Application of Epanechnikov kernel smoothing technique in disability data

Jumi Kalita*
Department of Statistics,
L.C. Bharali College,
781011 Guwahati, India
Email: jumikalita@yahoo.co.in
*Corresponding author

Pranita Sarmah
Department of Statistics,
Gauhati University,
781014 Guwahati, India
Email: pranitasarma@gmail.com

Abstract: Statistical data contains noise. Smoothing is used to smooth out these noises and present the data as a meaningful one. Kernel methods are nonparametric smoothing tools that can reveal structural features in the data which may not be possible with a parametric approach. This paper applies Epanechnikov kernel method of data smoothing to smooth out the dropout rates of the children with disabilities in the special educational institutions. The continuation probabilities and dropout rates of these children in the special educational institutions are indicators of effectiveness of such education systems. The dropout rates before and after smoothing are graphically presented. The distributions of the crude and smoothed rate are examined. It has been observed that under chi-squared test the smoothed data follows log logistic distribution while the crude data follows triangular distribution.

Keywords: dropout rate; continuation probability; smoothing; kernel smoothing; Epanechnikov kernel.

Reference to this paper should be made as follows: Kalita, J. and Sarmah, P. (2017) ‘Application of Epanechnikov kernel smoothing technique in disability data’, Int. J. Intelligent Systems Design and Computing, Vol. 1, Nos. 1/2, pp.198–204.

Biographical notes: Jumi Kalita is an Assistant Professor of Statistics in Lalit Chandra Bharali College, Guwahati, India. She has been teaching statistics to undergraduates for the last sixteen years. Her area of research for doctoral studies was statistical analysis of disability data of children. Her areas of interest are reliability, non-parametric statistics, data analysis, etc.

Pranita Sarmah is a Professor in the Department of Statistics in Gauhati University, where she has been engaged in post-graduate teaching and research for the last few decades. Her areas of specialisation are – reliability, stochastic process, Bayesian analysis, etc.
1 Introduction

Statistical data may contain noise in varying degrees and analysis of such noisy data may give misleading conclusions. Neurobiological data are a combination of the signals of interest and other signals from processes not directly related to the topics of interest. The presence of such noise can adversely affect the statistical analysis performed on the data (Albo et al., 2004). Smoothing is one of the basic tools in statistical applications to filter out noise contained in data. In statistics, to smooth a data set is to create an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures phenomena (Holčapek and Tichý, 2011).

Over the years, smoothing has been explicitly used by many researchers in diverse areas of study as a way of achieving noise-filtered data. Some of the many diverse applications are: reduction of scattering amongst data points from the signal measured by a motorised photo detector (Levesque, 2008), to attain the local adaptivity of a regression function together with a global regularisation using weighted smoothing splines (Sturdivant and Hosmer, 2007) in processing of human face images (Ganguly et al., 2014) etc.

In this paper, the drop-out rates of the children with disability in special schools are smoothed to provide better estimate of the rate, using Epanechnikov kernel smoothing technique.

Special education is offered to the children exhibiting high levels of physical, emotional, sensory or intellectual disability, which present a challenge to learning. Specially trained educators teach the children employing special methods at a pace suitable to the needs of these children. Data were collected from some such special schools of Assam, India and the drop-out rates of the children from these schools were determined using Nelson-Aalen estimator for cumulative hazard rate (Klein and Moeschberger, 1997). This gives a crude estimate of the hazard rate. In survival analysis, hazard rate plays an important role and is unique for all the distributions. These are used to understand the probabilistic pattern of an event (Kalita and Sarmah, 2012).

As Nelson-Aalen estimator provides only crude estimate, smoothing technique is used to get a better estimate of the same. In this context, the hazard rate ($\hat{H}(t)$) is referred to as the drop-out rate and $t$ is the age of the children in special institutions. Once smoothed estimates are obtained, the probabilistic structure of drop out time from the special institutions can be identified. This was expected to lead to proper analysis of drop out of the children along with reason behind dropping out, strength-weakness of the special education system and there from to arrive at corrective measures in the system if necessary.

2 Data source

For the study, 518 sets of data were collected from selected institutions (namely, NIPCCD, Monbikash Kendra, Shishu Sarathi, etc.) from Assam, India and categorised according to disability and need. These institutions were created especially for the children with disabilities to provide education to them.
3 Materials and method

Kernel methods are nonparametric smoothing tools and they can reveal structural features in the data which a parametric approach may not do. Structural features are based on the Nelson-Aalen estimator \( \hat{H}(t) \) and its variance. For right-censored and left-truncated data it gives better fit (Klein and Moeschberger, 1997).

The Nelson-Aalen estimator of the cumulative hazard rate is defined as

\[
\hat{H}(t) = \begin{cases} 
0 & \text{if } t \leq t_i \\
\sum_{i: t_i \leq t} \frac{d_i}{Y_i} & \text{if } t_i \leq t 
\end{cases}
\]

where
- \( d_i \) number of drop-outs at time \( t_i \)
- \( Y_i \) number of individuals who are at risk of leaving the institution at time \( t_i \)

The variance of Nelson-Aalen estimator which is due to Aalen is given by

\[
\sigma_{\hat{H}(t)}^2 = \sum_{i: t_i \leq t} \frac{d_i}{Y_i^2}
\]

The Kernel-smoothed estimator of \( h(t) \) (hazard rate) is a weighted average of the crude estimates over event times close to \( t \), the time point. The event times in the range \( t - b \) to \( t + b \) are included in the weighted average which measures \( h(t) \), where \( b \) is a bandwidth. The bandwidth is measured either to minimise some measure of the mean-squared error or to give a desired degree of smoothness. The weights are controlled by the choice of a Kernel function, \( K() \), defined on the interval \([-1, +1]\), which determines how much weights is given to points at a distance from \( t \). \( \Delta \hat{H}(t_i) \) provides a crude estimate of \( h(t) \).

Among the different kernels, with a minimum asymptotic mean squared error (AMISE), Epanechnikov kernel is optimal (Wand and Jones, 1995). Information on the children with disability is not available prior to time of detecting of the problem; it is reported after detection by concerned specialists. With this characteristic, data on children with disability become left truncated. At the same time it is right censored because, after a specific period, children left these schools and data on them become unavailable. For this reason, Epanechnikov kernel method of smoothing is selected for this study.

It gives progressively heavier weight to points close to \( t \) and it is defined as

\[
K(x) = 0.75\left(1 - x^2\right) \text{ for } -1 \leq x \leq 1
\]

Again the Kernel-smoothened estimator of \( h() \) based on the kernel \( K() \) for the time point \( t, b \leq t \leq t_d - b \), is given by

\[
\hat{h}(t) = b^{-1} \sum_{i=1}^{D} K\left(\frac{t-t_i}{b}\right) \Delta \hat{H}(t_i)
\]

and its variance is
Application of Epanechnikov kernel smoothing technique in disability data

\[ \sigma^2 \left[ \hat{h}(t) \right] = b^{-2} \sum_{i=1}^{D} K \left( \frac{t-t_i}{b} \right)^2 \Delta \hat{V} \left[ \hat{H}(t_i) \right] \]  

(5)

For \( t < b \) the above symmetric kernels become inapplicable, since no event times less than 0 are observable. For this modified kernels are used, which accounts for the restricted range of data. The modified kernel for Epanechnikov kernel is given as (Klein and Moeschberger, 1997):

\[ K_q(x) = K(x)(\alpha_E + \beta_E x), \quad q = \frac{t}{b}, \quad -1 \leq x \leq q \]  

(6)

where

\[ \alpha_E = \frac{64(2 - 4q + 6q^2 - 3q^3)}{(1+q)^2(19-18q+3q^2)} \]  

(7)

and

\[ \beta_E = \frac{240(1-q)^2}{(1+q)^2(19-18q+3q^2)} \]  

(8)

For \( t_D - b < t < t_D \), \( q = (t_D - t)/b \), the modified kernels stated above are used with \( x \) replaced by \((-x)\).

Along with this, the mean integrated square error (MISE) is defined as

\[ E = \left\| f_n - f \right\|^2 = E \int (f_n(x) - f(x))^2 \, dx \]  

(9)

where, \( f \) is the unknown density, \( f_n \) is its estimate based on a sample \( n \) independently and identically distributed random variables. \( E \) denotes the expected value with respect to that sample.

In this study, the probabilistic pattern of continuation probabilities of these children in special schools is obtained by using a software Easyfit. EasyFit is a data analysis and simulation application allowing to fit probability distributions to sample data, select the best model, and apply the analysis results to make better decisions. It can be used as a stand-alone Windows application. This requires the continuation probabilities corresponding to the drop-out rates. For this purpose, the expression of continuation probability \( C = \exp(-\hat{H}(t)) \) is used. This \( C \) is calculated for both crude and smoothed \( \hat{H}(t) \).

5 Tables and calculation

For smoothing, the appropriate value of \( b \) (bandwidth) is required and is obtained by minimising the MISE function given by

\[ g(b) = \sum_{i=1}^{M+1} \left[ \frac{\hat{h}^2(u_i) + \hat{h}\hat{2}(u_{i+1})}{2} \right] - 2b^{-1} \sum_{i=1}^{M} k \left( \frac{t_i}{b} \right) \Delta \hat{H}(t_i) \Delta \hat{H}(t_j) \]  

(10)
where, \( u_i \)'s are the grid points such that \( T_L = u_1 < u_2 < \ldots < u_M = T_U \).

This study covers the age of the children from 2 years to 14 years. So values of \( g(b) \) are compared in the range for \( b = 1, 2, 3 \) and 4 years and it is found that \( b = 2 \) years minimises the function \( g(b) \) (Figure 1).

Figure 1  Values of \( g(b) \) in the range of \( b = 1 \) to 4 (see online version for colours)

Table 1 along with the graph presents the smoothed dropout rates along with crude rates for children with disability with bandwidth 2.

<table>
<thead>
<tr>
<th>( t_i )</th>
<th>Crude dropout rate, ( \Delta \hat{H}(t_i) )</th>
<th>Smoothed dropout rate, ( \hat{h}(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.00579</td>
<td>0.00599</td>
</tr>
<tr>
<td>3</td>
<td>0.01359</td>
<td>0.01337</td>
</tr>
<tr>
<td>4</td>
<td>0.02362</td>
<td>0.01892</td>
</tr>
<tr>
<td>5</td>
<td>0.02218</td>
<td>0.02018</td>
</tr>
<tr>
<td>6</td>
<td>0.01856</td>
<td>0.02088</td>
</tr>
<tr>
<td>7</td>
<td>0.02731</td>
<td>0.02457</td>
</tr>
<tr>
<td>8</td>
<td>0.0324</td>
<td>0.03239</td>
</tr>
<tr>
<td>9</td>
<td>0.04464</td>
<td>0.04097</td>
</tr>
<tr>
<td>10</td>
<td>0.05374</td>
<td>0.05354</td>
</tr>
<tr>
<td>11</td>
<td>0.07407</td>
<td>0.06239</td>
</tr>
<tr>
<td>12</td>
<td>0.06933</td>
<td>0.0702</td>
</tr>
<tr>
<td>13</td>
<td>0.08309</td>
<td>0.07615</td>
</tr>
<tr>
<td>14</td>
<td>0.09063</td>
<td>0.0956</td>
</tr>
</tbody>
</table>

6 Discussion and conclusions

Figure 2 presents the comparative study between the crude and smoothed drop-out rates. Using the smoothed version of continuation probability, \( C = \exp(-\hat{H}(t)) \), corresponding to the drop out time for the children with disability, it is observed that the drop-out time follows the distributions mentioned in Table 2 along with their parameters.
Figure 2  Crude and smoothed drop-out rates for all types of children with disability (see online version for colours)

![Crude and smoothed drop-out rates](image)

Table 2  Distribution type and parameters before and after smoothing

<table>
<thead>
<tr>
<th>Test statistic used</th>
<th>Crude density</th>
<th>Smoothed density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distribution</td>
<td>parameters</td>
</tr>
<tr>
<td>Chi-squared</td>
<td>Triangular</td>
<td>$m = 0.00619$, $a = 0.00193$ and $b = 0.06491$</td>
</tr>
</tbody>
</table>

It is observed that the smoothed data follows log logistic distribution under chi-squared test, while the crude data follows Triangular distribution. The log-logistic distribution is a continuous probability distribution for a non-negative random variable. It is used in survival analysis as a parametric model for events whose rate increases initially and decreases later (Bennett, 1983). The log-logistic distribution provides the parametric form of the hazard where the mortality (failure) rate reaches a peak after some finite period and then slowly declines. The contribution made by a right censored observation to the likelihood is equal to the value of the continuation function at the time of censoring and this can be evaluated explicitly for the log-logistic distribution (Bennett, 1983). As smoothed data follows log-logistic distribution, and the data are also right censored survival data, we can infer that the smoothed distribution is appropriate for the data.

So, Epanechnikov kernel smoothing technique has successfully delivered the trends for drop-out rates in the case of children with disabilities in this study, by reducing noise and suppressing erratic data.

Acknowledgements

The authors acknowledge the help and cooperation received from the authorities and staffs of NIPCCD, Monbikash Kendra, and Shishu Sarathi at Guwahati, Assam, India while collecting data on children with disabilities.
References


