Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 4000–4020
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Transformations for semi-continuous data

Galit Shmueli[a,b,*], Wolfgang Jank[a,b], Valerie Hyde[b,1]

[a] *Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, United States*

[b] *Applied Mathematics and Scientific Computation Program, University of Maryland, College Park, MD 20742, United States*

Received 12 December 2006; received in revised form 23 July 2007; accepted 22 January 2008
Available online 7 February 2008

Abstract

Semi-continuous data arise in many applications where naturally-continuous data become contaminated by the data generating mechanism. The resulting data contain several values that are "too frequent", and in that sense are a hybrid between discrete and continuous data. The main problem is that standard statistical methods, which are geared towards continuous or discrete data, cannot be applied adequately to semi-continuous data. We propose a new set of two transformations for semi-continuous data that "iron out" the too-frequent values thereby transforming the data to completely continuous. We show that the transformed data maintain the properties of the original data, but are suitable for standard analysis. The transformations and their performance are illustrated using simulated data and real auction data from the online auction site eBay.

1. Introduction and motivation

Standard statistical methods can be divided into methods for continuous data and those for discrete data. However, there are situations in which observed data do not fall in either category. In particular, we consider data that are inherently continuous but get contaminated by inhomogeneous discretizing mechanisms. Such data lose their basic continuous structure and instead are spotted with a set of "too-frequent" values. We call these semi-continuous data.

Semi-continuous data arise in various settings. Reasons range from human tendencies to enter rounded numbers or to report values that conform to given specifications (e.g., reporting quality levels that conform to specifications), to contamination mechanisms related to the data entry, processing, storage, or any other operation that introduces a discrete element into continuous data. Examples are numerous and span various applications. In accounting, for example, a method for detecting fraud in financial reports is to search for figures that are "too common", such as ending with 99 or being "too round". In quality control sometimes data are manipulated in order to achieve or meet certain criteria. For example, Bzik (2005) describes practices of data handling in the semiconductor industry that deteriorate the performance of statistical monitoring. One of these is replacing data with "physical limits": in reporting contamination levels negative values tend to be replaced with zeros and values above 100% are replaced with 100%. However,

* Corresponding author at: Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, United States.
E-mail address: gshmueli@rhsmith.umd.edu (G. Shmueli).

[1] Author names are listed in reverse-alphabetical order. This paper is part of the Doctoral Dissertation of the third author.

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.01.025

negative and >100% values are feasible due to measurement error, flawed calibration, etc. Another questionable practice is "rounding" actual measurements to values that are considered ideal in terms of specifications. This results in data that include multiple repetitions of one or more values. We will use the term "too frequent" to describe such values. We encountered two particular studies where semi-continuous data were present. The first is a research project by IBM on customer wallet estimation, where the marketing experts who derived the estimates tended to report round estimates thereby leading to data with several too-frequent values (Perlich and Rosset, 2006). A second study, which motivated this work, studies consumer surplus in the online marketplace eBay (www.eBay.com). Here, the observed surplus values had several too-frequent values, most likely due to the discrete set of bid increments that eBay uses and the tendency of users to place integer bids. The top panels in Fig. 1 show the frequencies of the values in samples from each of these two datasets. In the eBay surplus data (top panel) the value \$0 accounts for 6.35% of the values, and values such as \$0.01, \$0.50, \$1.00, \$1.01, \$1.50, \$2.00, \$2.01, \$3.00 are much more frequent than their neighboring values. In the IBM customer wallet estimates (middle panel) too-frequent values are 0, 100, 200, 250, 300, and 400.

We will use the surplus data throughout the paper to illustrate and motivate the proposed transformations. We therefore describe a few more details about the mechanism that generates the data. Consumer surplus, used by economists to measure consumer welfare, is the difference between the price paid and the amount that consumers are willing to pay for a good or service. In second-price auctions, such as those on the famous online marketplace eBay, the winner is the highest bidder; however, s/he pays a price equal to the second highest bid. Surplus in a second-price auction is defined (under some conditions) as the difference between the price paid and the highest bid. Bapna et al. (in press), who investigate consumer surplus in eBay, found that although surplus is inherently continuous, observed surplus data are contaminated by too-frequent values, as shown in the top left panel of Fig. 1.

The main problem with semi-continuous data is that they tend to be unfit for use with many standard statistical analysis methods.

Semi-continuous data can appear as if they come from a mixture of discrete and continuous populations. Graphs such as scatter plots and frequency plots might indicate segregated areas or clouds. Such challenges arise in the surplus data described above. Fig. 2 displays two probability plots for log(surplus + 1): the first is a lognormal fit and the second is a Weibull fit. It is obvious that neither of these two models approximate the data well (other distributions fit even worse) because there are too many 0 values in the data. Furthermore, using log(surplus + 1) as the response in a linear regression model yields residual plots that exhibit anomalies that suggest a mixture of populations. Fig. 3 shows two residual plots exhibiting such behavior; these plots indicate clear violation of the assumptions of a linear regression model.

One possible solution is to separate the data into continuous and discrete parts, to model each part separately, and then to integrate the models. For example, in a dataset that has too many zeros (zero-inflated) but otherwise positive, continuous values, we might create a classification model for zero/nonzero and then a prediction model for the positive data. This approach has two practical limitations: first, partitioning the data leads to loss of statistical power, and second, it requires the "too-frequent" values to be concentrated in a limited area or in some meaningful locations. To solve the first issue one might argue for a mixture model. Although there exists a plethora of such models for mixed continuous populations or for mixed discrete populations (e.g., "zero-inflated" models, Lambert, 1992), we have not encountered models for mixtures of continuous and discrete data. Moreover, there is a major conceptual distinction between mixture and semi-continuous data: unlike mixture data, semi-continuous data are inherently generated from a single process. Therefore treating them as a mixture of populations is artificial. The ideal solution, of course, would be to find the source of discretization in the data generation mechanism and to eliminate it or account for it. However, in many cases this is impossible, very costly, or very complicated. We therefore strive for a method that "unites" the apparent "sub-populations" so that the data can be integrated into a single model and treated with ordinary models for continuous data.

Our proposed solution is a set of two transformations which yield continuous data. We distinguish between two cases: one, where the goal is to obtain data that fit a particular continuous distribution (e.g., a normal distribution, for fitting a linear regression model); and two, where there is no particular parametric distribution in mind, but the data are still required to be continuous. The first approach, when the goal is simply to obtain continuous data, is based on jittering. As in graphical displays, where jittering is used to better see duplicate observations, our data transformation adds a random perturbation to each too-frequent value, thereby ironing out the anomalous high frequencies. We call this the *jittering transform*. The second approach, suitable when the data should fit a particular distribution, is based on