Comput Stat (2012) 27:381–392
DOI 10.1007/s00180-011-0261-5

ORIGINAL PAPER

Boxplot for circular variables

Ali H. Abuzaid · Ibrahim B. Mohamed · Abdul G. Hussin

Received: 30 October 2008 / Accepted: 23 May 2011 / Published online: 23 June 2011

Abstract A boxplot is a simple and flexible graphical tool which has been widely used in exploratory data analysis. One of its main applications is to identify extreme values and outliers in a univariate data set. While the boxplot is useful for a real line data set, it is not suitable for a circular data set due to the fact that there is no natural ordering of circular observations. In this paper, we propose a boxplot version for a circular data set, called the circular boxplot. The problem of finding the appropriate circular boxplot criterion of the form $\nu \times CIQR$, where $CIQR$ is the circular interquartile range and ν is the resistant constant, is investigated through a simulation study. As might be expected, we find that the choice of ν depends on the value of the concentration parameter κ. Another simulation study is done to investigate the performance of the circular boxplot in detecting a single outlier. Our results show that the circular boxplot performs better when both the value of κ and the sample size are larger. We develop a visual display for the circular boxplot in S-Plus and illustrate its application using two real circular data sets.

Keywords Circular boxplot · Boxplot · Resistant constant · Outlier · Overlapping

1 Introduction

A visual display is a useful and an informative technique for describing a data set. It includes a histogram, a pie chart, a Q-Q plot and a boxplot. Tukey (1977) developed the boxplot, a simple and flexible graphical tool in exploratory data analysis. It entails

A. H. Abuzaid · I. B. Mohamed (✉)
Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: imohamed@um.edu.my

A. G. Hussin
Center for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia

the measurements of the smallest value, the lower quartile Q_1, the median ϕ, the upper quartile Q_3 and the largest value. One of its main applications is to identify extreme values and outliers in a univariate data set.

Extensive research has been conducted on the use of the boxplot in the labelling of outliers. To identify outliers in a real line data set, most studies use 1.5 as the value for the resistant constant ν in the boxplot criterion $\nu \times IQR$, where IQR is the interquartile range. In other words, any observations smaller than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$ are labelled as “outliers”. Hoaglin et al. (1986) investigated the performance of the boxplot for outlier labelling by considering different values of ν. The value $\nu = 1.5$ is considered to be the best choice in avoiding masking problems while $\nu = 3$ is considered to be extremely conservative. On the other hand, Ingelfinger et al. (1983) suggested the use of $\nu = 2$ while Sim et al. (2005) demonstrated that the choice of $\nu = 1.5$ or $\nu = 3$ was in general inappropriate for normal samples and was completely inappropriate for skewed distributions. The discussion above signifies the importance of choosing the most suitable value of ν for different data sets with different underlying distributions.

Our focus here has directions in 2-dimensions. In many diverse applications, data are measured in degrees or radians; for instance, wind directions and animal navigation. Such data are known as circular data. Fisher (1993) noted that circular plots have existed since 1858, when Florence Nightingale drew the plot of mortality data in the British Army during the Crimean War. Such a plot is also known as a rose diagram or wind rose diagram. In addition, Graedel (1977) used the boxplot to describe wind speed in different sectors of a wind rose diagram. However, in general, the boxplot is not directly applicable to a circular data set. In the literature, no special boxplot framework for a circular data set has been found. In this paper, we address the problem by proposing a boxplot for a circular data set which we will call a circular boxplot. We describe in detail the construction of this circular boxplot and develop subroutines in an S-Plus environment to display it. The circular boxplot can be used to detect outliers in circular samples. To date, there are several methods available to detect outliers in circular data sets. Collett (1980) presented four different numerical tests of discordance in circular data, namely the C, D and L statistics and an improved version of the M statistic originally proposed by Mardia (1975). Recently, Abuzaid et al. (2009) proposed the A statistic based on the summation of the circular distances from the point of interest to all other points while Abuzaid et al. (2008) used numerical and graphical tools to detect outliers in a circular regression model. However, the circular boxplot is simpler and more appealing compared to the other outlier detection techniques described above.

This paper is organized as follows. The next section discusses the proposed construction of the circular boxplot. Simulation and numerical studies are carried out in Sect. 3 to estimate the appropriate values of ν. Then, in Sect. 4, we investigate the power of performance of the circular boxplot for different values of ν and sample sizes. Two numerical examples are discussed in Sect. 5. The first example gives an application of the circular boxplot to the frogs’ directions data set as given in Ferguson et al. (1967), whereas the second utilizes the plot to identify outliers in a circular regression based on the resulting circular residuals discussed in Hussin et al. (2004).