 24.76 | 21.17 | 19.60 | 20.13 | 16.96 | 16.25 | 11.43 | 9.49 | 10.55 | 14.29 | 17.81 |
| 03 to 04 | 18.75 | 17.43 | 15.16 | 15.41 | 12.20 | 10.84 | 8.17 | 7.67 | 10.16 | 14.29 | 13.32 |
| 04 to 05 | 10.56 | 11.81 | 12.70 | 11.77 | 10.67 | 7.45 | 6.01 | 6.76 | 11.33 | 0.00 | 9.99 |
| 05 to 06 | 7.70 | 6.91 | 9.66 | 8.16 | 7.68 | 5.48 | 4.46 | 5.55 | 7.03 | 0.00 | 7.22 |
| 06 to 07 | 4.04 | 3.70 | 5.33 | 5.30 | 4.38 | 3.47 | 3.13 | 3.13 | 3.52 | 14.29 | 4.26 |
| 07 to 08 | 2.47 | 2.64 | 4.97 | 3.44 | 3.18 | 2.31 | 2.52 | 2.37 | 4.30 | 0.00 | 3.19 |
| 08 to 09 | 1.08 | 1.39 | 2.25 | 1.77 | 2.01 | 1.23 | 1.73 | 2.22 | 1.17 | 0.00 | 1.73 |
| 09 to 10 | 0.64 | 0.72 | 1.43 | 1.28 | 1.33 | 1.16 | 1.40 | 2.02 | 1.95 | 0.00 | 1.25 |
| 10 to 20 | 1.44 | 1.30 | 2.02 | 1.92 | 2.25 | 3.06 | 3.22 | 3.53 | 7.03 | 0.00 | 2.36 |
| 20 to 30 | 0.57 | 0.24 | 0.39 | 0.59 | 1.41 | 1.23 | 1.40 | 1.36 | 1.17 | 0.00 | 0.89 |
| 30 to 40 | 0.28 | 0.48 | 0.29 | 0.73 | 1.02 | 0.81 | 0.92 | 0.76 | 0.39 | 0.00 | 0.66 |
| 40 to 50 | 0.15 | 0.19 | 0.05 | 0.08 | 0.20 | 0.20 | 0.23 | 0.20 | 0.00 | 28.57 | 0.15 |
| 50 to 60 | 0.08 | 0.14 | 0.14 | 0.15 | 0.22 | 0.18 | 0.14 | 0.05 | 0.78 | 0.00 | 0.15 |
| **Total** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **Number** | **3882** | **2083** | **6583** | **5930** | **4976** | **5453** | **4443** | **1982** | **256** | **7** | **35595** |

From the table it can be seen that around 78 per cent of the patients stayed in the hospital for less than 5 days. Across age groups, differences are observed in LOS for more than 5 days. Longer stay in the hospital is observed in case of patients of more than 49 years. From this, one can conclude that age has an impact on LOS.

**Difference in LOS among Illness categories:**

The average LOS and other descriptive statistics according to illness categories are presented in Table 5. It is observed from the table that there is a variation among mean LOS across the diseases. The average length of stay in case of 'Eye' and 'Ear 'is quite low.

**Gender difference in LOS**

The LOS separately for males and females are given in Table 6 and the corresponding graph is given in Fig 9. It can be seen that the movements between the genders are more or less similar. Further, the quartiles of both the data sets are the same, which suggests that there are no overall differences in the LOS between males and females.

**Table 5. Descriptive Statistics of LOS according to Illness categories**

| *ILLNESS CATEGORIES* | *N* | *P* | *MEAN* | *SD* | *SKEWNESS* | *KURTOSIS* |
|---|---|---|---|---|---|---|
| ARTHROPATHY (M00-M99) | 1098 | 0.0308 | 5.6494 | 6.1118 | 3.460 | 15.331 |
| BLOOD (D50-89) | 191 | 0.0054 | 4.6204 | 5.1346 | 4.008 | 18.485 |
| CIRCULATORY (I00-I99) | 2400 | 0.0674 | 4.7283 | 5.7512 | 4.372 | 25.399 |
| DIGESTIVE (K00-K93) | 4298 | 0.1207 | 3.7592 | 4.0247 | 5.511 | 44.034 |
| EAR (H60-H95) | 442 | 0.0124 | 2.8552 | 4.0292 | 8.423 | 92.953 |
| ENDOCRINE (E00-E90) | 625 | 0.0176 | 6.0760 | 8.6982 | 3.057 | 10.127 |
| EYE (H00-H59) | 5211 | 0.1464 | 1.2298 | 2.7105 | 10.718 | 136.478 |
| GENITOURINARY (N00-N99) | 3549 | 0.0997 | 4.3123 | 6.6198 | 5.077 | 30.194 |
| INFECTIONS (A00-B99) | 5545 | 0.1558 | 4.1777 | 3.0938 | 5.255 | 53.245 |
| INJURY (S00-T98) | 3037 | 0.0853 | 3.9300 | 5.2012 | 4.500 | 26.288 |
| MENTAL (F00-F99) | 46 | 0.0013 | 4.5870 | 6.0363 | 2.950 | 10.154 |
| NEOPLASM (C00-D48) | 1703 | 0.0478 | 4.4706 | 6.3215 | 3.952 | 20.261 |
| NERVOUS (G00-G99) | 432 | 0.0121 | 5.1366 | 5.9261 | 4.127 | 23.344 |
| PERINATAL (P00-P96) | 141 | 0.0040 | 5.1525 | 6.0190 | 4.631 | 24.238 |
| PREGNANCY (O00-O99) | 3134 | 0.0880 | 4.7198 | 3.1514 | 4.224 | 37.234 |
| RESPIRATORY (J00-J99) | 2968 | 0.0834 | 4.0185 | 4.4708 | 6.352 | 54.460 |
| SKIN (L00-L99) | 775 | 0.0218 | 4.3258 | 6.4435 | 4.307 | 27.712 |
| **ALL COMBINED** | **35595** | **1.0000** | **3.8574** | **4.8892** | **5.236** | **37.952** |

**Table 6. Percentage distribution of length of stay by gender**

| *LOS (Days)* | *Gender* | | |
|---|---|---|---|
| | *Female* | *Male* | *Total* |
| 00 to 01 | 13.72 | 15.98 | 14.84 |
| 01 to 02 | 21.42 | 22.93 | 22.16 |
| 02 to 03 | 17.06 | 18.58 | 17.81 |
| 03 to 04 | 13.58 | 13.07 | 13.32 |
| 04 to 05 | 10.93 | 9.04 | 9.99 |
| 05 to 06 | 7.71 | 6.72 | 7.22 |
| 06 to 07 | 4.55 | 3.97 | 4.26 |
| 07 to 08 | 3.78 | 2.59 | 3.19 |
| 08 to 09 | 1.87 | 1.59 | 1.73 |
| 09 to 10 | 1.44 | 1.06 | 1.25 |
| 10 to 20 | 2.19 | 2.53 | 2.36 |
| 20 to 30 | 0.92 | 0.86 | 0.89 |
| 30 to 40 | 0.62 | 0.70 | 0.66 |
| 40 to 50 | 0.13 | 0.18 | 0.15 |
| 50 to 60 | 0.09 | 0.21 | 0.15 |
| Total | 100.00 | 100.00 | 100.00 |
| Number | 17999 | 17596 | 35595 |
| AVG LOS | 3.9206 | 3.7929 | 3.8574 |
| SD | 4.6525 | 5.1194 | 4.8892 |
| SKEWNESS | 5.036 | 5.368 | 5.236 |
| KURTOSIS | 36.276 | 38.605 | 37.952 |
| $Q_1$ | 1.5000 | 1.5000 | 1.5000 |
| $Q_2$ | 2.5000 | 2.5000 | 2.5000 |
| $Q_3$ | 4.5000 | 4.5000 | 4.5000 |

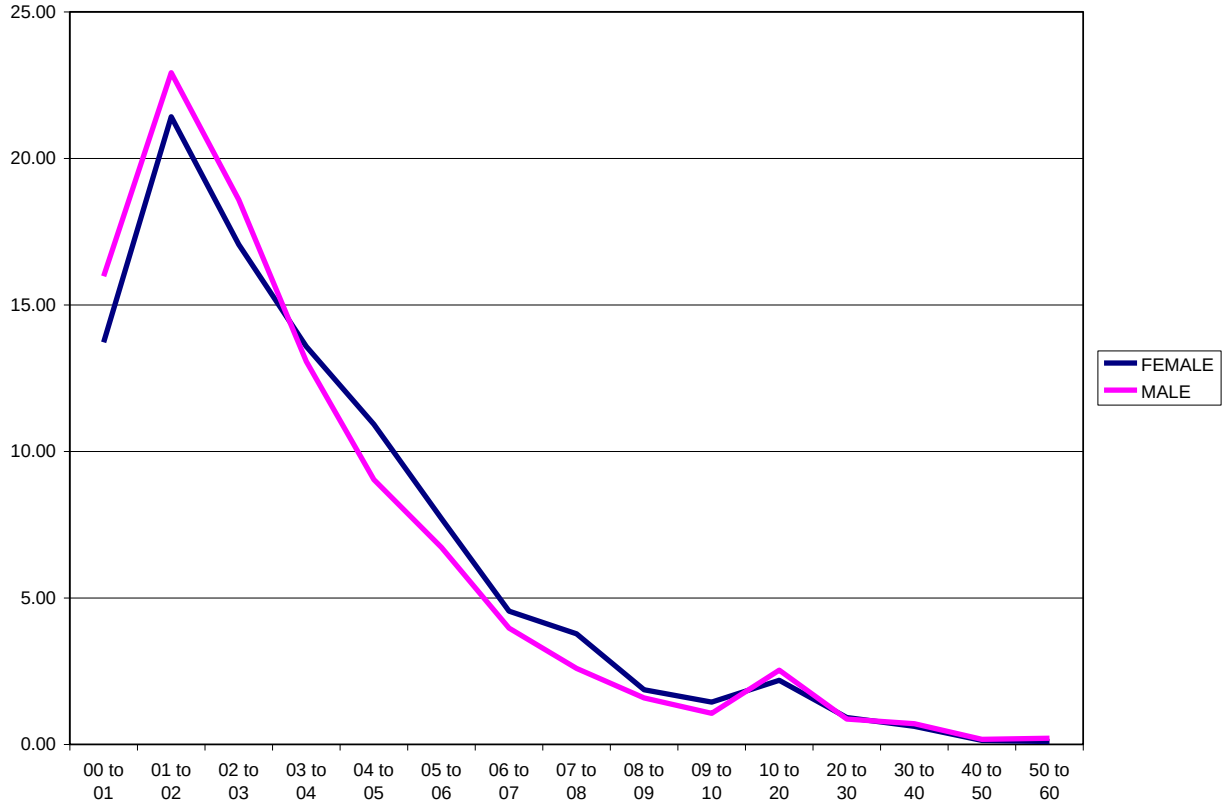


**Figure 9. Distribution of length of stay by gender**

A statistical test has also been conducted to ascertain whether differences exist in LOS across the above factors namely, age, gender and illness group. For comparing the means of two independent samples, it is common to use t-test. However, the same is not appropriate if the underlying distribution happens to be skewed. One usual way to overcome with this problem is to apply t-test to the data after the data is transformed into logarithms. Chand (1950), Lehman (1975) etc. showed that such tests may have type I error rates that are very different from the nominal levels when the two population variances are not equal leading to inappropriateness in the test procedure. To take care of this, Zhou et al (1997) proposed two methods namely, the likelihood-based test and the bootstrap-based test. Following Zhou et al (1997), let Xi and Yi, be the outcomes of the ith female and jth male patient with means μ1 and μ2 respectively.

$\log X_i \sim N(\mu_1 , \sigma^2_1)$ and,

$\log Y_i \sim N(\mu_2 , \sigma^2_2)$

The maximum likelihood estimators for $\mu_1$ and $\mu_2$ are:

$\mu_1 = (\sum \log X_i) / n_1$, over 1 to $n_1$ and,

$\mu_2 = (\sum \log Y_i) / n_2$, over 1 to $n_2$

The unbiased estimators of $\sigma^2_1$ and $\sigma^2_2$ are defined as:

$S^2_1 = [\sum \log( X_i - \mu_1)^2] / (n_1 -1)$, over 1 to $n_1$ and,

$S^2_2 = [\sum \log (Y_i - \mu_2)^2] / (n_2 - 1)$, over 1 to $n_2$

Zhou et al (1997) proposed the test statistics

$$Z = \frac{(\mu_2 - \mu_1) + (1/2)(S_2^2 - S_1^2)}{SQRT[(S_1^2/n_1) + (S_2^2/n_2) + (1/2)\{S_1^4/(n_1 - 1)\} + (1/2)\{S_2^4/(n_2 - 1)\}]}$$

When sample sizes ($n_1$ and $n_2$) are both large, the distribution of Z can be approximated to standard normal, N (0, 1) under $H_0$.

A positive value of Z-statistic will imply greater mean for males than the females and vice-versa. The Z-statistic for common diseases between males and females works out to +0.205 and is not statistically significant at 5 per cent level of significance. It may be mentioned here that ALOS of females after deleting pregnancy records works out to 3.7521 days, lower than the males (3.7929 days). The Z-values for various illness groups are furnished in Table 7.

**Table 7. Z-Values for Average LOS Differences**

| *ILLNESS CATEGORIES* | *G* | *N* | *μ* | *$S_i$* | *Z – Values (P-Values)* |
|---|---|---|---|---|---|
| ARTHROPATHY (M00-M99) | M | 529 | 1.240322 | 0.873541 | - 2.6740 |
| | F | 569 | 1.355578 | 0.938122 | (0.00750)* |
| BLOOD (D50-89) | M | 93 | 1.114220 | 0.804944 | - 0.7067 |
| | F | 98 | 1.187327 | 0.834117 | |
| CIRCULATORY (I00-I99) | M | 1545 | 1.146087 | 0.883606 | - 0.0593 |
| | F | 855 | 1.157649 | 0.873412 | |
| DIGESTIVE (K00-K93) | M | 2407 | 1.018852 | 0.776368 | - 0.1547 |
| | F | 1891 | 1.015456 | 0.786151 | |
| EAR (H60-H95) | M | 198 | 0.760224 | 0.716367 | +0.7947 |
| | F | 244 | 0.752481 | 0.642592 | |
| ENDOCRINE (E00-E90) | M | 304 | 1.026563 | 1.165706 | - 1.5072 |
| | F | 321 | 1.267967 | 1.108587 | |
| EYE (H00-H59) | M | 2804 | - 0.189223 | 0.686482 | - 0.3653 |
| | F | 2407 | - 0.198830 | 0.711569 | |
| GENITOURINARY (N00-N99) | M | 1601 | 0.978200 | 0.921215 | +1.0271 |
| | F | 1948 | 1.018228 | 0.834465 | |
| INFECTIONS (A00-B99) | M | 3132 | 1.248540 | 0.636185 | +1.1352 |
| | F | 2413 | 1.228420 | 0.634162 | |
| INJURY (S00-T98) | M | 1980 | 0.946610 | 0.852669 | - 0.1483 |
| | F | 1057 | 0.959608 | 0.843922 | |
| MENTAL (F00-F99) | M | 26 | 0.855723 | 1.251158 | +0.2792 |
| | F | 20 | 0.935266 | 1.074839 | |
| NEOPLASM (C00-D48) | M | 521 | 0.937008 | 1.170511 | +2.8610 |
| | F | 1182 | 0.927292 | 0.979753 | (0.00422)* |
| NERVOUS (G00-G99) | M | 230 | 1.327556 | 0.826458 | +1.5616 |
| | F | 202 | 1.220575 | 0.785959 | |
| PERINATAL (P00-P96) | M | 85 | 1.432186 | 0.713004 | +2.7563 |
| | F | 56 | 1.279262 | 0.464341 | (0.00584)* |
| RESPIRATORY (J00-J99) | M | 1720 | 1.115731 | 0.725028 | +0.3602 |
| | F | 1248 | 1.105896 | 0.723559 | |
| SKIN (L00-L99) | M | 421 | 1.002500 | 0.889086 | +0.0138 |
| | F | 354 | 0.967407 | 0.926557 | |
| **ALL COMBINED** | **M** | 17596 | 0.883398 | 0.931644 | +0.4206 |
| | **F** | 14865 | 0.876375 | 0.933562 | |

*Significant P-values

It can be seen from the above table that for all diseases, there are no significant differences in the average LOS between males and females. However, significant differences were observed for specific diseases.

The differences in average length of stay for males and females are found to be statistically significant for three diseases viz. Arthropathy (0.75 per cent level), Neoplasm (0.422 per cent level) and Perinatal (0.584 per cent level). In case of Arthropathy, the ALOS of females is significantly higher than their males counterpart, while it is opposite in the case of Neoplasm and Perinatal categories.

## 4. Regression Model for LOS

The above analysis suggests that while gender may not have a significant impact on LOS, the type of disease and age could impact the LOS. In order to test this, a regression set up in logarithmic form with LOS as dependent variable and age, gender & diseases as independent variables was attempted. In order to take care of the length of stay of zero-day, 0.5 day is added to each LOS following Lim and Tongkumchum (2009). Further, we discard cases of LOS of more than 59 days [i.e. (LOS+0.5) more than 59.5 days].Besides the above determinants, LOS can be influenced by the type of hospital and its discharge policy. But as data on these are not available, the following regression set up is used.

$$\text{Ln (LOS+0.5)} = \beta_0 + (\beta_1 * \text{age}) + (\beta_2 * \text{gender}) + (\beta_3 * \text{age} * \text{gender}) + \sum(\beta_j * \text{illness}) + \text{Error term}$$

$j = 4$ to 20.

Age is in years; gender = 0 if female, 1 if male and illness is categorical variable ($\beta_j$), which takes 1 or 0 according as whether the patient suffers from $j^{th}$ disease or not. The co-efficient, $\beta_3$ , is the co-efficient for the interaction between age and gender. The regression analysis is carried out in the SPSS software. The estimated coefficients are given in Table 8.

**Table 8. Estimates of regression co-efficient**

| *Variable* | *Co-efficient* | *Standard Error* | *t-ratios* |
|---|---|---|---|
| Constant | + 0.934 | 0.004 | + 239.016 |
| Age | + 0.008 | 0.000 | + 95.639 |
| Gender | + 0.008 | 0.004 | + 1.799 |
| Age * Gender | + 0.000 | 0.000 | + 0.707 |
| Arthropathy | - 1.984 | 0.006 | - 320.193 |
| Blood | - 1.921 | 0.014 | - 141.099 |
| Circulatory | - 2.056 | 0.005 | - 429.882 |
| Digestive | - 1.225 | 0.004 | - 322.912 |
| Ear | - 0.815 | 0.009 | - 89.105 |
| Endocrine | - 0.911 | 0.008 | - 115.372 |
| Eye | - 0.878 | 0.004 | - 220.223 |
| Genitourinary | - 0.333 | 0.004 | - 82.410 |
| Injury | + 0.218 | 0.004 | + 51.900 |
| Mental | + 0.240 | 0.027 | + 8.770 |
| Neoplasm | +0.226 | 0.005 | + 42.588 |
| Nervous | +0.428 | 0.009 | + 46.116 |
| Perinatal | + 0.762 | 0.016 | + 48.130 |
| Pregnancy | + 0.633 | 0.004 | + 145.419 |
| Respiratory | + 1.052 | 0.004 | + 250.278 |
| Skin | + 2.131 | 0.007 | + 299.628 |

The estimated $R^2$ = 0.96 suggests the appropriateness of the model. The coefficient of interaction term is near zero. As observed in the exploratory analysis, the gender is statistically insignificant and age turns out to be statistically significant. Illness turned out to be highly significant. The plot of the residuals is close to standard normal distribution.

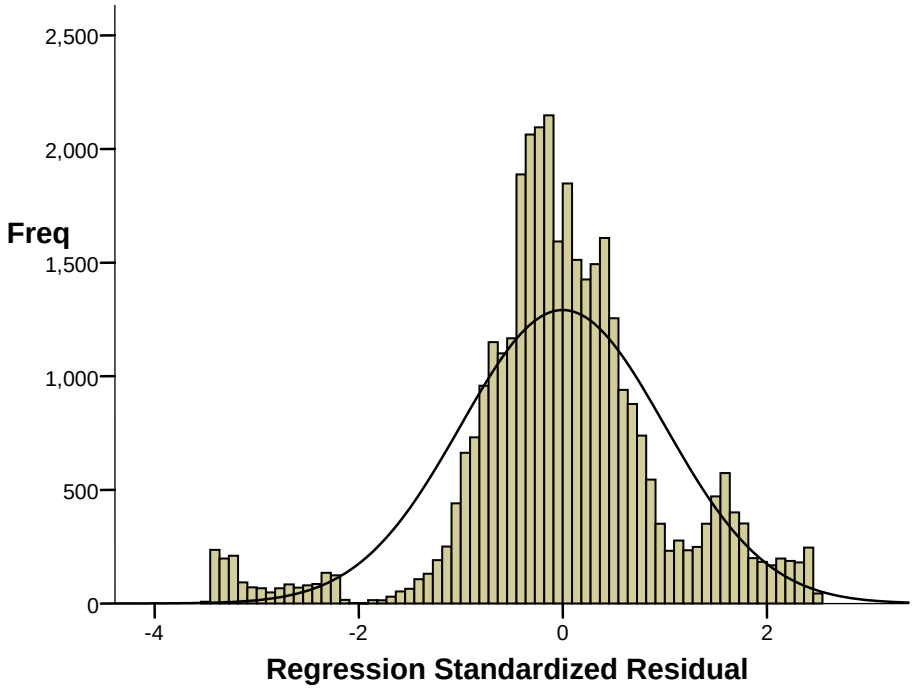


**Figure 10. Histogram of residuals and comparison with Standard Normal (0,1)**

## 5. Discussion and Conclusions

The length of stay at hospital by patients contains important information, which can be used for policy making and future projections on resource requirements by government as well as hospitals. This is also an important ingredient for insurance companies from the point of view of severity of claims.

The disease elasticity of LOS will be of help to hospital management in estimating the release of beds. Insurers can use such studies in calculating the likely losses in the case of hospitalization given the probability of a disease. However, the above can be achieved only when the regression equations are stable across samples and over time. As the results obtained in this paper are relevant to the sample, which is quite representative with respect to geographical area, hospitals, diseases and gender, repetitive experiments will be more meaningful.

The second line of research could be with respect to diseases and hospitals, which throws more insights into the hospital management policies and treatments. For insurer, if premium data are also available, actuaries can gain in fine-tuning the rates.

## References

Atienza, N (2005), "Fitting the variable 'Length of Hospital stay' with mixtures from different distribution families". University of Seville, Spain.

Brys G, Hubert M and Struyf A (2003), "A robust measure of skewness", *Journal of Computational and Graphical Statistics*, 13, 996-1017.

Chand, U (1950), "Distributions related to comparison of two means and two regression coefficients", *Annals of Mathematical Statistics*, 21, 507-522.

Fisher, R & Tippett, L (1928), "Limiting forms of the frequency distribution of the largest or smallest member of a sample", Proceedings of the Cambridge Philosophical Society 24, 180-190.

Hellervik, A & Rodgers, G J (2006), A power law distribution in patient's length of stay at hospital.

Jenkinson, A F (1955), "The frequency distribution of the annual maximum (or minimum) values of meteorological events", *Quarterly Journal of the Royal Meteorological Society*, 81, 158-171.

Kulinskaya, E, Kornbrot, D and Gao, H (2005), "Length of stay as a performance indicator: robust statistical methodology", *IMA Journal of Management Mathematics*, 16(4), 369-381.

Lehman, E L (1975), *Nonparametrics: Statistical methods based on ranks*, San Francisco: Holden-Day.

Lim, A & Tongkumchum P (2009), "Methods for analyzing hospital length of stay with application to inpatients dying in southern Thailand", *Global Journal of Health Science*, Vol.1, No.1, 27-38.

Maclean L C & Richman A (2001), "Resource absorption in a health service system", *Health Care Management Science*, 4, 337-345.

Marazzi, A, Paccaud, F, Ruffieux, C & Beguin, C (1998), Medical Care 36, 915.

Martin S & Smith P (1996), "Explaining variations in inpatient length of stay in the National Health Service", *Journal of Health Economics*, 15, 279-304.

Zhou, X H, Gao, S and Hui S L (1997), "Methods for comparing the means of two independent log-normal samples", *Biometrics*, 53, 1129-1135.

♦♦♦