Paper 130-2010

# Kernel Density Estimation as an Alternative to Binning in the Analysis of Survey Data

David J. Corliss, The University of Toledo, Department of Physics and Astronomy, Toledo, OH

## Abstract

In the statistical analysis of survey data, a large number of data points having a continuously distributed observed variable may be grouped into ranges of constant width, a process known as "binning". For example, stars may be grouped into ranges defined by a number of solar masses. In binning data, a certain amount of information about the object is often lost: any information at a higher degree of accuracy than needed to place it into a bin is discarded. A methodology is proposed for the determination of population distributions allowing for full retention of the measured value for each observation in cases where the uncertainties are expected to be Gaussian. If the uncertainties are normally distributed, a Gaussian function may be determined for each measurement, with the observed value as the mean and the uncertainty as the standard deviation. The Gaussian distribution for each observation are then summed, creating a continuous probability density distribution. Where observation data lie within the limitations of this technique, all available data can be incorporated into the final population distribution without loss of information due to binning.

Keywords: Kernel Density Estimation, KDE, Binning, Astrostatistics

## Binning Data

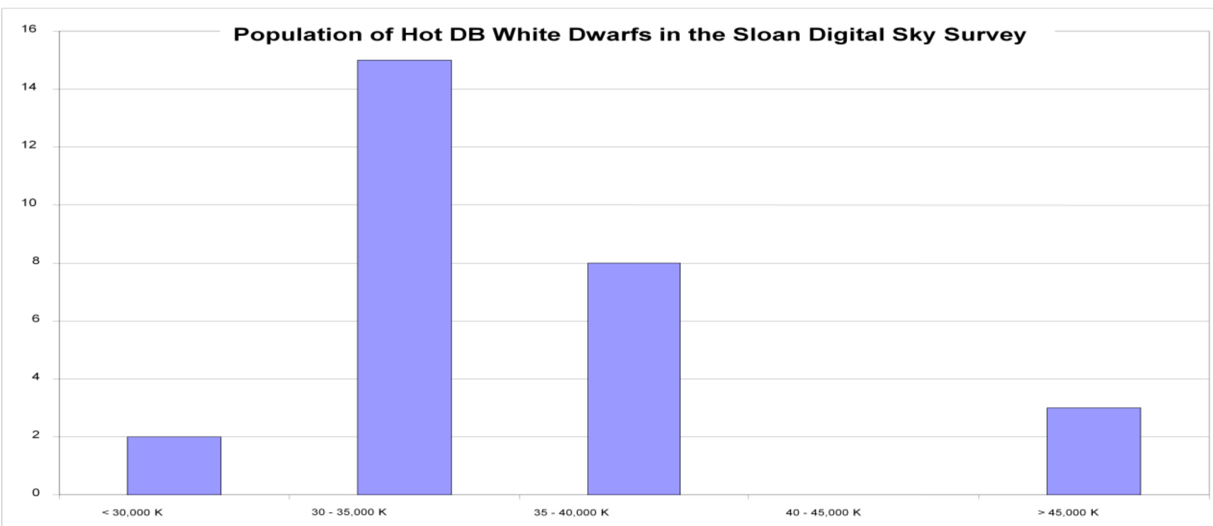


Figure 1 – Population Distribution of hot DB white dwarfs described by Eisenstein et al. 2006

Binning of data is commonly used in the analysis of a continuous variable. While this can simplify management of the data, the actual situation can be more complex. Information at a higher degree of accuracy than needed to place it into a bin is discarded. Also, binning creates a new source of systematic error, as observational uncertainty leads to uncertainty of whether a point is placed into the correct bin. Thus, the uncertainty in observational measurement is carried over into an uncertainty in the number of records contained in each bin.

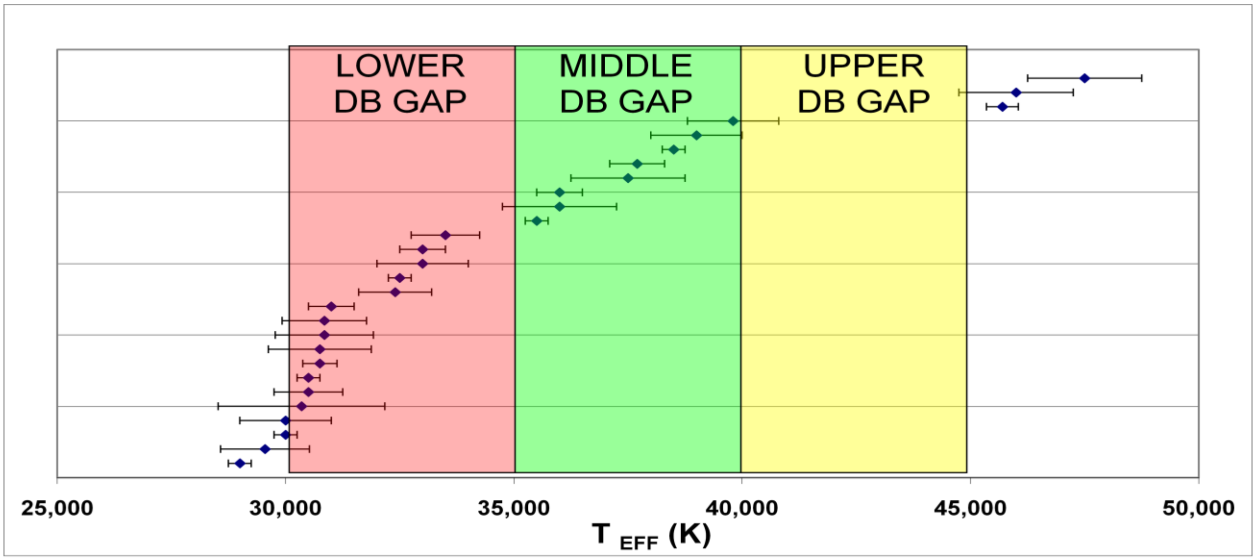


Figure 2A – Population Distribution of hot DB white dwarfs described by Eisenstein et al. 2006b

## Kernel Density Estimation

Kernel Density Estimation (KDE) is a commonly encountered as a process for smoothing data. This is accomplished by replacing each data point with a Gaussian distribution. A Gaussian distribution can be completely defined by only two values: the mean μ and the standard deviation σ. In Kernel Density Estimation, the Gaussian distribution used to replace each observation takes with the observed value for the mean of the Gaussian distribution and sets σ equal to the standard deviation of the observed value.

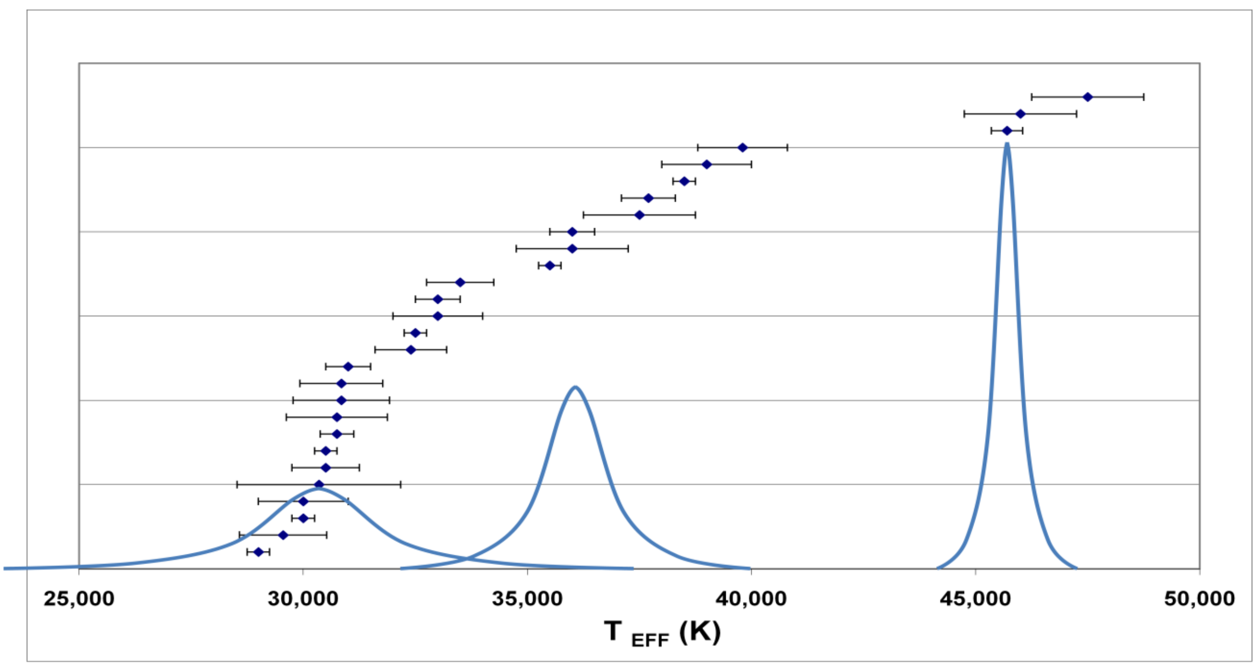


Figure 2B – Population Distribution of hot DB white dwarfs described by Eisenstein et al. 2006 b

The KDE is evaluated numerically by calculating the value of each Gaussian at a large number of evenly spaced values of the independent variable. These values are summed over all observation, providing a continuous probability distribution.

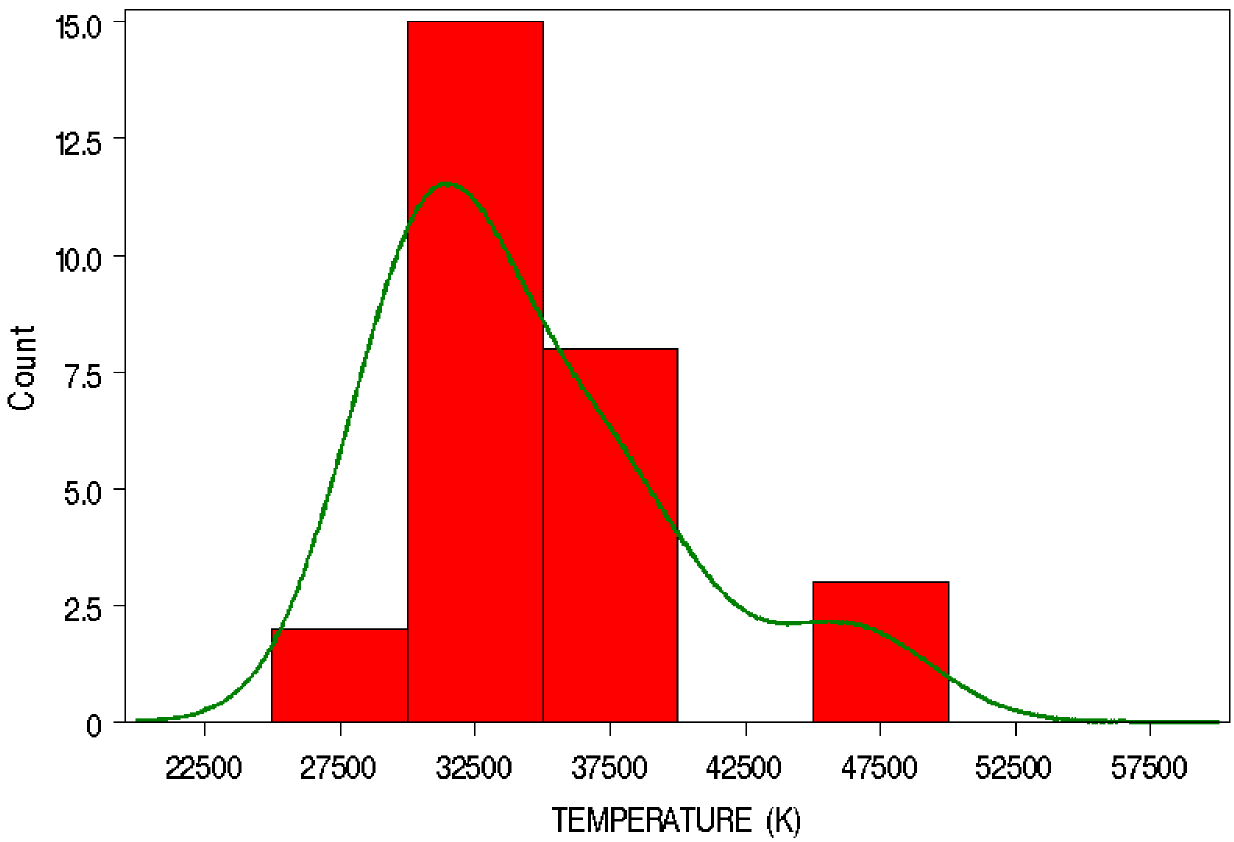


**Figure 3 - Hot DB White Dwarfs in Eisenstein et al. 2006: Histogram and KDE Plot**

## PROC KDE

The SAS procedure PROC KDE implements Kernel Density Estimation using a single value for the standard deviation for all observations. This is appropriate for survey sample limited only by the number of records. However, a more general case may be presented by experimental data where the amount of uncertainty can vary from one observation to the next. In the SAS program, a distinct Gaussian distribution is created for each observation. The mean μ for each Gaussian is given by the observed value and σ is equal to the experimental uncertainty, given as the standard deviation of the value of the individual observation. In order to address this general case, the SAS code uses only base SAS and not PROC KDE.

## SAS Source Code

```
data work.records;
   seq_num = _n_;
   dummy = 1;
   input name $20. mu 6.0 sigma 4.0;
   cards;
J084916.1+013721    29000 250
J093759.5+091653    29550 975
```

```
J090232.1+071929    30000 250
J153852.3-012133    30000 1000
J093041.8+011508    30350 1825
J215514.4-075833    30500 750
J141349.4+571716    30500 250
J141258.1+045602    30750 375
J222833.8+141036    30750 1125
J234709.3+001858    30850 1075
J143227.2+363215    30850 925
J212403.1+114230    31000 500
J095256.6+015407    32400 800
J154201.4+502532    32500 250
J123750.4+085526    33000 1000
J164703.4+245129    33000 500
J084823.5+033216    33500 750
J001529.7+010521    35500 250
J090456.1+525030    36000 1250
J211149.5-053938    36000 500
J040854.6-043354    37500 1250
J140159.1+022126    37700 600
J092544.4+414803    38500 250
J134524.9-023714    39000 1000
J074538.1+312205    39800 1000
J113609.5+484318    45700 350
J081546.0+244603    46000 1250
J081115.0+270621    47500 1250
run;

proc sort data=work.records;
   by mu;
run;

proc sort data=work.records;
   by dummy;
run;


**** minimum and maximum x-values ****;

data work.x_min;
   set work.records;
   x_min = mu - (2 * sigma);
   keep dummy x_min;
run;

proc sort data=x_min;
   by x_min;
run;

data work.x_min;
   set work.x_min;
   by dummy;
   if first.dummy;
run;
```

```
data work.x_max;
   set work.records;
   x_max = mu + (2 * sigma);
   keep dummy x_max;
run;

proc sort data=x_max;
   by x_max;
run;

data work.x_max;
   set work.x_max;
   by dummy;
   if last.dummy;
run;


data work.min_max;
   merge work.x_min work.x_max;
   by dummy;
   x_range = x_max - x_min;
run;

data work.records;
   merge work.records work.min_max;
   by dummy;
run;


**** KDE Process ****;

data work.final;
   set work.records;
   by dummy;
   x = x_min;
   y_i = (1/(sigma * ( SQRT(2 * constant('pi') ) ) ) ) *
         EXP((-0.5)*( ( (x - mu) / sigma )**2));
   retain y 0;
   y = y + y_i;
   if last.dummy then output work.final;
   keep x y;
run;

%macro kde(iter);
   %do i=1 %to &iter;
      data work.tot;
         set work.records;
         by dummy;
         x = x_min + ((&i. / &iter.) * x_range);
         y_i = (1/(sigma * ( SQRT(2 * constant('pi') ) ) ) ) *
               EXP((-0.5)*( ( (x - mu) / sigma )**2));
         retain y 0;
         y = y + y_i;
         if last.dummy then output work.tot;
```

```
         keep x y;
      run;

      data work.final;
         set work.final work.tot;
      run;
   %end;

%mend kde;

%kde(500);
```

## Summary and Conclusions

Kernel Density Estimation creates a continuous probability density distribution by summing over Gaussian distributions for each data point, Where μ is the observed value and σ is the σ of the individual measurement. This process prevents loss of information from relatively accurate measurements being placed into larger bins, incorporating the uncertainty associated with measured values into population distributions and provides a viable alternative to binning in developing population distributions for survey and other data.

## Acknowledgments

This research was performed at the University of Toledo Department of Physics and Astronomy under Dr. Nancy Morrison. Work on addressing issues of binning survey data using a continuous probably distribution based on replacing each point with a gaussian distribution began while in attendance at Summer School for Astrostatistics at Penn State University in June 2009. The suggestion to address these issues by leveraging the SAS KDE procedure was made by Dr. John Sall at the Midwest SAS User Group conference in October, 2009. This paper was presented in summary form at a meeting of the American Astronomical Society in January, 2010, where session members made useful comments and improvements.

## References

Babu, G. Jogesh, Summer School in Statistics for Astronomers V lecture Notes, Pennsylvania State University 2009
Barnes, George R., Cerrito, Patricia B., The Visualization of Continuous Data Using PROC KDE and PROC CAPABILITY , SUGI, 26, 2001
Corliss, David J., MS Thesis, Wayne State University, 2008
Eisenstein, D.J., et al., 2006, ApJS, 167, 40 (Eisenstein et al. 2006a)
Eisenstein, D.J., et al., 2006, ApJ, 132, 676 (Eisenstein et al. 2006b)
Sall, John – Personal Communication re. the SAS KDE Procedure

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the author at:

David Corliss
Marketing Associates
777 Woodward Avenue, Suite 500
Detroit, MI 48226
(313) 202 - 6323
dcorliss@marketingassociates.com