Generalized robust statistics method for estimating average length of stay in hospitals

B. Senthilkumar and R. Ramakrishnan
National Institute of Epidemiology, Ayyapakkam, Chennai - 600 077, India
senthilkumarbalan@gmail.com, drramakrishnan.nie@gmail.com

Abstract
Hospital length of stay is an important performance indicator for hospital management and a key measure of efficiency in health care. However, it is difficult to analyze as most distributions pertaining to hospital length of stay are asymmetric due to outliers. The objective of the present study is to suggest a new method to estimate average length of stay (ALOS). Data on LOS in the year 2010 from a corporate hospital was considered for analysis. ALOS was calculated using generalized robust statistics method (GRSM) and other existing methods. The comparison of results shows that GRSM is a better alternative over other existing methods.

Keywords: Health management; hospital stay; health care.

Introduction
Length of stay (LOS) is one of the performance indicators in hospital administration. Also the type of reimbursement or health insurance plan plays a significant role in the patient's length of stay in hospitals. LOS in hospitals is a crucial variable for the quality of life of the patients and their families. Furthermore, it is the most important component in the consumption of hospital resources (Lave & Leinhard, 1976). The log transformation on LOS data is known to generate a near normal distribution which does not require sophisticated methods such as Box Cox transformations (Kulinskaya et.al., 2001). A few outliers can completely distort both LOS means and comparisons based on them (Mc Millan & Hyzy, 2007) have reviewed the different approaches to quality improvement in the intensive care units, including measures such as mortality and length of stay.

A common practice is to trim values according to various empirical rules, but there is little theoretical support for choosing between alternative procedures. Reduction of LOS of a patient and predicting the LOS in the hospital will help the hospital administration. Hence, hospitals can improve their quality of service to the patients with the available resources which in turn improves the patient's satisfaction level and reduction in the cost.

The LOS is difficult to analyze because most distributions of LOS are asymmetric, with a long right tail and some very large observations called outliers. Since the values and the frequency of outliers typically fluctuate from sample to sample, the mean and the related inferences are very unstable. These features vitiate the reliability of many statistical summaries pertaining to hospital data.

The objective of the present study is to suggest a better method to estimate ALOS.

Methodology for estimating ALOS
The different methods for estimating ALOS are explained in the following section.

Method I
This method gives the simple arithmetic mean of LOS. This is the common method which is used to estimate ALOS.

Average length of stay (in days) = Total discharge days / Total discharges

where,
Total discharge days is the sum of the number of days spent in the hospital for each inpatient who is discharged during the time period regardless of when he / she is admitted,
Total discharges is the number of inpatients discharged from the hospital during the time period.
This figure includes deaths but not births.

Method II
This method is different from method I in the sense that it considers admissions rather than discharges.

Average length of stay (in days) = Total inpatient days of care / Total admissions

where,
Total inpatient days of care is the sum of each daily inpatient census for the time period.
Total admissions is the total number of individuals admitted into inpatient units during the said time period examined. As in method I, births are excluded in this method also. This method is commonly applied in hospitals in California to estimate ALOS. They have considered inpatient days and total number of admissions in a particular month.

Method III
Method III is given by the formula,

Average length of stay (in days) = Median (Patient LOS)

Weissman (1997) analysed the ICU LOS data and found that the average of the mean LOS of all diagnostic groups was higher than the average of the median and mode, reflecting the rightward skew ness of the LOS frequency distribution. He confirmed that the median was the most useful descriptor of central tendency for their data. Wright et.al., (2003) did a study on factors influencing the length...
has a mixed generalized Erlang distribution to estimate the predictive distribution of LOS in hospitals. They also addressed the problem of optimizing the number of beds in hospitals with respect to average cost per unit time. They have applied box plot graphs to calculate ALOS. In the box plots it is observed that the data are very highly skewed with a large number of outliers. Method VI: Truncated or trimmed mean

A truncated mean is defined as the arithmetic mean of the data contained between boundaries. The computation of the boundaries is based on three numbers: a measure of position, a measure of scale and a measure of factor. The boundaries are set at a certain distance from the position, the distance being the product of scale times the factor. The boundaries are computed on the original scale or transformed scale. A sample truncated mean is an estimate of the population mean. The notations for the different truncation rules and the corresponding lower boundary ‘a’ and upper boundary ‘b’ are given in Table 1and Table 2 respectively:

Average length of stay = Total inpatient days (within the boundaries) / Total admissions, a < LOS < b. Ruffieux et al. (2000) has discussed the calculation of ALOS using truncated or trimmed means for a hospital data from European countries.

Method VII: Approximated quartile based truncated mean (AQTMM)

It is a mathematical method for deriving lower and upper boundaries using statistical principles. The AQTMM is a particular truncated mean. A major feature of the AQTMM is that it takes into account the distribution of the sample data which is a mixture of a regular distribution that includes rare but expected long stays and a contaminating distribution that describes the irregular, exceptional and unexpected stays. The regular distribution is described with the help of a parametric model. The AQTMM is aimed at estimating the mean of the regular distribution, after removal of extreme cases that are likely to belong to the contaminating distribution. This procedure is a flexible approach to the management of both a variety of distributions and proportion of outliers.

Table 3. Formula for lower (a), upper (b) boundaries and factors f1 and f2 based on statistical distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>a</th>
<th>b</th>
<th>f1</th>
<th>f2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>ln(a) = ln(Q2) - f1*ln(Q3)/ln(Q2)</td>
<td>ln(b) = 1.36s + 0.2s²</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td>f2*ln(Q3)/ln(Q2)</td>
<td>0.153s + 0.024s²</td>
<td>1.71 - 0.437s - 0.071s²</td>
<td></td>
</tr>
<tr>
<td>Lognormal</td>
<td>1.72 - 0.55s</td>
<td>1.725</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where s = ln(Q3) - ln(Q2).

of hospital stay of patients with heart failure. They have applied median as a measure for calculating ALOS. They found a list of factors influencing ALOS for patients who have LOS more than the median length of stay. Ramakrishnan et al. (2009) discussed about calculating ALOS while defining the quality indicators for ICU. They found that arithmetic mean overestimates LOS, as outliers both ways influence the mean LOS very adversely. They also found that median of LOS can circumvent this problem.

Method IV

Method IV considers logarithmic transformations as variation among the total discharge days or total inpatient days can be minimized by this transformation. Average length of stay (in days) = Antilog (Σlog (total inpatient days) / total admissions)

Method V: BOX PLOT graphs

For the data with large observations a common practice is to remove outliers in estimating the ALOS. An informal approach to identify outliers is based on visual inspection of the data using box plot graphs. However, more formal procedure is the determination of lower and upper boundaries for the ALOS using some predefined truncation rules. The truncation rules have the practical advantages of objectivity and simplicity. After truncating the outliers ALOS can be calculated. Concepcion Austin et al. (2003) have modeled the distribution of LOS in hospital for geriatric patients. They assumed that LOS

Table 1. Notations for truncation rules

<table>
<thead>
<tr>
<th>Truncation rule</th>
<th>Transformation</th>
<th>Position</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tlmi</td>
<td>Logarithmic</td>
<td>Median</td>
<td>Interquartile range</td>
</tr>
<tr>
<td>Tlmm</td>
<td>Logarithmic</td>
<td>Median</td>
<td>Median absolute deviation</td>
</tr>
<tr>
<td>Tlqi</td>
<td>Logarithmic</td>
<td>Quartile</td>
<td>Interquartile range</td>
</tr>
<tr>
<td>Tlms</td>
<td>Logarithmic</td>
<td>Mean</td>
<td>Standard error</td>
</tr>
</tbody>
</table>

Table 2. Formula for lower and upper boundaries

<table>
<thead>
<tr>
<th>Truncation rule</th>
<th>Lower boundary a</th>
<th>Upper boundary b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tlmi</td>
<td>ln(a) = ln(Q1) - 1.5</td>
<td>ln(b) = ln(Q3) + 1.5</td>
</tr>
<tr>
<td>Tlmm</td>
<td>ln(a) = ln(Q1) - 1.15</td>
<td>ln(b) = ln(Q3) + 1.15</td>
</tr>
<tr>
<td>Tlqi</td>
<td>ln(a) = ln(Q1) - 3mad(ln(x))</td>
<td>ln(b) = ln(Q3) + 3mad(ln(x))</td>
</tr>
<tr>
<td>Tlms</td>
<td>ln(a) = average</td>
<td>ln(b) = average+3SD</td>
</tr>
</tbody>
</table>

where Q1, Q2 and Q3 denotes the first, second (median) and the third quartiles, mad{.} is the median absolute deviation, SD is the standard deviation and x is the particular LOS.

Table 3. Formula for lower (a), upper (b) boundaries and factors f1 and f2 based on statistical distributions

The lower and upper boundaries of the data can be different based on the LOS distributions like the lognormal, the Weibull or the gamma models which are given in Table 3.

Ruffieux et al. (2000) have found a set of new methods for calculating ALOS which are termed as Approximated Quartile based Truncated Mean (AQTMM). Based on the distribution of the data the upper and lower boundaries are found. The numbers below the lower
boundary and above the upper boundary are ignored and ALOS is estimated for the remaining data.

Table 4. Estimates based on measures of central tendencies

<table>
<thead>
<tr>
<th>Method</th>
<th>ALOS (in days)</th>
<th>a (in days)</th>
<th>b (in days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>3.3337</td>
<td>0</td>
<td>71.66</td>
</tr>
<tr>
<td>II</td>
<td>2.9110</td>
<td>0</td>
<td>71.66</td>
</tr>
<tr>
<td>III</td>
<td>1.64</td>
<td>0</td>
<td>71.66</td>
</tr>
<tr>
<td>IV</td>
<td>1.2665</td>
<td>0</td>
<td>71.66</td>
</tr>
<tr>
<td>V</td>
<td>-</td>
<td>0</td>
<td>71.66</td>
</tr>
</tbody>
</table>

Table 5. Estimates based on truncated means

<table>
<thead>
<tr>
<th>Method</th>
<th>ALOS (in days)</th>
<th>a (in days)</th>
<th>b (in days)</th>
<th>P1 (%)</th>
<th>P2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiqi</td>
<td>2.0223</td>
<td>0</td>
<td>-</td>
<td>7.8993</td>
<td>8.57%</td>
</tr>
<tr>
<td>Timi</td>
<td>4.6645</td>
<td>2.09</td>
<td>60.13%</td>
<td>12.7649</td>
<td>4.1%</td>
</tr>
<tr>
<td>Timm</td>
<td>1.4629</td>
<td>0.4709</td>
<td>5.74%</td>
<td>3.4308</td>
<td>25.41%</td>
</tr>
<tr>
<td>Tiqi</td>
<td>2.4286</td>
<td>0.1957</td>
<td>2.68%</td>
<td>12.21</td>
<td>4.37%</td>
</tr>
<tr>
<td>Tims</td>
<td>1.8560</td>
<td>0.2822</td>
<td>3.41%</td>
<td>5.6848</td>
<td>12.99%</td>
</tr>
<tr>
<td>Method VII AQT M</td>
<td>1.9086</td>
<td>0.0567</td>
<td>1.37%</td>
<td>6.45</td>
<td>11.05%</td>
</tr>
</tbody>
</table>

Table 6. Estimates based on advanced methods

<table>
<thead>
<tr>
<th>Method</th>
<th>ALOS (in days)</th>
<th>a (in days)</th>
<th>b (in days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIII</td>
<td>2.2174</td>
<td>0</td>
<td>71.66</td>
</tr>
<tr>
<td>IX</td>
<td>2.7571</td>
<td>0</td>
<td>71.66</td>
</tr>
</tbody>
</table>

Method VIII: Boot strap method

When the distribution of the data is not known, boot strap method could be applied to get an estimate of ALOS and the respective boundaries. From the available data of size n, there is a possibility to generate n groups of observations. This can be done by using statistical software like SPSS, Systat, MatLab etc. Average for each group of n observations can be calculated, which would result in n averages. Then average length of stay is calculated as the average of all the n averages.

Method IX: Robust statistics

A simple method to calculate ALOS is the robust statistics method. In this method 5th percentile and 95th percentile are calculated. The numbers below 5th percentile and above the 95th percentiles are considered equal to the respective percentile values. The average for all the n numbers is calculated without eliminating the outliers. This method is called robust statistics method. Marazzi et al. (1998) have applied robust statistical method for calculating ALOS. New robust procedures based on M-estimators are used in their study. Kulinskaya et al. (2005) have elucidated the factors like age, sex and Health Resource Groups classification influencing LOS. They have applied robust statistical method for estimating ALOS. Robust statistical methods address the problem of non normal data in a variety of ways, all of which down weight outliers. Robust methods predict the behavior of the outliers from the behavior of the more central values which ensures the accurate estimate of ALOS.

Method X: Generalized robust statistics

The generalized robust statistics method is introduced by the authors as a new method to estimate ALOS. This is a simple and accurate method based on robust statistics. The method computes lower boundary 'a' and upper boundary 'b' which are given as follows:

\[ \ln(a) = \text{average} \{\ln(x) - 3SD\{\ln(x)\}\} \]
\[ \ln(b) = \text{average} \{\ln(x) + 3SD\{\ln(x)\}\} \]

The data below the lower boundary 'a' and above the upper boundary 'b' are treated as 'a' and 'b' respectively.

Now the generalized robust statistics method is given by the average of all these n values.

Analysis of the data

The estimates of ALOS, a and b are calculated for the data on LOS from Jan.2010 to Dec.2010 from a 600+ bed corporate hospital in Chennai. This data base contains 5242 hospital stays in Critical Care Unit (CCU) with different diagnosis. The fitting of distribution of the LOS data is done using easy fit software. The statistical software like SPSS and MatLab are used to calculate other statistical measures.

Estimates of ALOS, lower (a) and upper (b) boundaries for the said data by using Method I to Method V are tabulated in Table 4. All the observations are considered in these five Methods. It is observed that ALOS is more for Method I compared to Method II. Method III gives lesser ALOS compared to Methods I and II, while using logarithmic transformation in Method IV leads to the least value of ALOS.

Box plot graph which is Method V for the said data is shown in Fig.1.

Similarly ALOS, lower (a) and upper (b) boundaries and the proportion of stays lesser than a (P1) and greater than b (P2) by using Method VI and Method VII are given in Table 5.

It is observed that nearly 64% of observations are truncated in Timi and 30% of observations in Timm which led to the usage of only 36% and 70% of the data.
respectively in Method VI. As Method VI and Method VII use different truncation rules giving different values of a and b and hence $P_1$ and $P_2$ respectively.

Estimates of ALOS, lower (a) and upper (b) boundaries for the said data by using Method VIII and Method IX are tabulated in Table 6.

The estimates of ALOS the lower and upper boundaries by using generalized robust statistics method, that is, Method X are given below

ALOS (in days): 1.5978
Lower boundary (a): 0
Upper boundary (b): 71.66

It can be observed that all the observations are considered in Methods VIII, IX and X. It is difficult to use method VIII, bootstrap, without good statistical software. In method IX, there is a possibility of having many patients with more length of stay than 95th percentile. There is a chance for increase in average value in this method. But Method X yields the lowest ALOS which indicates that generalized robust statistics method is better compared to bootstrap method and robust statistics method. Also the computations are simple which do not require sophisticated techniques or statistical packages. The numbers outside the range will be minimal. However, there is limitation in terms of generalizing the findings. The methodology was subjected to one year data of one large corporate hospital. Data from other health facilities, both from private and government need to be subjected to the methodology and validated.

Conclusion
This paper assesses statistical methods for the analysis of hospital length of stay. Commonly used methods produce varied results. However, Generalized Robust Statistics Method (GRSM) can be an objective yet simple technique for computing average length of stay.

References