Paper 176-26

The Visualization of Continuous Data Using PROC KDE and PROC CAPABILITY

George R. Barnes, Patricia B. Cerrito, Jewish Heart and Lung Institute, Department of Mathematics, University of Louisville, Louisville, Kentucky 40292

ABSTRACT

Kernel density estimation provides a very useful means of investigating entire populations. Traditionally, interest has focused on means and inference. This interest has lead to an unfortunate tendency to expect the entire population to be average. For example, if one medical treatment on average reduces the hospital length of stay by 10.40 versus 9.36 with a p-value 0.0235, the indication is that the treatment makes a full day difference in patient outcome. However, when the kernel density functions are examined, by shifting the curve representing Treatment B by 0.5 day, the two distributions become nearly identical. PROC KDE stores the data values so that such shifts can be identified graphically. PROC CAPABILITY will also make comparisons of treatments, but the graphics are displayed separately. Subpopulations can be examined to determine where probability density functions differ, and conditional density functions can be examined in detail using both the bivariate density capabilities of Proc KDE and also by examining the 2-dimensional cross sections. Several examples from a data mining analysis of open heart surgery demonstrating the potential of the two will be examined. The examples will demonstrate how visualization can provide essential information.

INTRODUCTION

Continuous data are usually visualized by histograms for one variable and by scatterplots for two variables. SAS/GRAPH has the capability for providing contour plots, surface plots, and 2-and 3-dimensional scatter plots. However, these methods give very rough pictures of the population distributions. It is possible to use kernel density estimators to smooth the distributions and give a good representation of the data. It is the purpose of this paper to demonstrate the use of PROC KDE and PROC CAPABILITY to find these estimates and to apply them to examine the relationships between variables.

Definition of Kernel Density Function

In many statistical tests, the assumption is made that the data sample is large or from a normally distributed population. However, that assumption is not reasonable unless the population is reasonably homogeneous. Unfortunately, most large populations are heterogeneous. In this case, the assumption of normality must be questioned. In the past, unless an inferential test proved that the population was not normal, the assumption of normality was used. In many instances, the assumption of normality was not questioned at all. In the past, there has not been a quick, reasonable way to estimate the underlying population distribution from the data. The empirical distribution function was not entirely satisfactory since it did not use all information available from the data. However, PROC CAPABILITY first incorporated the kernel density estimation into SAS/QC. With Version 7, PROC KDE became available in SAS/STAT. The two techniques greatly enhance the data visualization capabilities of SAS.

The kernel density estimate is defined by the equation:

$$\hat{f}(x) = \frac{1}{na_n}\sum_{j=1}^{n} K\left(\frac{x - X_j}{a_n}\right)$$

where n is the sample size, K is a known density function, and a_n is a constant depending upon the size of the sample that controls the amount of smoothing in the estimate. Note that for most standard density functions K, where x is far in magnitude from any point X_j, the value of $K\left(\frac{x - X_j}{a_n}\right)$ will be very small. Where many data points cluster together, the value of the density function will be high because the sum of $x\text{-}X_j$ will be small and the probability defined by the kernel function will be large. However, where there are only scattered points, the value will be small. K can be the standard normal density, or the uniform density. Simulation studies have demonstrated that the value of K has very limited impact on the value of the density estimate. It is the value of the bandwidth, a_n, that has substantial impact on the value of the density estimate. The true value of this bandwidth must be estimated, and there are several methods available to optimize this estimate. PROC KDE differs from PROC CAPABILITY in the algorithms used to estimate the bandwidth.

PROC KDE uses only the standard normal density for K but allows for several different methods to estimate the bandwidth, as discussed below. The default for the univariate smoothing is that of Sheather-Jones plug in (SJPI):

$$h = C_3\left\{\int f''(x)^2\,dx, \int f'''(x)^2\,dx\right\}C_4(K)h^{5/7}$$

where C_3 and C_4 are appropriate functionals. The unknown values depending upon the density function f(x) are estimated with bandwidths chosen by reference to a parametric family such as the Gaussian as provided in Silverman:

$$\int f''(x)^2 dx = \sigma^{-5}\int \phi''(x)^2 dx \approx 0.212\sigma^{-5}$$

However, the procedure uses a different estimator, the simple normal reference (SNR), as the default for the bivariate estimator:

$$h = \hat{\sigma}\left[4/(3n)\right]^{1/5}$$

along with Silverman's rule of thumb (SROT):

$$h = 0.9\min[\hat{\sigma}, (Q_1 - Q_3)/1.34]n^{-1/5}$$

and the oversmoothed method (OS):

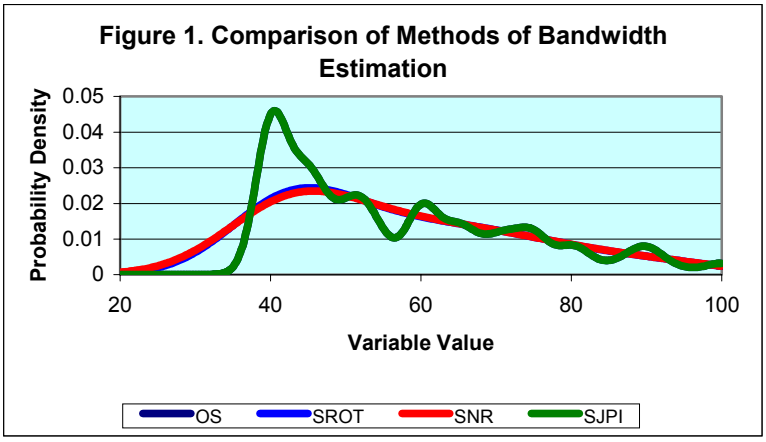
$$h = 3\hat{\sigma}\left[1/70\sqrt{\pi}n\right]^{1/5}$$

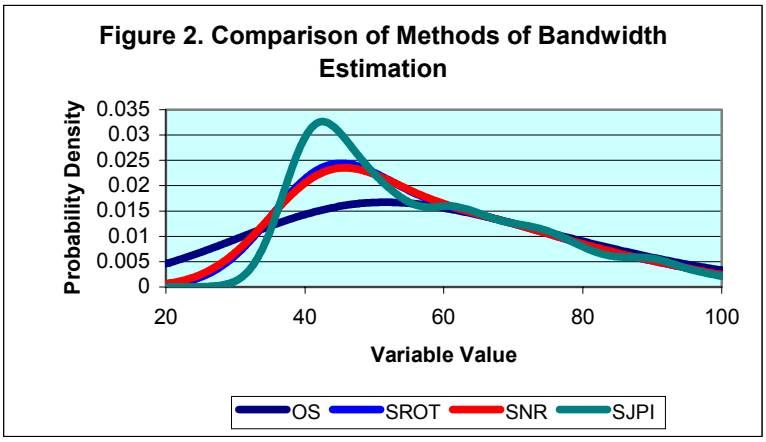
The basic code for PROC KDE is as follows:

```
proc kde data=work.cabg gridl=1 gridu=20
method=srot out=outkde;
var packyears;
run;
```

PROC KDE will accept one or two variables depending upon whether the density is univariate or bivariate. The density estimate is stored in an output file containing the values of x from gridl (minimum) to gridu(maximum). The default number of datapoints is 401. This can be changed with the option ng=*number*. In addition, the default bandwidth estimate can be altered with the option bwm=*number*. The number is a multiple of the default bandwidth. A comparison of the four methods is provided in Figure 1.

It can be noted that OS, SROT, and SNR OS it not clearly visible since the other two curves are so close) give nearly identical estimates whereas the default is more jagged. The bandwidth can be smoothed when compared to the estimates given in Figure 1 (as they are rather jagged) as shown in Figure 2. The different





impact on the different methods of estimation is very different. The SJPI becomes considerable smoother than previously in Figure 1. In contrast, PROC CAPABILITY is more limited. There are several choices for the density function K but the choice for the bandwidth is restricted to the mean integrated squared error or a constant value:

$$h_{MISE} = \left\{\int t^2 K(t)dt\right\}^{-2/5}\left\{\int K(t)^2\right\}^{1/5}\left\{\int f''(x)^2 dx\right\}^{-1/5} n^{-1/5}$$

The basic codes is equal to

```
proc capability data= WORK.LUNG;
var PACKYEARS;
HISTOGRAM /kernel( k=NORMAL c=MISE
color=BLUE l=1)cfill=GRAY;
run;
```

Note that the procedure sketches the histogram and the kernel density (Figure 3) but does not save the output so that the graph can be manipulated. It is more difficult to limit the points of the kernel density function so that outliers are usually given weight in the curve. However, note that PROC CAPABILITY generally does not have as much probability weight in the tails as do the outcomes using PROC KDE.

Application to Medical Data
PROC KDE

The use of kernel density estimation can greatly enhance a statistical investigation of observational data. Consider a database with 2150 observations and 171 variables. The primary objective of the analysis is to determine whether a relatively new type of surgical technique can enhance patient outcomes. To date, there are no blinded, randomized, controlled trials to examine new surgical techniques. Because of the issue of informed consent, the patient needs to know the surgical procedure that will be performed. There is alsoa possibility of bias in the choice of patients who undergo the new technique. In this particular example, patients received open heart surgery without the use of the bypass machine. The procedure was performed on a beating heart. Data mining procedures were used to examine the interrelationships between variables. An attempt was made to case-match patients examined by pre-existing conditions such as diabetes, obesity, and renal failure. In addition, an attempt was made to match the severity of the problem. The study presented here will focus on an examination of the relationship of smoking variables to patient outcomes. Several analyses of variance were performed. Patients were asked if they were current smokers. They were also asked to estimate their number of pack years (1 pack/day x 1year=1 pack year). The outcome variable examined was the length of hospital stay from admission to discharge. The number of pack years was dichotomized at the value of 50 pack years.

In the initial ANOVA, only the main effects were examined. All three variables were statistically significant (number of pack years, p=0.0006; current smoking, p=0.0481; surgical technique, p<0.0001). The least squares means indicated that there was over a two-day difference between the two techniques. It must be remembered that at this point all possible biases had net yet been examined. Also, because of the large sample size, it is possible that any variable could be statistically significant. Therefore, it is useful to examine the probability density functions. Consider the comparisons for each of the two smoking variables (Figure 4). The two distributions indicate that the p-values may be inflated.

Figure 3. Kernel Density Estimator Using PROC CAPABILITY

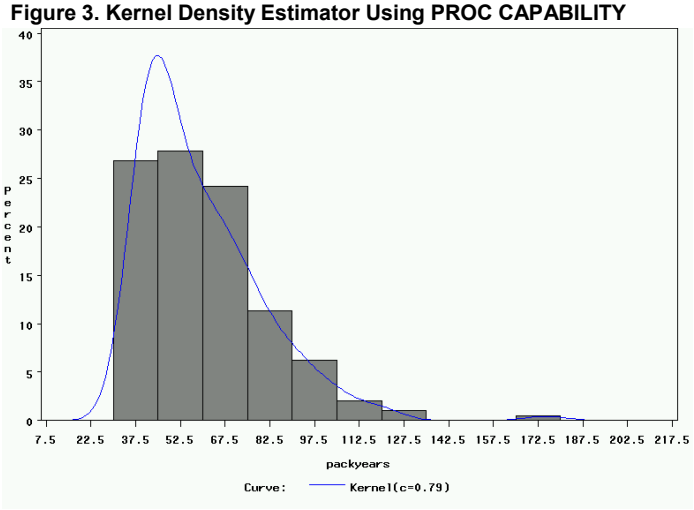
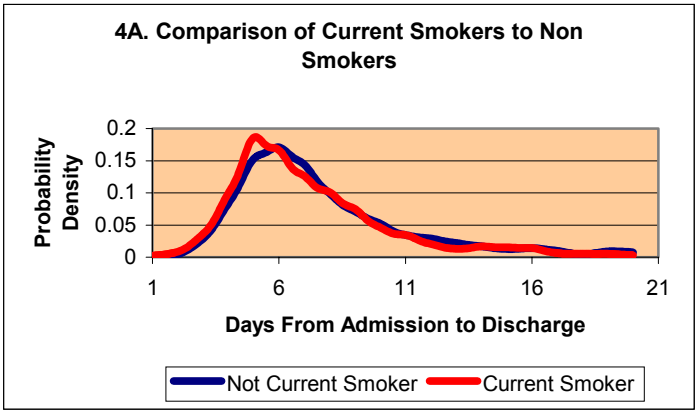


Figure 4. Comparison of Smoking Variable versus Quantity of Smoking Variable



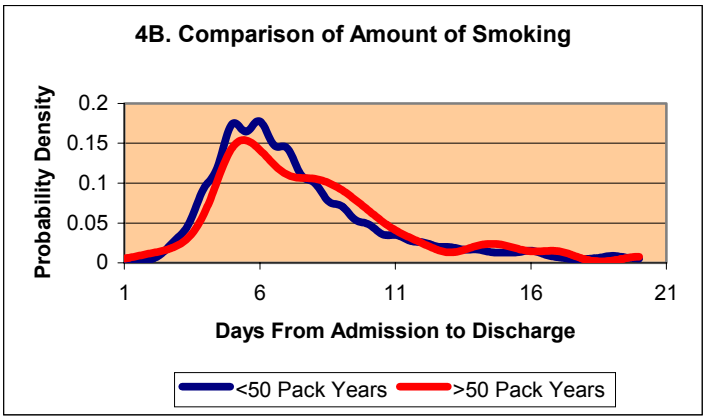


Figure 5 indicates the difference in density functions for the two surgical procedures. At first glance, the two curves look very different. However, the two subpopulations are very different in size. The standard technique was performed on 1912 patients. The SAS code:

```
proc kde data=work.cabg gridl=1 gridu=20
out=outkde;
var losadmitdischarge;
by  technique;
run;
```

cannot always optimize the bandwidth for both subsamples. Therefore, it is possible to use a different method to estimate the bandwidth and improve upon the curve:

```
proc kde data=work.cabg gridl=1 gridu=20
method=srot  out=outkde2;
var losadmitdischarge;
by  cardioplegia;
run;
```

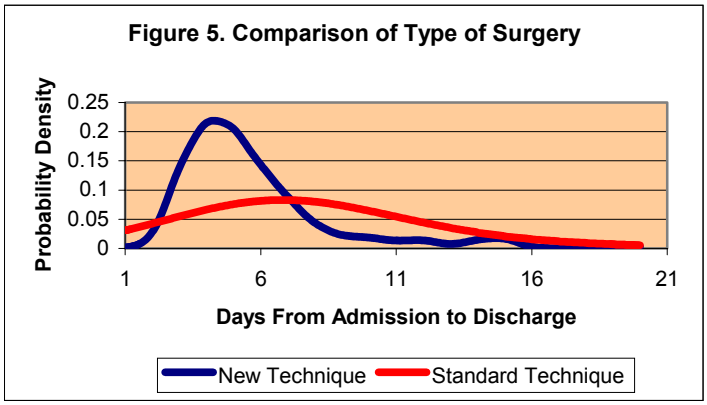
The result of the above procedure is given in Figure 6. The separation between the two curves indicates that the newer procedure does reduce the length of hospital stay. However, it is not yet know whether smokers (or heavy smokers) were offered this new technique. Therefore, the two way interactions were also examined using both ANOVA and kernel density estimation. Then the following p-values from the Type III SS are given in Table 1.

Note that the surgical procedure ceases to be statistically significant. However, when examining the density curves comparing the surgical technique to smoking, the separation between the curves remains clear. Therefore, kernel density estimators can find shifts in probability density that cannot be detected by comparing the averages of two populations. Similarly, the interaction of

Table 1. Type III SS P-Values

Variable	*P-Value*
Surgical Technique	0.8896
Smoking	0.0004
Number of Pack Years	0.0607
Smoking*Surgical Technique	0.6942
Number of Pack Years * Surgical Technique	0.3772
Number of Pack Years * Smoking	<0.0001

smoking with surgical technique is not statistically significant. However, as demonstrated in Figure 7, there remains a clear split between the new surgical technique and the more standard technique in terms of hospital length of stay. The values that are statistically significant are smoking and the number of pack years. Therefore, both variables were also considered with respect to the type of surgery (Figure 8). As shown in the graph, there is a difference between the two techniques. Patients with a lifetime of more than 50 pack years of smoking have a considerable increase in the overall length of stay with the new surgical technique. Similarly, with less than 50 pack years, patients are much more likely to have a shorter



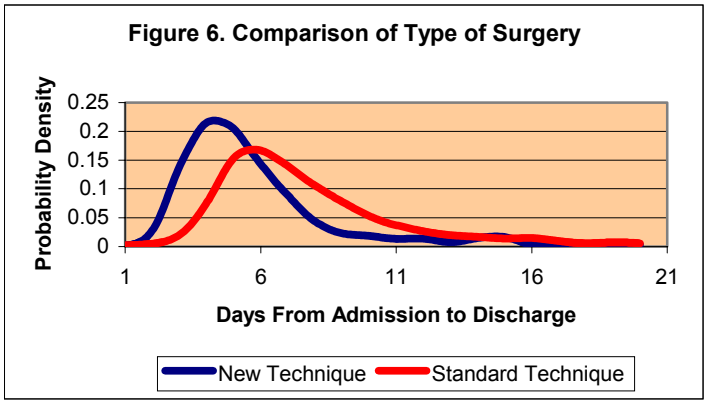


Figure 7. Comparison of Smoking and Type of Surgery

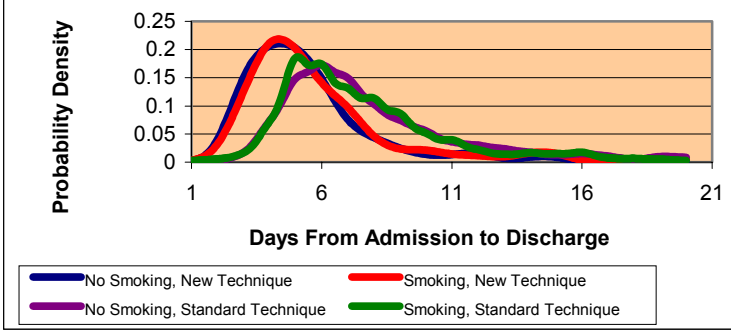
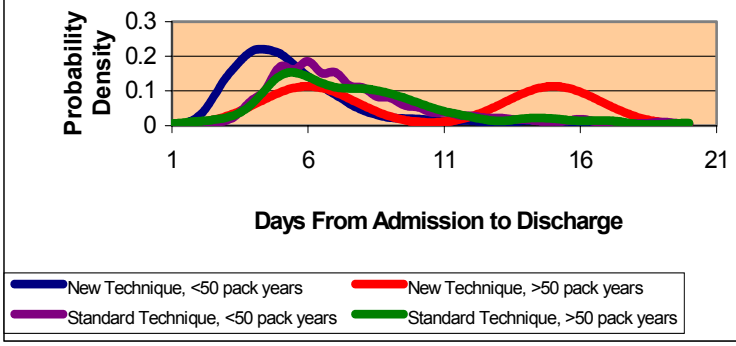


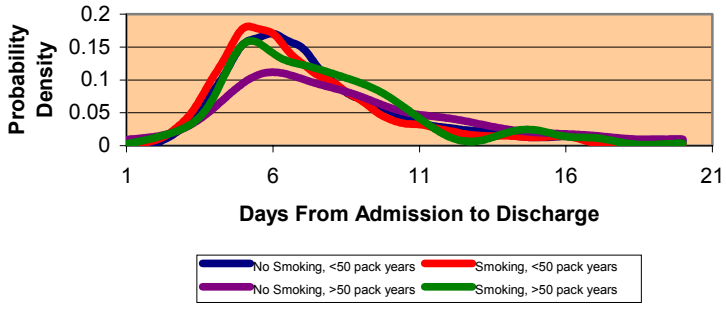
Figure 8. Comparison of Amount of Smoking and Type of Surgery



length of stay using the new technique. Therefore, it can be conjectured that moderate smokers reduce costs under the new technique while severe smokers should use the more traditional technique.

With the newer technique and more than 50 pack years, there is a true bimodality to the probability density. Figure 9 demonstrates that the likelihood of being discharged at day 6 is 1.5 times more likely for smokers with less than 50 pack years than non-smokers with more than 50 pack years.

Figure 9. Comparison of Amount of Smoking and Current Smoking



Bandwidth Esimation

One of the primary difficulties in using the kernel density estimator is optimizing the value of the bandwidth. Although there are several methods available in PROC KDE as discussed previously, there needs to be some interaction with the investigator. Therefore, the following option can be added to the basic KDE code:

```
proc kde data=work.cabg gridl=1 gridu=20
method=srot out=outkde bwm=2;
var packyears;
run;
```

The bwm stands for bandwidth multiplier. The optimal bandwidth determined by SAS can be multiplied by the given factor, in this case two. Figure 10 gives an example with the default bandwidth along with different multipliers. The default bandwidth appears to be too large with a curve that is over-smoothed. As the bandwidth decreases, the tail probability also decreases. In the example in Figure 10, it is suggested that either 1/2 or 1/5 of the optimal bandwidth yields a better estimator.

Curves that are under-smoothed generally cannot be used to predict outcomes using new data. Curves that are over-smoothed will tend to misrepresent the probability. Therefore, it is up to the investigator to examine the results that seem inherently reasonable.

PROC CAPABILITY

There is not as much flexibility with PROC CAPABILITY as there is with PROC KDE. For example, consider the resulting curve using the same data as that used in Figures 5 and 6. The SAS code used is as follows:

```
proc capability data= WORK.CABG
var LOSADMITDISCHARGE;
COMPHIST /kernel( k=NORMAL c=MISE )
class=Technique nrows=2 ncols=1;
run;
```

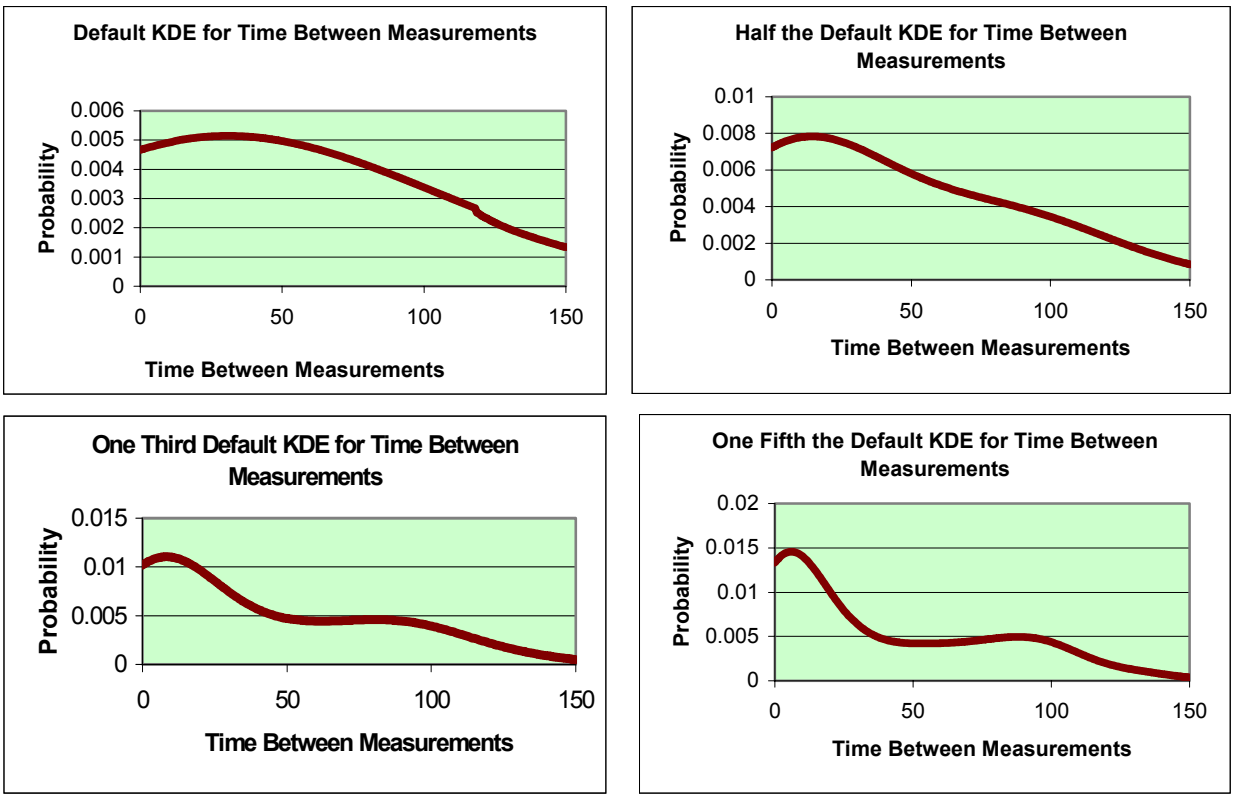
with the result provided in Figure 11. In this case, the default curve has very heavy tails. It is possible to limit this tail by using the WHERE statement:

```
proc capability data= WORK.CABG
var LOSADMITDISCHARGE;
COMPHIST /kernel( k=NORMAL c=MISE )
class=Technique nrows=2 ncols=1;
Where LOSADMITDISCHARGE le 20;
run;
```

The result is given in Figure 12.

The ease by which PROC KDE and PROC CAPABILITY can be used in SAS can greatly contribute to the ability to estimate and visualize continuous data. These procedures are particularly relevant when examining large, heterogeneous populations where it is not reasonable to assume that

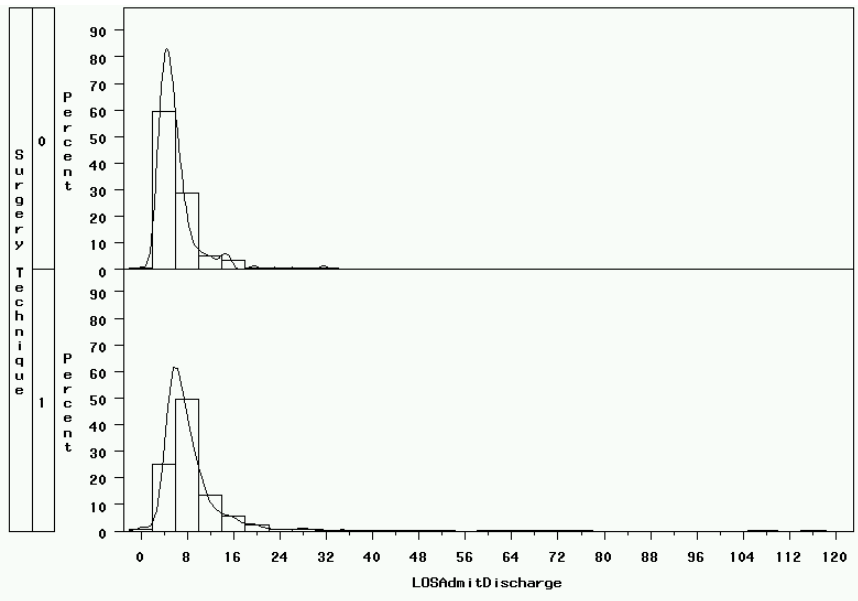
Figure 10. Different Bandwidths for the Same Data Using PROC KDE.



the underlying probability distribution is normal. These procedures are also more appealing than histograms as the kernel density procedure smoothes out the rough edges of the histogram.

By using conditional probability and bivariate density estimates, it is possible to explore in-depth the relationships between variables. The kernel density can provide much more information than can be gained from the use of the general linear model to compare differences in means.

Figure 11. Kernel Density Comparison Using PROC CAPABILITY



Confidence Limits

The kernel density gives the value of the probability density function. The estimated integral of the kernel density will provide an exact estimate of the probability at any point x. Through the use of any standard method from calculus for numerical integration, it is possible to estimate the probability. In particular, it is possible to examine upper and lower confidence limits for the distribution To find a 95% confidence limit around a point x, it is necessary to integrate the value of the following integral, solving for x_o. Using

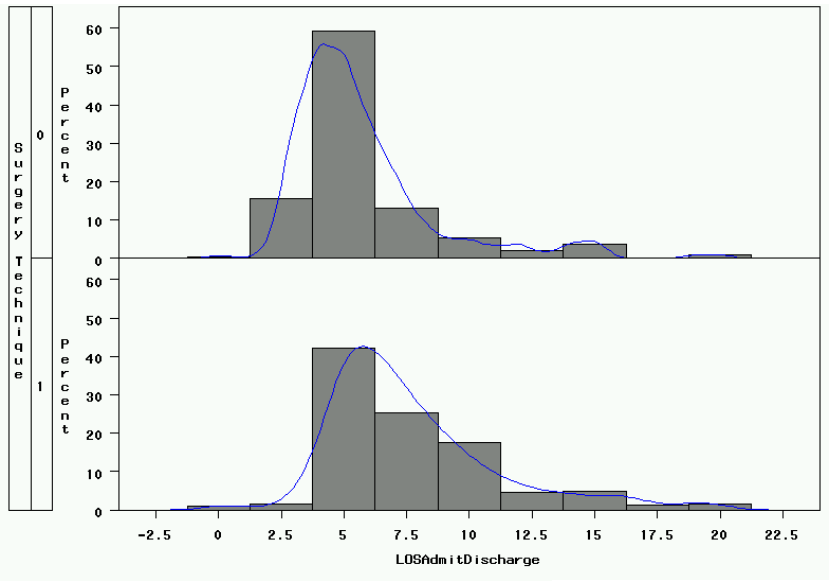
$$\int_{x-x_0}^{x+x_0} f(x)dx = 0.95$$

Simpson's Rule, the value of the integral is equal to

$$F(x_0) = \frac{\Delta x}{3}\left[f(x-x_0) + \sum_{i=1}^{n-1} f(x_i) + f(x+x_0) \right]$$

where $x_1,\ldots,x_{n-1}$ consists of all non-inclusive points computed using PROC KDE that are between x-xo and $x+x_o$. The value Δx is equal to the distance between any two com-

Figure 12. PROC CAPABILITY Result Corresponding to Figure 6.



puted points. The value x_0 can be found by computing F (x_0) for all values of x_0 computed in PROC KDE. These probabilities can be computed more accurately than estimating confidence limits by using the discrete values in a histogram.

Bivariate Density Estimation

Another option available in PROC KDE is to estimate a bivariate density function. Since the number of pack years was initially collected as a continuous variable, it is possible to compare the pack years more directly to the length of stay using the code (Figure 13). The following code is used:

```
proc kde data=work.cabg gridl=20,1
gridu=100,20 ngrid=400,400 out=outkde1;
var pack_years losadmitdischarge;
run;
```

Note that the number of patients with pack years larger than 50 begins to decline substantially. Similarly, there is a substantial drop in the number of days to discharge beyond 10 (Figure 14).

Discussion

It is useful to estimate probability distributions as many large, non-homogeneous populations will exhibit a multi-modal distribution. It is also useful to compare populations, providing information concerning shifts in probability rather than restricting attention to averages. The kernel density estimator is extremely flexible in providing visual insight to large populations. It does have limitations in that it does not work well with small samples and it relies on optimizing the bandwidth.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support of the Jewish Heart and Lung Institute, Louisville, Kentucky 402022 for providing the data (in accordance with IRB approval) and for partially funding this project.

CONTACT INFORMATION

George R. Barnes, Patricia B. Cerrito
Department of Mathematics
University of Louisville
Louisville, Kentucky 40292
502-852-6826
Fax: 502-852-7132
Email george.barnes@louisville.edu, pcerrito@louisville.edu
Web:www.math.louisville.edu

TRADEMARK CITATION

Figure 13. Bivariate Density Estimate Using PROC KDE Comparing Pack Years to Length of Stay

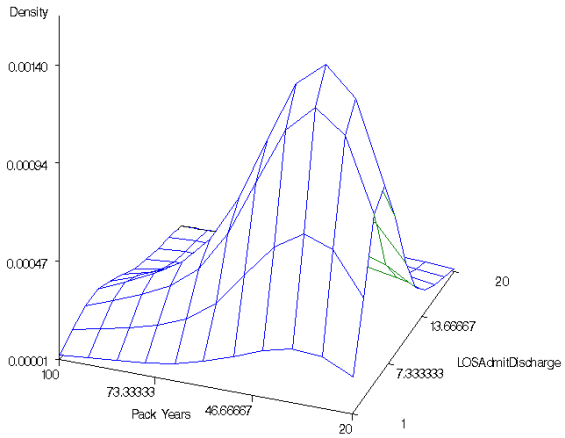


Figure 14. Bivariate Density Estimator Rotated