Chapter 2
Data Visualization and Data Summary

INTRODUCTION

Data visualization is an important concept in data mining. It enables you to understand the data, and to begin to formulate questions that you need to answer in order to make reasonable decisions. In traditional statistics, models and hypothesis tests provide proof; visualization is used to accompany the model in an attempt to explain it. In the data mining approach, visualization may provide essential information about the patterns in the data.

Patients in a clinic or geographic area are very heterogeneous. Therefore, the distribution of patient factors will not have a normal distribution. Generally, the distribution will have a heavy tail since every patient population will have those extreme patients who need extraordinary care; there will be more patients who need considerable care than those whose treatment can be discontinued early. Thus, unlike the normal distribution assumption, distributions of patient populations will not be symmetric. Therefore, great care must be used when considering a model that assumes normality, or even symmetry. We will look at alternative methods of analysis and visualization that do not depend upon the assumption of normality.

DOI: 10.4018/978-1-60566-752-2.ch002
In this chapter, we look at bar graphs of patient length of stay. As will be clearly demonstrated, these distributions are not normally distributed. Therefore, we will need some way to model these patient distributions. We will show a method, called kernel density estimation, that makes no assumption about the shape of the underlying population distribution; instead, it provides a way to define a model that can be used to examine the entire population instead of just its average.

Probability is usually defined in terms of a density function. The most common such density function is the bell-shaped curve. The Probability that a value \( x \) is less than some pre-determined value, \( X \) (\( P(x \leq X) = C \)), is equal to the integral between \( -\infty \) and the value \( X \) of the density function. In the past, tables were provided with the integral value of the standard normal density function. More recently, most statistical or graphing calculators provide the values as more and more mathematical tables are no longer printed. A brief summary of kernel density estimation is available online. (Anonymous-kde, 2008)

Kernel density estimation follows the same principal as integral calculus. The area under a curve can be estimated using a series of rectangles, with the total area assumed to be the sum of the areas of the rectangles. As the width of the rectangles becomes smaller, the sum of the rectangles comes closer and closer to the actual area under the curve. As the width of the rectangles approaches zero, the sum of rectangles approaches the area. There are a number of calculus formulas available that can be used to estimate the area under the curve. The simplest, but still relatively effective method is known as Simpson’s Rule. (Anonymous-simpson, 2008)

**BACKGROUND**

**Bar Graphs of Population Distributions**

We start with Figure 1, the bar graph of the hospital length of stay for all patients with diabetes. In the National Inpatient Sample, that includes just over 1 million patient stays. Note that the distribution has