Data Transformation for Normalization

Amitava Mitra
Auburn University, USA

INTRODUCTION

As the abundance of collected data on products, processes and service-related operations continues to grow with technology that facilitates the ease of data collection, it becomes important to use the data adequately for decision making. The ultimate value of the data is realized once it can be used to derive information on product and process parameters and make appropriate inferences.

Inferential statistics, where information contained in a sample is used to make inferences on unknown but appropriate population parameters, has existed for quite some time (Mendenhall, Reinmuth, & Beaver, 1993; Kutner, Nachtsheim, & Neter, 2004). Applications of inferential statistics to a wide variety of fields exist (Dupont, 2002; Mitra, 2006; Riffenburgh, 2006).

In data mining, a judicious choice has to be made to extract observations from large databases and derive meaningful conclusions. Often, decision making using statistical analyses requires the assumption of normality. This chapter focuses on methods to transform variables, which may not necessarily be normal, to conform to normality.

BACKGROUND

With the normality assumption being used in many statistical inferential applications, it is appropriate to define the normal distribution, situations under which non-normality may arise, and concepts of data stratification that may lead to a better understanding and inference-making. Consequently, statistical procedures to test for normality are stated.

Normal Distribution

A continuous random variable, Y, is said to have a normal distribution, if its probability density function is given by the equation

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-(y-\mu)^2 / 2\sigma^2], \quad (1)$$

where μ and σ denote the mean and standard deviation, respectively, of the normal distribution. When plotted, equation (1) resembles a bell-shaped curve that is symmetric about the mean (μ). A cumulative distribution function (cdf), F(y), represents the probability $P[Y \leq y]$, and is found by integrating the density function given by equation (1) over the range $(-\infty, y)$. So, we have the cdf for a normal random variable as

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} \exp[-(x-\mu)^2 / 2\sigma^2]dx. \quad (2)$$

In general, $P[a \leq Y \leq b] = F(b) - F(a)$.

A standard normal random variable, Z, is obtained through a transformation of the original normal random variable, Y, as follows:

$$Z = (Y - \mu)/\sigma . \quad (3)$$

The standard normal variable has a mean of 0 and a standard deviation of 1 with its cumulative distribution function given by F(z).

Non-Normality of Data

Prior to analysis of data, careful consideration of the manner in which the data is collected is necessary. The following are some considerations that data analysts should explore as they deal with the challenge of whether the data satisfies the normality assumption.

Data Entry Errors

Depending on the manner in which data is collected and recorded, data entry errors may highly distort the distribution. For instance, a misplaced decimal point on an observation may lead that observation to become an outlier, on the low or the high side. **Outliers** are observations that are "very large" or "very small"

compared to the majority of the data points and have a significant impact on the **skewness** of the distribution. Extremely large observations will create a distribution that is right-skewed, whereas outliers on the lower side will create a negatively-skewed distribution. Both of these distributions, obviously, will deviate from normality. If outliers can be justified to be data entry errors, they can be deleted prior to subsequent analysis, which may lead the distribution of the remaining observations to conform to normality.

Grouping of Multiple Populations

Often times, the distribution of the data does not resemble any of the commonly used statistical distributions let alone normality. This may happen based on the nature of what data is collected and how it is grouped. Aggregating data that come from different populations into one dataset to analyze, and thus creating one "superficial" population, may not be conducive to statistical analysis where normality is an associated assumption. Consider, for example, the completion time of a certain task by operators who are chosen from three shifts in a plant. Suppose there are inherent differences between operators of the three shifts, whereas within a shift the performance of the operators is homogeneous. Looking at the aggregate data and testing for normality may not be the right approach. Here, we may use the concept of **data stratification** and subdivide the aggregate data into three groups or populations corresponding to each shift.

Parametric versus Nonparametric Tests

While data from many populations may not necessarily be normal, one approach for dealing with this problem, when conducting parametric tests, is to determine a suitable transformation such that the transformed variable satisfies the normality assumption. Alternatively, one may consider using nonparametric statistical tests (Conover, 1999; Daniel, 1990) for making inferences. The major advantage of nonparametric tests is that they do not make any assumption on the form of the distribution. Hence, such tests could be used for data that are not from normal distributions. There are some disadvantages to nonparametric tests however. One significant disadvantage deals with the **power** of the test. The power of a statistical test is its ability to identify and reject a null hypothesis when the null is false. If the assumptions associated with a parametric test are satisfied, the power of the parametric test is usually larger than that of its equivalent nonparametric test. This is the main reason for the preference of a parametric test over an equivalent nonparametric test.

Validation of Normality Assumption

Statistical procedures known as **goodness-of-fit tests** make use of the empirical cumulative distribution function (cdf) obtained from the sample versus the theoretical cumulative distribution function, based on the hypothesized distribution. Moreover, parameters of the hypothesized distribution may be specified or estimated from the data. The test statistic could be a function of the difference between the observed frequency, and the expected frequency, as determined on the basis of the distribution that is hypothesized. Goodness-of-fit tests may include chi-squared tests (Duncan, 1986), Kolmogorov-Smirnov tests (Massey, 1951), or the Anderson-Darling test (Stephens, 1974), among others. Along with such tests, graphical methods such as **probability plotting** may also be used.

Probability Plotting

In probability plotting, the sample observations are ranked in ascending order from smallest to largest. Thus, the observations $x_1, x_2, \ldots, x_n$ are ordered as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)}$ denotes the smallest observation and so forth. The **empirical cumulative distribution function (cdf)** of the ith ranked observation, $x_{(i)}$, is given by

$$F_i = \frac{i - 0.5}{n}. \tag{4}$$

The **theoretical cdf**, based on the hypothesized distribution, at $x_{(i)}$, is given by $G(x_{(i)})$, where $G(\cdot)$ is calculated using specified parameters or estimates from the sample. A probability plot displays the plot of $x_{(i)}$, on the horizontal axis, versus F_i and $G(x_{(i)})$ on the vertical axis. The **vertical axis** is **so scaled** such that if the data is from the hypothesized distribution, say normal, the plot of $x_{(i)}$ versus $G(x_{(i)})$ will be a straight line. Thus, departures of $F(\cdot)$ from $G(\cdot)$ are visually easy to detect. The closer the plotted values of $F(\cdot)$ are to the fitted line, $G(\cdot)$, the stronger the support for the null hypothesis. A test statistic is calculated where large