Paper 273-25

The Use of Kernel Density Estimators to Monitor Protocol Compliance

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

ABSTRACT

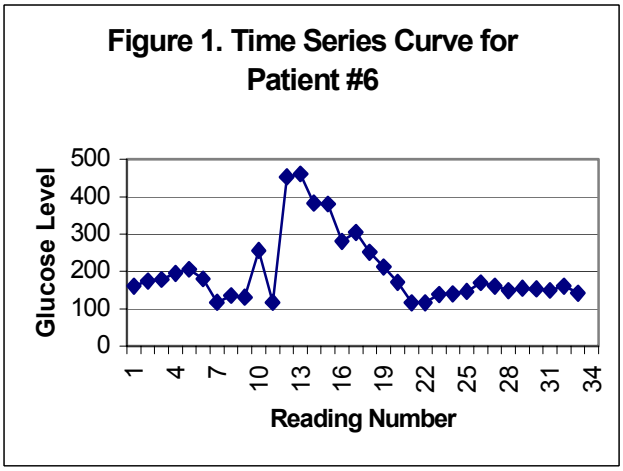
The use of kernel density estimation to develop instrument panels will be demonstrated in the continuous monitoring of medical outcomes. Kernel density functions can be quickly derived using SAS/QC®, and can be used to examine conditional probabilities. In version 7, SAS/STAT® contains PROC KDE that can also be used to estimate the kernel density. Specifically, kernel density functions were used to determine whether a new protocol was being implemented properly and whether the intervention had impact upon the patients it was intended to serve. The purpose of this protocol was to systematically monitor and control glucose levels in diabetic patients undergoing open heart surgery. The motivation for the protocol was to reduce infection rates, and average length of stay. Kernel density was used to determine the distributions of glucose levels, and to determine whether protocol interventions at high patient glucose levels resulted in a subsequent reduction. After analysis, it was concluded that the protocol was being properly implemented and that glucose levels were more likely to decrease once they climbed above 150 mg/dL. Initial results also indicated a reduction in the infection rate for diabetic patients.

INTRODUCTION

The protocol under investigation was designed to reduce infection rates in open heart surgery for diabetic patients. It was initiated in January, 1999 and subsequently modified in March and July. In January, all patients who were known diabetic patients were placed on the protocol. In March, all patients with an initial glucose reading greater than 180 were added to the protocol. In July, the protocol was extended to three days past surgery. As the requirements for monitoring and intervention were quite stringent, there was an initial question of protocol compliance that had to be answered prior to any examination of the impact on infection rates. It was successfully demonstrated that hospital personnel were complying with protocol.

The protocol called for the patient's glucose levels to be monitored every hour before and after surgery, and every 30 minutes during surgery. The monitoring began at 5:30 am regardless of the scheduled time of surgery. As soon as the glucose level climbed above 150 mg/dL, a combination insulin-glucose drip was initiated. The amount of insulin and glucose was determined by the level of blood glucose. The amount increased by preset increments if the glucose level continued to increase; it was decreased once the glucose level peaked and began to decrease. Complicating the issue is the fact that glucose levels routinely climb above 150 mg/dL during the use of the heart-lung bypass machine. Therefore, patients identified as diabetic were routinely put on an insulin-glucose IV drip. After postoperative recovery, the patient was moved to the nursing floor and the monitoring continued every hour until the patient was stabilized and removed from IV nourishment. At that point, glucose was monitored by stick approximately every 4 hours until discharge.

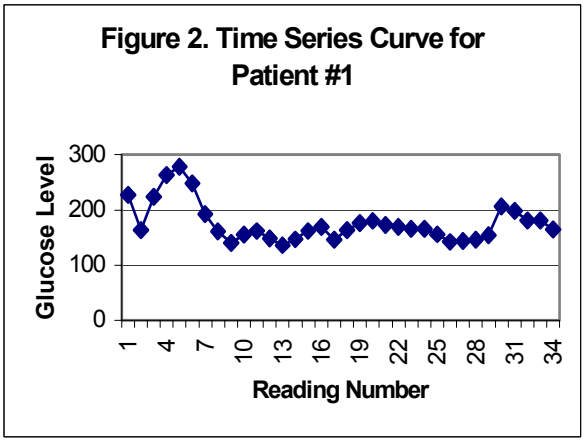
In March, the protocol was expanded to include all patients who might be diabetic even if not diagnosed. All patients were tested prior to surgery, and if the sugar level was high (ie above 150 mg/dL), patients were also placed on the protocol. In July, patients were continued on the IV insulin-glucose drip for 3 days after surgery with continued hourly monitoring.



Data Collection

PROTOCOL COMPLIANCE

Initial examination of the data took place after 60 patients had completed the protocol. The data were collected from several sources. A general database was used to gather information concerning surgery: types of procedures, length of surgery, complications, etc. A second database contained information about infection rates. The third database used in the protocol contained the monitored glucose levels. The three databases had to be merged by patient identification number prior to analysis. The glucose levels for each patient are recorded as a time series as shown in Figure 1.



It is somewhat questionable whether the use of averages could be used to examine the protocol since it was primarily concerned with high levels of glucose.

Contrast the time series of patient#6 with that of patient#1 (Figure 2). Patient #1 has an initial spike that quickly returns to normal range while patient #6 has a serious and long-lasting spike. According to the protocol, patient #6 requires much more aggressive intervention than patient #1. The data were complicated in that

the insulin readings were not equally spaced. While repeated measures analysis of variance using PROC MIXED could be used to analysis the data, it cannot provide the depth of detail that can be found using kernel density estimation. Therefore, the kernel density was used to examine the data.

KERNEL DENSITY ESTIMATION

Until recently, kernel desntiy estimators have not generally been used in clinical analysis. They are difficult to compute and they are even more difficult to use in statistical inference. However, with their introduction into SAS/QC, kernel density estimators have a place in clinical research. More recently, PROC KDE has been added to the SAS/Stat module in SAS. Although inference is still not part of the package, it is possible to compute inferential statistics. However, in examining an ongoing, non-randomized protocol where inference is of less importance than the distributions themselves, kernel estimators can provide a wealth of information.

One of the major drawbacks in the use of analysis of variance and logistic regression is the required assumption that data values are independent. Repeated measures analysis of variance can use time-dependent data but only if the time interval is categorical and fixed. Kernel density estimation is just as effective with weakly dependent data as it is with independent data. More necessary is the assumption that the data are from stationary distributions. The kernel density estimator in one dimension is defined by

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

where K(.) is a known density function and a_n is a constant called the bandwidth or window width. The bandwidth depends upon n since $a_n \to 0$ as $n \to \infty$. There are additional conditions on the bandwidth as well (Silverman, 1986). The most difficult part of using kernel density estimation is the estimation of this bandwidth. Therefore, as the sample size increases, the optimal choice of the bandwidth must decrease. It is known that the choice of K does not significantly alter the estimator f(x). The optimal bandwidth to reduce the mean integrated squared error is equal to
This cannot be computed because it depends on the unknown

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

distribution of the density function f(x). However, There are many different approaches to estimating the bandwidth (Silverman, 1986; Jones, Marron, and Sheather, 1996). The SAS/QC histogram procedure allows for several choices of the kernel density function including the triangular or quadratic kernel.

PROC KDE uses the Gaussian kernel only. There are four bandwidth estimator functions available in PROC KDE. 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 also uses the simple normal reference (SNR), 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):

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

In addition to these four methods for estimating the bandwidth, Jones, Marron, and Sheather (1996) compared the method of cross-validation. The authors of this paper demonstrated through empirical example that the SJPI method provides a better estimate. It is further demonstrated in Altman and Leger (1995) that cross-validation is not really practical in kernel density estimation, particularly with large samples.

It is, however, suggested that the SJPI method be employed and then examined by using multiples of the SJPI to avoid both over– and under-smoothing of the bandwidth. In this way, the SJPI can be adjusted depending upon the needs of the problem. It is the practical experience of the authors that the computed optimal bandwidth, regardless of the method used, has a tendency to oversmooth the result, particularly when the sample size is large. This can be done instead of comparing outcomes using the four different methods as provided by PROC KDE.

It is also suggested that the outcome from PROC KDE be compared to that of PROC CAPABILITY. It is the experience of the authors that PROC CAPABILITY does not have as great a tendency to over-smooth. It also provides a method to adjust the bandwidth. In addition, PROC CAPABILITY can super-impose a known distribution function as well as the histogram of the data. Therefore, the kernel density estimator can be compared to known distributions.

In the context of this problem, data for one patient and in one patient environment are weakly dependent. However, as the protocol is constant over the time interval under examination, stationarity can be assumed. The kernel density function can be applied and it is useful to examine the compliance with the protocol and its impact using kernel density estimators.

Since there are multiple time units to examine for each individual patient, and the time units are not constant across any patient group, an initial examination of the distribution requires the use of a method than can use weakly dependent time data. It has been clearly demonstrated that the kernel density estimator is just as effective with weakly dependent data as it is for independent data (Masry, 1987). In this particular data set, the raw glucose values are not of as great importance as the difference between time *t* and time *t+1*. Therefore, the difference was computed and the distribution determined using kernel density estimation. Diabetic patients on the protocol were compared to a cohort of diabetic patients with glucose values recorded prior to the implementation of the protocol (Figure 3). There were approximately 60 patients on protocol; 150 patients were not on protocol. Only differences where the initial value at time t

exceeded 150 mg/dL were examined for the distribution. Therefore, the conditional probability distribution (f(x), y>150) was estimated where x is the glucose level at time t=1 and y is the glucose level at time t=0.

The SAS code for this procedure is

```
goptions reset=(axis, legend, pattern,
symbol, title, footnote) htext= ftext=
ctext= target=  gaccess= gsfmode= inter-
pol= ;
goptions device=WIN     graphrc;
proc capability data= WORK.COMPARE;
var DIFF;
COMPHIST /
kernel( k=NORMAL c=MISE  )
class=Protocol
nrows=2
ncols=1;
Where Glucose>150;
run;
```

Note that PROC CAPABILITY has an automatic routine for drawing the kernel density function superimposed over the histogram of the data.

The kernel density function from PROC CAPABILITY is contrasted with that from PROC KDE. Using the BY command in the procedure permits comparisons. For this graph, the bandwidth estimator yielded an estimator that was too rough. The BMW option (to examine multiples of the default bandwidth) was exercised and the bandwidth doubled to yield the results from Figure 4.

There is a default of 401 points for a univariate kernel density estimator; a default of a 60 by 60 grid for the bivariate case. There is an option available to change these values. Also, outliers in the data tend to stretch these points beyond what is necessary for a reasonable estimator. The code for the graphs in Figure 4 is equal to

```
proc sort data=work.compare;
by protocol;
run;
proc kde data=work.compare gridl=-125
gridu=125 bwm=2 out=outkde;
var diff;
where glucose>150;
by Protocol;
run;
```

The density points are saved in the file outkde and can be used in a scatterplot. The data must be sorted if the BY command is used in PROC KDE.

Note that in both curves, the modal class is approximately 0. However, in the patient group not on protocol, the mode via kernel density estimation is slightly higher. A comparison of means could not be performed because of the weakly dependent nature of the data. There is a definite shift in the glucose levels with more values below 0 on protocol. The results are even more dramatic when the data are restricted to values where the initial time point exceeds 200 (Figure 5).

Figure 3. Comparison of Protocol Versus Non-Protocol Patients and Glucose Levels

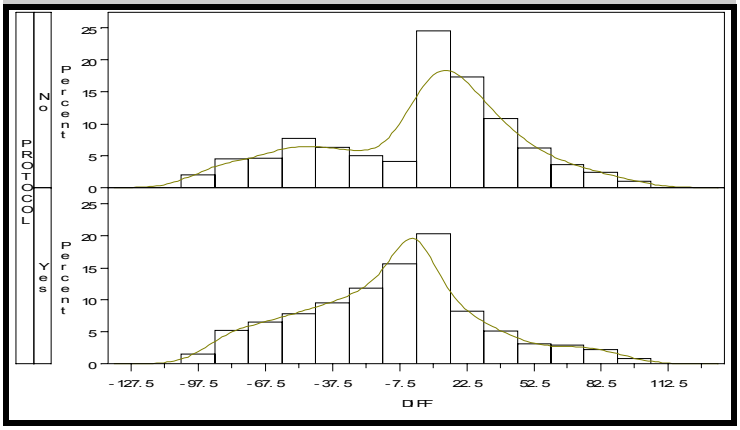


Figure 4. PROC KDE Comparative Kernel Density Estimators.

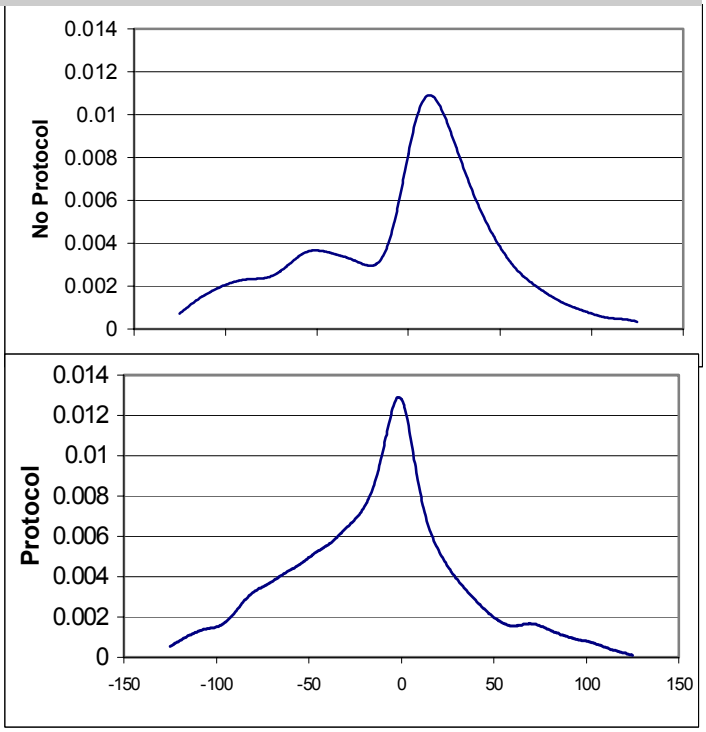


Figure 5. Comparison of Glucose Values When the Difference

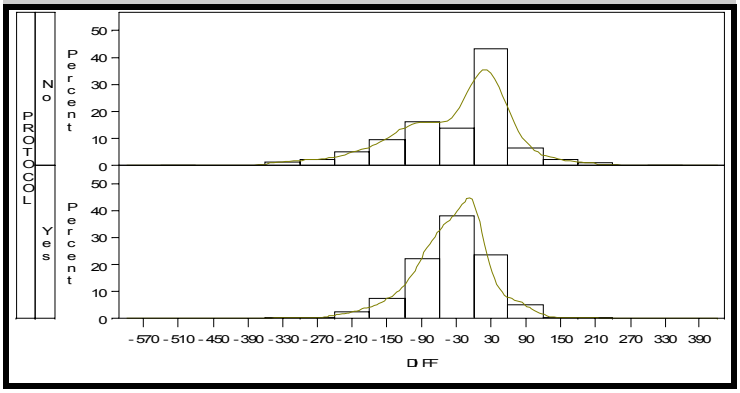


Figure 7. Bivariate Kernel Density of the No Protocol Data

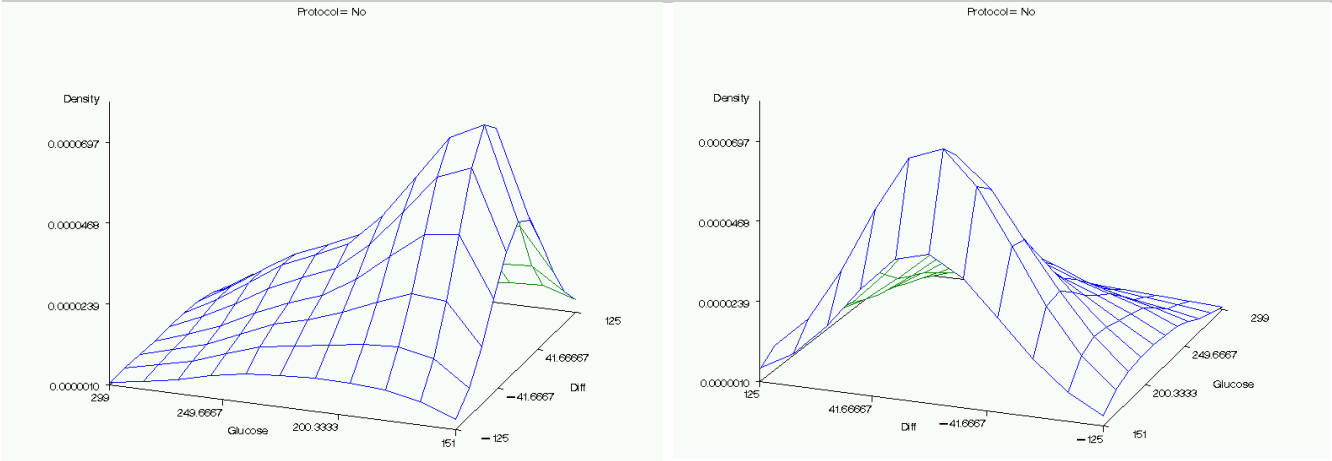
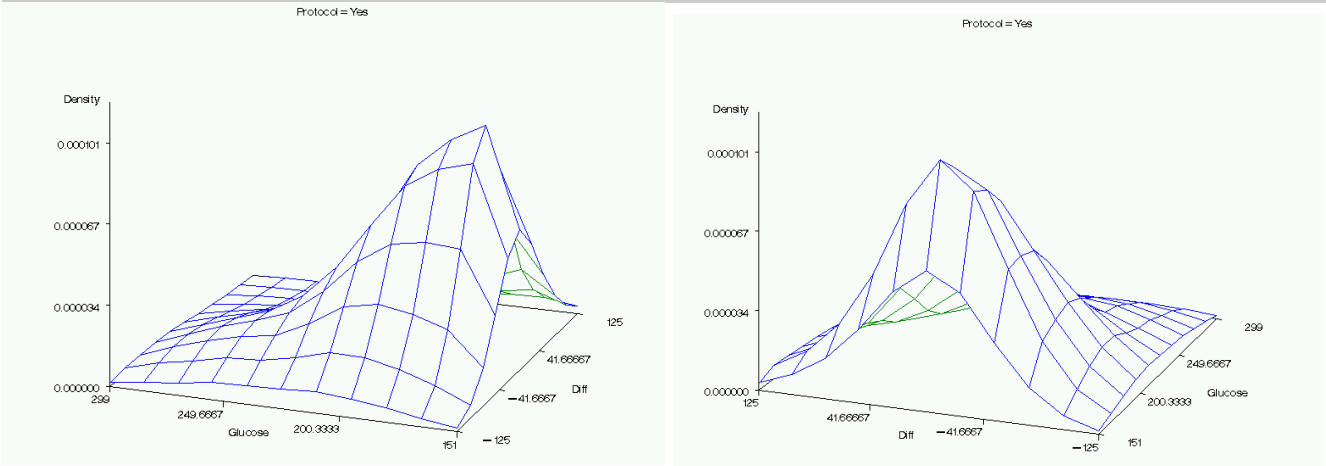


Figure 8. Bivariate Kernel Density of the Protocol Data



At this level, the difference in glucose level shifts from a positive to a negative mode under the protocol. With this distribution, there is an 80% change that once the glucose level climbs above 200, the next glucose level will decrease using the protocol compared to a 44% chance without the protocol. The likelihood of a negative difference increases to 90% when examining values with the initial point above 250 mg/dL.

It is possible to examine the bivariate distribution to explore the relationship between glucose level at time t and difference at time t+1 (Figures 7,8) . These curves were generated by surface plot after using the following SAS code:

```
proc sort data=work.compare;
by protocol;
run;
proc kde data=work.compare gridl=-125 gridu=125
bwm=2  ngrid=150,150 out=outkde;
var diff glucose;
where glucose>150 and glucose<300;
by Protocol;
run;
```

The default grid of 60 by 60 was increased to provide a better graph.

Inference With Kernel Density Estimators

PROC KDE does not permit inference. Therefore, it is not possible to test the hypothesis as to whether the two curves in comparison are statistically different. The primary problem with using kernel density estimators is that the estimator itself is degenerate under the null hypothesis. Therefore, the kernel density estimator must be modified to remove the degeneracy while at the same time preserving its properties.

This type of inference is not possible without the use of intensive computation. However, as discussed in Rosenblatt (1975), it is possible to find a test statistic based upon the kernel density estimator that is more powerful than statistics based upon the empirical distribution function. The output from PROC KDE can be modified to create a statistic that can be used in hypothesis testing. In addition, test statostocs of goodness-of-fit can be asymptotically normal under both the null and alternate hypothesis so that the power can easily be computed.

As proposed in Ahmad and Cerrito (1993), the null hypothesis H_0:f=f_0 for a known density function $f_0(x)$ can be examined using a modified kernel density function. Under the null hypothesis, the value of the kernel density function is equal to zero. Therefore, a measure of the goodness-of-fit is for the modified value δ_1 to approach zero.

$$\delta_1 = \int (f - f_0)^2 = \int f^2 - 2\int f f_0 + \int f_0^2$$

However, since the density function f is unknown, this values has to be estimated by equation (1).

$$\hat{\delta_1} = \left(m^2 a_m^p\right)^{-1} \sum_{i=1}^{m}\sum_{j=1}^{m} K\left[\frac{(X_i - X_j)}{a_m}\right] \quad (1)$$

$$-\left[2 \Big/ \sum_{i=1}^{m} C_{i,m}(\gamma)\right] \sum_{j=1}^{m} C_{i,m}(\gamma) f_0(X_i)$$

$$+\int f_0^2(x)dx$$

For any real number, $0<\gamma<1$, the triangular array $\{C_{i,m}(\gamma)\}$ consists of known real numbers satisfying (2).

$$m\sum_{i=1}^{m} C_{i,m}^2(\gamma) \Big/ \left[\sum_{i=1}^{m} C_{i,m}(\gamma)\right]^2 \xrightarrow{m\to\infty} C^2(\gamma) > 1 \quad (2)$$

A good choice for this array is $C_{i,m}(\gamma) = 1+\gamma$ for all m odd and $C_{i,m}(\gamma) = 1-\gamma$ for m even. Note that if $C_{i,m}(\gamma)$ is constant then $C^2(\gamma)$ is equal to one. Therefore, the series of functions cannot be constant. Under the null hypothesis, the value:

$\sqrt{m}\hat{\delta_1} \Big/ \sigma_{01}(\gamma)$ is asymptotically standard normal where σ_{01} is equal to:

$$\sigma_{01}^2(\gamma) = 4\left(C^2(\gamma) - 1\right)\left\{\int f_0^3(x)dx - \left(\int f_0^2 dx\right)^2\right\}$$

In the two sample case testing H_0:f=g where both f and g are unknown, then the L_2-norm is equal to

$$\delta_2 = \int f^2 + \int g^2 - 2\int fg$$

Again, this norm must be estimated (3).

$$\hat{\delta_2} = \left(m^2 a_m^p\right)^{-1} \sum_{i=1}^{m}\sum_{j=1}^{m} K\left[\frac{(X_i - X_j)}{a_m}\right] \quad (3)$$

$$+\left(n^2 b_n^p\right)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} K\left[\frac{(Y_i - Y_j)}{b_n}\right]$$

$$-\left(ma_m^p \sum_{j=1}^{n} D_{j,n}(\gamma)\right)^{-1} \sum_{i=1}^{m}\sum_{j=1}^{n} D_{j,n}(\gamma) K\left[\frac{(X_i - Y_j)}{a_m}\right]$$

$$-\left(nb_n^p \sum_{i=1}^{m} C_{i,m}(\gamma)\right)^{-1} \sum_{i=1}^{m}\sum_{j=1}^{n} C_{i,m}(\gamma) K\left[\frac{(Y_j - X_i)}{b_n}\right]$$

For this estimate, it must be assumed that the random sample $(X_1, X_2, \ldots, X_m)$ is independent of the random sample $(Y_1, Y_2, \ldots, Y_n)$. This can be satisfied if the two samples are taken from completely different subpopulations. Under the null hypothesis,

$$\sqrt{mn/(m+n)}\left(\delta_2 - \hat{\delta}_2\right)$$

Is asymptotically normal with mean 0 and variance

$$\sigma_{02}^2 = 4\gamma^2\left\{\int f^3 - \left(\int f^2\right)^2\right\}$$

in the special case where $C_{i,n}(\gamma) = 1+\gamma$ if i is odd and $C_{i,n}(\gamma) = 1-\gamma$ if i is even. The restrictions for the series $\{D(\gamma)\}$ is the same as that for the series $\{C(\gamma)\}$.

Since PROC KDE does not allow for the estimation of δ_1 or δ_2, the values estimated using PROC KDE must be adapted. The biggest problem is that these estimators fix both X_i and X_j whereas PROC KDE chooses equally spaced intervals for values of x in computing the density estimator. This can be done by limiting the number of points:

```
proc kde data=work.test gridl=0 gridu=150
ngrid=150 out=outkde;
var time;
where time<150;
```

In many instances, the data collected is limited by some measurement accuracy; in this case by integer amounts. It is possible to limit the grid to this measurement accuracy. The output from PROC KDE yields (for the estimated bandwidth a) the value in:

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

For all $x \neq X_j$ for some j, the points can be filtered out of the output table leaving only points of the form:

$$\left\{\frac{1}{na}\sum_{j=1}^{m} K\left(\frac{x - X_j}{a}\right) \middle| x \in \left\{\min X_j, \max X_j\right\}, j = 1,\ldots,n\right\}$$

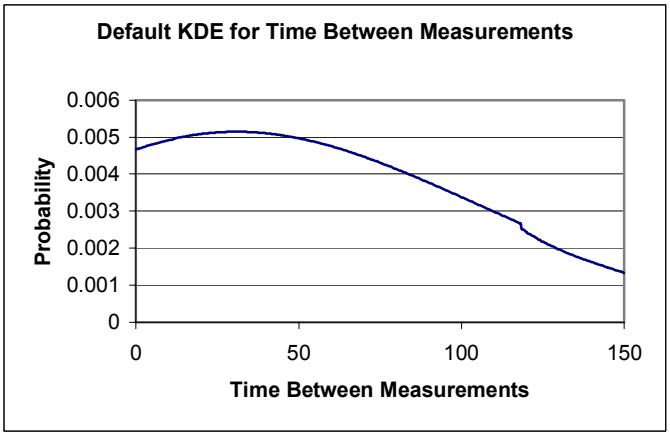
Leaving the values- $\frac{1}{na}\sum_{j=1}^{m} K\left(\frac{X_i - X_j}{a}\right)$ for I=1,…,m.

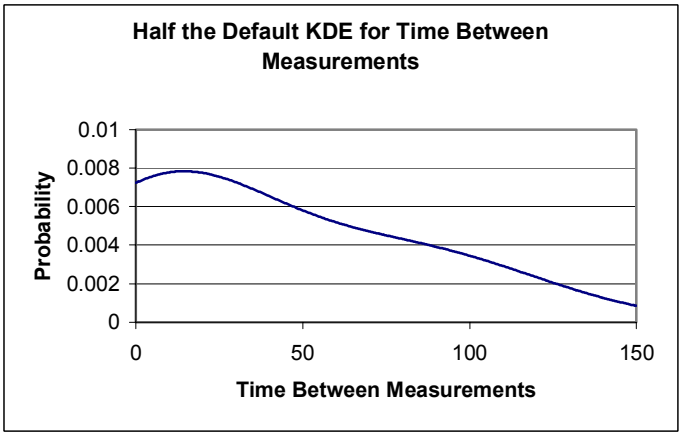
Once computed, these values can be summed and divided by m and a multiple of a so that δ_1 can be estimated.

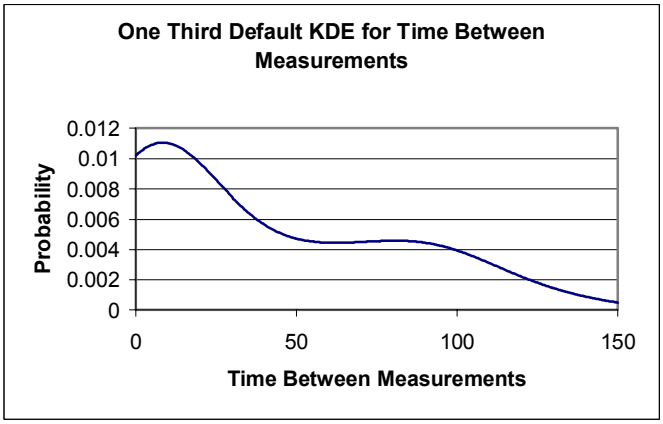
Similarly, the values for δ_2 can be computed repeating the process four times for points of the form (X_i, X_j), (Y_i, Y_j), and (X_i, Y_j). The series $\{C(\gamma)\}$ and $\{D(\gamma)\}$ can be equally defined.

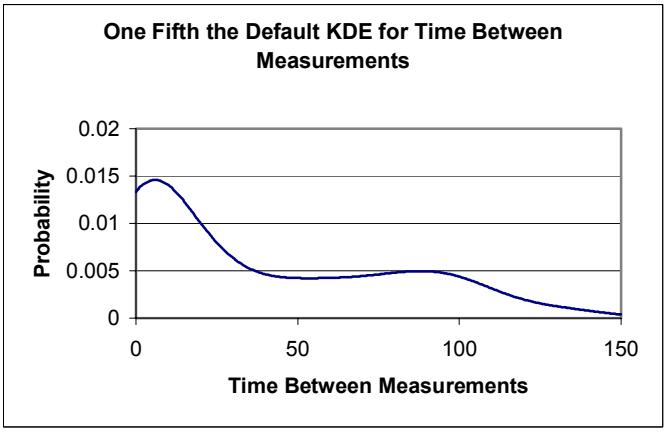
As there are numerous tests based upon the empirical distribution function or the chi-square goodness-of-fit that are much easier to compute (D'Agostino and Stephens, 1986), there must be some advantage to the use of kernel density estimators. This lies in the nature of the hypothesis test itself, particulary in the one-sample case. The purpose of testing H_0:f=f_0 is to determine whether the unknown distribution follows a specific, known pattern. Without knowing something about the power of the test, it is impossible to accept the null hypothesis; it is only possible to reject it.

Figure 9. Comparison of KDE Density Estimates for the Default Bandwidth, and Multiples of 0.5, 0.3, 0.2.









Unfortunately, power generally cannot be computed using the empirical distribution function because the form of the population distribution is unknown. Instead, the relative efficiency is computed based upon knowing the parametric family to which the unknown distribution belongs (Daniel, 1990; Stephens, 1993; D'Agostino and Stephens, 1986). More recently, a method of computing the power for empirical distribution tests was defined (Friedrich and Schellhaas, 1998) but the method depends upon recursion, also requiring substantial computation. In any case, the kernel density estimator uses more information from the data than does the empirical distribution function. Therefore, it has higher power than statistics based upon the empirical function. No such problem exists for the kernel density estimator. The distribution of

$$\sqrt{m}\left(\delta_1 - \hat{\delta}_1\right)$$

remains asymptotically normal with mean 0 and variance
Therefore, the power can be computed exactly, and can be esti-

$$\sigma_1^2(\gamma) = 4\left\{\int f^3(x)dx - \left(\int f^2(x)dx\right)^2\right\}$$

$$+4C^2(\gamma)\left\{\int f_0^2(x)f(x)dx - \left(\int f_0(x)f(x)dx\right)^2\right\}$$

$$-8\left\{\int f_0(x)f^2(x)dx - \int f_0(x)f(x)dx\int f^2(x)dx\right\}$$

mated by using the kernel estimator for the value of f in the variance equation. The one-sample case will be examined in an investigation of protocol compliance monitoring.

Further Examination of Glucose Data

COMPLIANCE WITH GLUCOSE MONITORING

It becomes important to monitor the level of compliance with the protocol through an examination of the data collected. Glucose levels were automatically recorded as to time and level. As the time of the glucose monitoring was recorded, it was examined to determine whether the glucose level was monitored within the time required by the protocol. PROC KDE was used to examine the distribution of monitoring times, comparing the results to the distribution provided by PROC CAPABILITY (Figures 9, 10).

In Figure 9, different multiples of the SJPI werew examined. The default value is too large and give an estimate that is over-smoothed. In this example, half the SJPI is also over-smoothed. Contrast this with the estimator provided by PROC CAPABILITY in SAS/QC (Figure 10).

This procedure can also superimpose the exponential density function. Judging by the histogram, the exponential density function is probably a better estimate than the kernel density since the kernel provides a decreasing estimate for negative values that cannot possibly exist in the actual data. Therefore, the hypothesis H_0:f=f_0 where f_0 is exponential can be tested. Only the points in the collected data were used to define the test statistic; the exponential tested was the one estimated by the PROC CAPABILITY. Using the SAS code provided on the previous page, the values needed in the estimator δ_1 were computed.

The output from PROC KDE was then merged with the original data set to eliminate all points not appearing in the original data. The code used to merge the two data sets is as follows (P 1):

```
proc sort data=WORK.KDE out=WORK._TABLE1_;
   by PROTOCOL ;
run;
proc sort data=WORK.KDE out=WORK._TABLE2_;
   by TIME ;
run;
data WORK.COMBINED;
   merge WORK._TABLE1_ (in=TABLE1) WORK.
_TABLE2_ (in=TABLE2) ;
   by TIME ;                          (P 1)
   if TABLE1;
run;
```

A combination of SAS code and use of spreadsheet computations seem to optimize the number of computations needed to compute δ_1 . To use the suggested values for C(γ):

$$C_{i,m} = \begin{cases} 1+\gamma & \text{if i is odd} \\ 1-\gamma & \text{if i is even} \end{cases}$$

The following SAS commands were used:

(P 2)

```
Testvalue=INT(Time/2)*2;
If (Testvalue eq Time) then Coeffi-
cient=0.5;
If (Testvalue ne Time) then Coefficient=1.5;
```

in a data statement to assign values to the variable Coefficient that are needed to compute δ_1. This is assuming that Time(i)=i. The statement defining Product1 was added:

```
Product1=Coefficient*theta*Exp(-
Parameter*Time)
```

The exponential density function were computed at the data points. For this example, the value of the parameter used was that that estimated using PROC CAPABILITY. The next step is to sum the density function values for all i using PROC MEANS. Using the results, Product2 was computed:

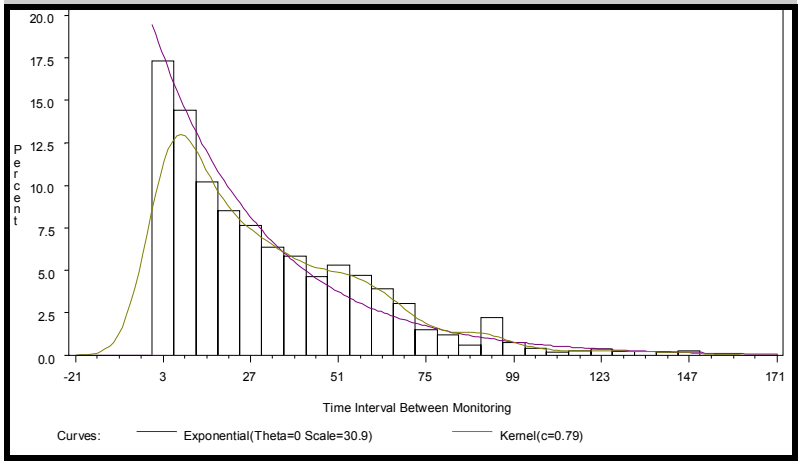
```
Product2=Density*SumofDensity
```

Where SumofDensity in this case is equal to the constant 57.9103243. What remains is to compute the sums of these variables: Product1, Product2 , and Coefficient as well as to compute the integrals of the exponential density squared and cubed. The final result yields a p-value<0.0001. The effect size is extremely small because the data set contains 3634 points. To increase the effect size, a random sample of 150, and then of size 50 was chosen from the data. In both cases, the p-value remains <0.0001. Therefore, it is clear that the distribution, while appearing to be exponential, does not satisfy the null hypothesis. A visual examination indicates that from time 50 to time 100, the distribution is heavier than that of the exponential distribution.

For the 2-sample case, the values can be com-

$$\frac{1}{ma_m^p}\sum_{i=1}^{m}\sum_{j=1}^{m} K\left(\frac{X_i - X_j}{a_m}\right), \frac{1}{nb_n^p}\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{Y_i - Y_j}{a_m}\right)$$

Figure 10. Time Between Glucose Readings Using PROC CAPABILITY



puted in the same manner as in the one-sample case. The SAS codes P1 and P2 can be used in the same way for the remaining two sums. The order of Table1 and Table 2 are reversed depending upon whether X or Y values are first listed. P2 is defined for the Time value first for X and then for Y. However, the Product2 values are modified:

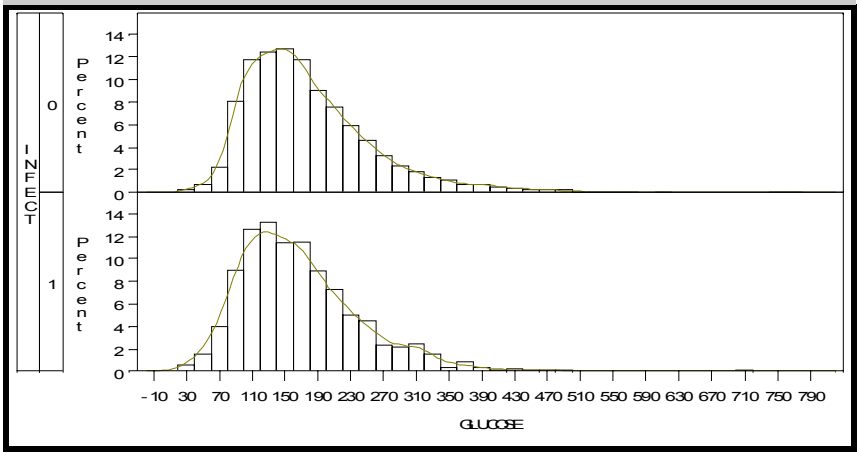
```
Product2=Density(Y)*Coefficient(X)
*SumofDensity(X);

Product3=Density(X)*Coefficient(Y)
*SumofDensity(Y);
```

GLUCOSE MONITORING AND INFECTION

First, a comparison was made between all glucose levels for infection. The density functions look very similar to the one in Figure 11. Note that in both cases, the functions are heavy-tailed, indicating that glucose levels can climb extremely high However, it does give an indication that the differences between infected and non-infected patients is not in the glucose values themselves and that various parameters need to be examined for diabetic patients. Therefore, on an initial examination, there does not appear to be a relationship between glucose levels and infection.

Figure 11. Kernel Density Estimators Comparing Infection to Non-Infection in All Patients.



Again, it is not feasible to use this data in an analysis of variance because of the dependency relationship between data points.

In the above code, Density(X) represents the values of the form

$$K\left(\frac{X_i - Y_j}{a_m}\right)$$

and Coefficient(Y) represents values in the coefficient series {D(γ)}. With these computations, the p-value<0.0001 using the total sample, as well as subsamples of size 50 and 150. It can be concluded that the distributions estimated in Figures 3 and 4 are not equal.

INFECTION RELATIONSHIP

It was also necessary to examine the parameter relationships (if any) between glucose levels and infection. For the first 60 patients on the protocol, no infections were found, indicating that the protocol can lead to a decrease in the nosocomial infection rate. However, the first 60 patients were not considered sufficient to adequately gauge the infection rate as infection in diabetic patients is generally under 10% of the population. An initial examination of the infected and non-infected population did not demonstrate a difference in the distributions (Figure 11). Again, the two density functions cannot be compared using standard methods because the data are from a strong mixing process (weakly dependent). Therefore, comparisons of the curves inferentially require kernel density estimation.

Data were made available to compare infection rates to glucose of diabetic patients from July through December, 1998. There were 609 patients without infection with glucose levels monitored in OR, and 20 diabetic patients with infection also monitored. Not all patients were represented in this sample as glucose levels were not collected on all patients. However, this analysis was not intended to be conclusive; it was intended to see if evidence did show that glucose and infection were related.

Various parameters were computed to determine whether relationships existed. Particular attention was given to peak glucose levels since diabetics have difficulty with high levels of glucose. Low values were also considered but discarded since there were very few measurements.

For each patient, the average and maximum glucose values were computed. The results were used in an analysis of variance to compute the difference between infected and non-infected patients.

The difference in peak values is statistically significant (p=0.0054) as is the difference in average values (p=0.0328) between the infected and non-infected samples (Table 1).

Note that the average peak values differ by almost 80 mg/dL; the difference in average is more modest. This is to be expected. However, the protocol is driven more by the peak glucose values than by the averages. Therefore, it is useful to examine the distributions of these peak values.

Figure 12. Comparison of Peak Glucose Values for Infected and Non-Infected patients.

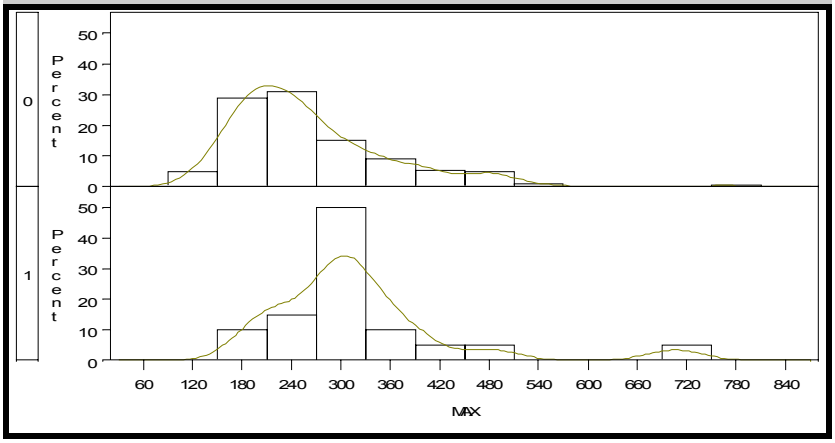
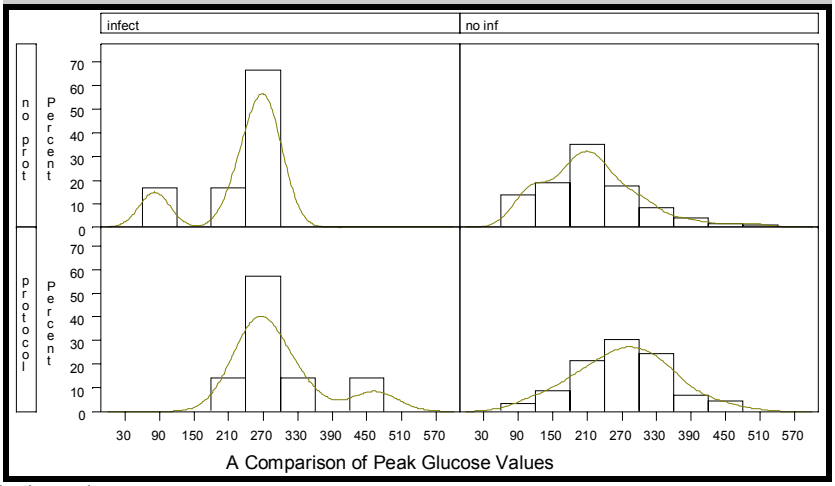


Figure 13. Comparison of Infection versus No Infection, and Protocol versus No Protocol.



As the kernel density estimator demonstrates, there is a clear difference between infected and non-infected patients (Figure 12). For infected patients, approximately 20% have a mean glucose value above 200; 70% have a mean glucose value above 150. For non-infected patients, only 5% have a mean glucose value above 200; only half have a mean glucose above 150. The difference between the two groups of patients becomes even clearer when examining the peak glucose levels for each patient. For infected patients, the peak values cluster around 300 (approximately half of the patients have peak values of 300 or more); for non-infected patients, the peak glucose values cluster around 240 with only 15% above 300.

The result is not conclusive since there are only 20 patients with infection in the data sample. However, it is sufficient to indicate that the protocol can have an impact on infection rates provided that it is successful in reducing the peak glucose levels for diabetic patients. Based upon this information, the protocol was implemented.

Table 1. Average and Peak Glucose Values for Infected and Non-Infected Patients.

Infection	**Average of the Peak Values**	**Average of the Mean Values**
No Infection	261.67	159.01
Infection	322.05	176.75

ANALYSIS OF INFECTIONS AND PEAK GLUCOSE VALUES FOR PATIENTS ON AND OFF PROTOCOL.

The protocol data from February to May, 1999 were collected and examined. There were a total of 436 patients examined with 13 infections at a rate of 2.98%. A total of 96 out of the 436 (22%) were identified as being on the protocol. Because of the small number of infected patients, it was difficult to demonstrate statistical significance. However, of this number, 7 were identified as on the protocol; 6 as not on the protocol. In addition, infection resulting from surgery is defined as occurring within 1 year after the surgery. Therefore, the number of patients infected was incomplete at the time of the analysis. The data were examined to determine whether differences exist. Any differences need to be validated with additional data to be conclusive.

The density function can also demonstrate that patients on the protocol still peak at higher glucose levels than patients not on the protocol (Figure 14). There is a very noticeable shift in the distribution. As a result, efforts were made to improve peak glucose levels for the diabetic patients. In July, 1999 the protocol was changed to require a 3-day post-surgical drip of insulin-glucose as part of the protocol.

It was demonstrated that patients with infection had higher peak glucose values than patients without infection (Figure 13). For patients without infection, 60% of the patients have peak glucose levels below 250, and 80% have peak glucose below 280. For infected patients, only 25% have peak glucose below 250, with 60% below 280. Therefore, the results of the previous analyses are reinforced here. Infection is strongly associated with high peak glucose values. Note that the average glucose value for each patient is indistinguishable between infected and non-infected patients.

Out of the 436 patients, comparing average glucose values, the difference is not significant (p=0.1887) for infection, nor for protocol (p=0.0638). Contrast this with the peak glucose levels for infection (p=0.3690) and protocol (p<0.0001). A summary is given in table 2.

In August, 1999, the data were again examined (Figure 15). It should be noted here that the glucose peaks for non-infected, protocol patients is only slightly higher than for the non-protocol patients. However, there is a difference in distributions for the protocol, infected patients.

Figure 14. Comparison of Peak Glucose Levels for Patients On and Off Protocol.

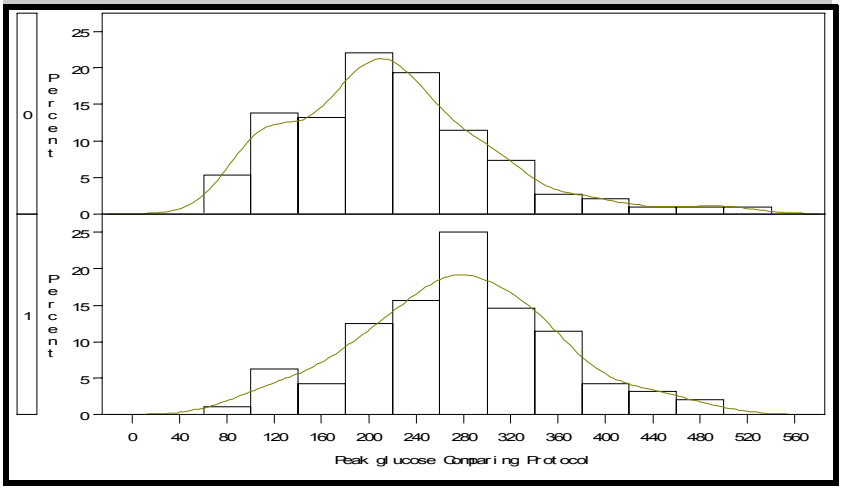


Figure 15. Comparison of Glucose and Protocol for Data January-August,

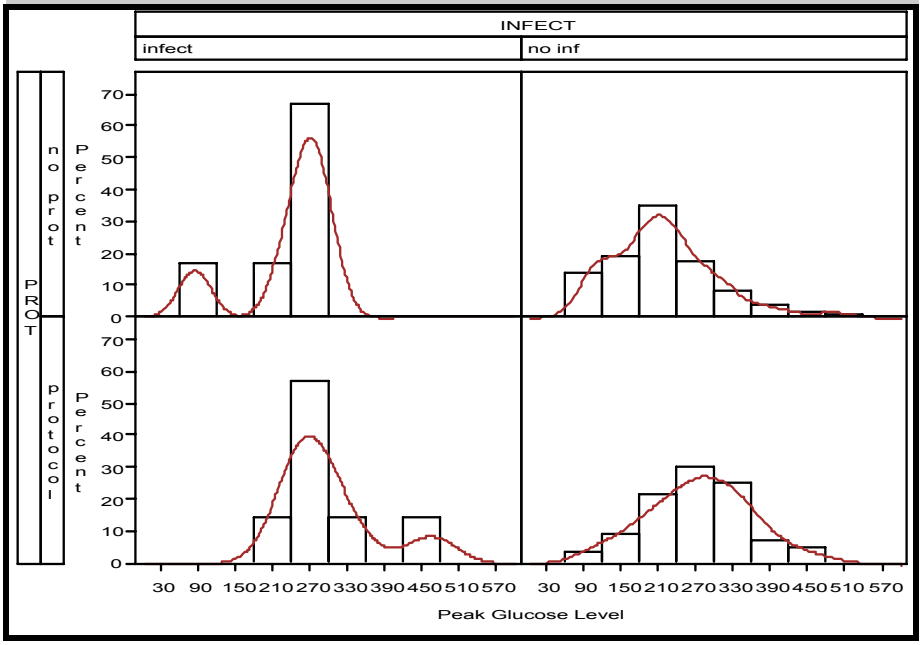


Table 2. Average and Peak Glucose Values for Infected and Non-Infected Patients, Comparing On and Off Protocol.

Infection	Protocol	Average of the Peak Values	Average of the Mean Values
Infection	On Protocol	300.00	184.05
Infection	Off Protocol	234.67	169.48
No Infection	On Protocol	274.22	165.10
No Infection	Off Protocol	217.96	155.82

The data were examined to determine the time of glucose peak. Location, in part, determines the time frmae of the glucose peaks. Therefore, the data were divided by patient location. Pre-op peak glucose values were compared to glucose values during surgery (Figure 16). It is clear from the distributions that patients on the protocol have a larger shift in peak glucose levels between pre-op and surgery. However, it is also of interest to note that in pre-op, the diabetic patients do not demonstrate much of a difference in peak glucose values when compared to the non-protocol patients. However, the results are not sufficient to demonstrate statistical significance in an analysis of variance.

In recovery, it is very clear that there is a difference between protocol and non-protocol patients (Figure 17). For non-protocol patients, there is a very quick reduction in the peak glucose level, coming down to normal range (less than 180 mg/dL). Also, for non-protocol patients, the peak glucose level is in a very narrow range. For patients on protocol, the glucose level does decrease, but not at as great an extent as non-protocol patients. Also, there

is considerable variability. The difference in recovery is statistically significant (p=0.0002).

CONCLUSION

Compliance with medical treatment is a very difficult parameter to monitor. Without such monitoring, it can be difficult to determine just how effective an experimental treatment actually is. Kernel density estimators can be used to continuously monitor an implemented protocol both for compliance and for patient benefit.

REFERENCES

1. Altman, Naomi; Léger, Christian. (1995). Bandwidth selection for kernel distribution function estimation. J. Statist. Plann. Inference 46. no. 2, 195--214.
2. Ahmad IA, Cerrito PB. (1993). Goodness of fit tests based on the L_2 norm of multivariate probability density functions. Nonparametric Statistics. 2:169-181.
3. D'Agostino RB, Stephens MA. Goodness-of-fit techniques. (1986). Statistics: Textbooks and Monographs, 68. Marcel Dekker, Inc., New York.
4. Daniel WW. (1990). Applied nonparametric statistics, 2nd Ed. PWS-Kent Publishing Company: Boston, MA.
5. Friedrich, T, Schellhaas H. (1998). Computation of the percentage points and the power for the two-sided Kolmogorov-Smirnov one sample test. Statist. Papers 39(4), 361--375.
6. Jones MC, Marron JS, Sheather SJ. (1996). A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association. 91 (433):401-407.
7. Masry, Elias (1987). Almost sure convergence of recursive density estimators for stationary mixing processes. Statist. Probab. Lett. 5(4), 249--254.
8. Rosenblatt M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann. Statist. 3:1-14.
9. Silverman, BW. (1986). Density estimation for statistics and data analysis. London: Chapman and Hall.
10. Stephens, Michael A. (1993). Aspects of goodness-of-fit. Statistical sciences and data analysis 395--405, VSP, Utrecht.

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. The authors also wish to acknowledge the aid of the infection control and protocol investigators: David Bybee,, MD Julio Melo, MD, and Veronica Pennington, RN. The authors also want to acknowledge the support of the National Science Foundation in the development of data mining techniques for clinical databases.

CONTACT INFORMATION

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

Figure 16. Comparison of Peak Glucose Values Pre-Op versus Surgery

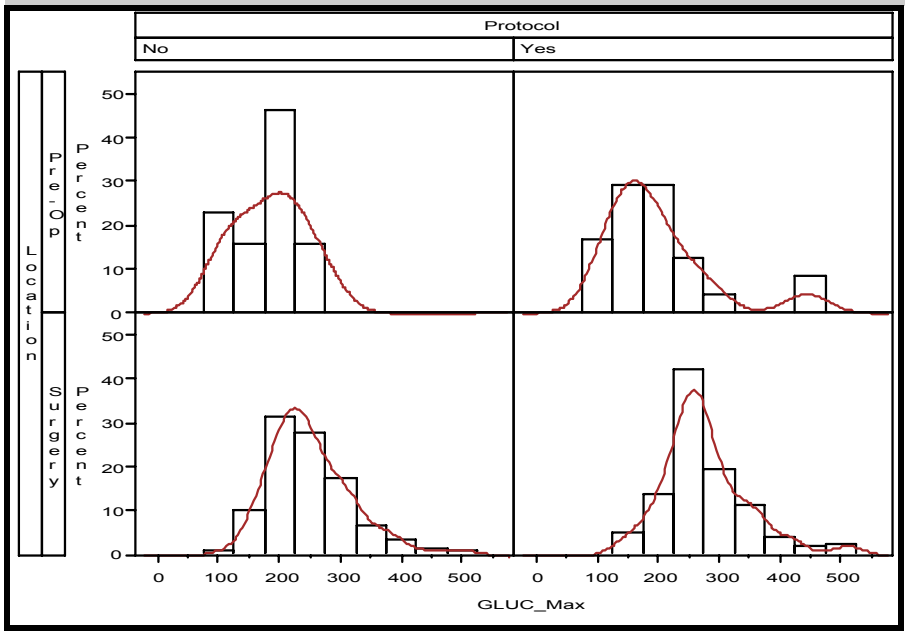
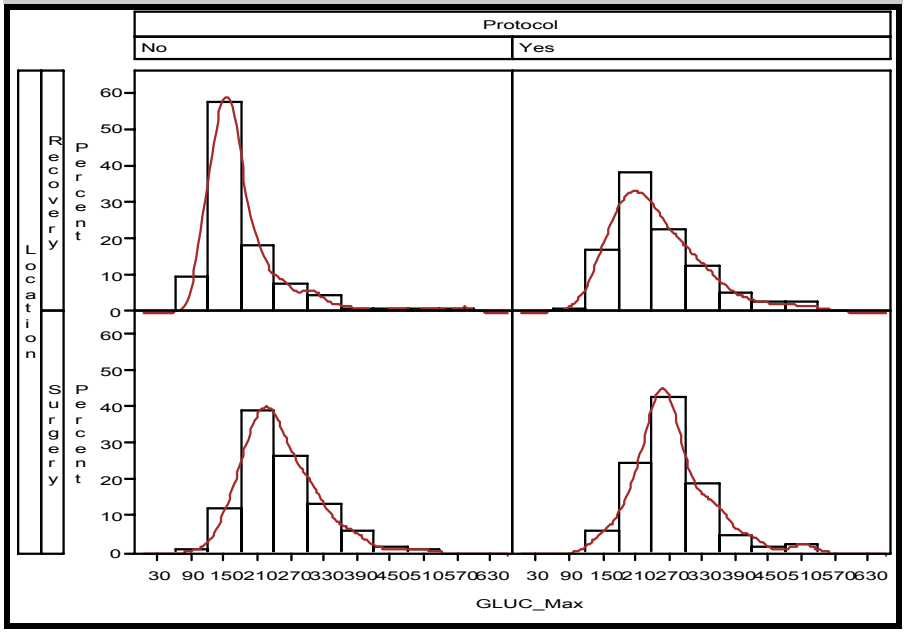


Figure 17. Comparison of Peak Glucose Levels During and After Surgery



TRADEMARK CITATION