Improved frequency table with application to environmental data

ABSTRACT

This paper proposes three different statistics to be used to represent the magnitude observations in each class when estimating the statistical measures from the frequency table for continuous data. The existing frequency tables use the midpoint as the magnitude of observations in each class, which results in an error called grouping error. Using the midpoint is due to the assumption that the observations in each class are uniformly distributed and concentrated around their midpoint, which is not always valid. In this research, construction of the frequency tables using the three proposed statistics, the arithmetic mean, median, and midrange and midpoint are respectively named, Method 1, Method 2, Method 3, and the Existing method. The four methods are compared using root-mean-squared error (RMSE) by performing simulation studies using three distributions, normal, uniform, exponential distributions. The simulation results are validated using real data, Glasgow weather data. The findings indicated that using the arithmetic mean to represent the magnitude of observations in each class of the frequency table leads to minimal error relative to other statistics. It is followed by using the median, for data simulated from normal and exponential distributions, and using midrange for data simulated from uniform distribution. Meanwhile, in choosing the appropriate number of classes used in constructing the frequency tables, among seven different rules used, the freedman and Diaconis rule is the recommended rule.

Keyword: Frequency table; Statistical measures; Midpoint; Number of classes