PharmaSUG 2015 - Paper DV01

Variable-Width Plot - SAS® GTL Implementation

Songtao Jiang, Boston Scientific Corporation, Marlborough, MA

ABSTRACT

If you do statistical analyses using S-PLUS, you may be familiar with BLiP graphical tool. The variable-width plot (VWP) included in this tool was developed to graphically present the one-dimension data. A unique feature is its capability of drawing the distribution plot with variable width proportional to the density at particular locations. Without certain limitations of some commonly used graphical methods, such as scatter plot, histogram, or boxplot, the VWP is a very useful tool for studying the data distributions. While the variable-width plot in the BLiP tool has been implemented in the S-PLUS, this type of plot is not available in SAS. In this paper, we implement VMP using SAS GTL. The properties and benefits are discussed by comparing it to other existing graphical methods.

KEYWORDS

Variable-width plot, BLiP, density function, GTL

INTRODUCTION

For sample distributions, the commonly used visualization tools including scatterplots, histograms, density plots, boxplots, and variations or combinations of these methods. Depending on different situations, one method may be better than another. The scatterplots emphasis more on the individual data point, but the histograms are more on the grouped data. While density plots present estimated data densities at different locations, the boxplots provide more details on the quartiles, mean, median, minimum and maximum. The variable-width plot introduced in the paper combines some of the features from scatterplots, histograms, density plots, and boxplots. The variable-width plots look similar to histogram but provide additional density information and percentiles of the data distribution.

METHODS

As described in the a few papers: the variable-width plot at particular location is proportional to the density estimation at that position (Warren W, 2003). It uses a computationally simple central-difference estimator (Rosenblatt 1956) in the program to produce graphs that look similar to box-percentile plots (Lee and Tu, 1997).

The unique feature of the VWP is that the widths of the individual bins of histogram vary depending on the sample data distribution. The bin widths of the histogram depend on the specified sample data percentiles. The heights of the bins are proportional to the densities at the locations of the sample data percentiles.

Two things need to be specified before constructing a VWP. First, the percentiles of the VWP, this determines the number of bins of the histogram, the locations of densities, and the bin widths. Second, the methods used to estimate the densities which determine the heights of the bins of the VMP.

For the illustration purpose, we select deciles in this paper. 10% of data points should fall in each partition of the sample data. A set of numbers (minimum, deciles, maximum) can be generated from these 10 partitions from the sample data, as the locations of the partitions:

$$\{min = x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} = max\}, \qquad where\ x_i\ (i = 1, \dots, 9)\ are\ deciles$$

The bin width is calculated based on these locations:

$$w_i = x_{i+1} - x_i, \qquad where\ i = 1, \dots, 10$$

Once the locations of the partitions have been determined, the densities at these locations are estimated by using the simple central-difference estimator (Rosenblatt 1956).

$$\hat{f} = \frac{\hat{F}\left(x_i + \frac{h}{2}\right) - \hat{F}\left(x_i - \frac{h}{2}\right)}{h}, \qquad where\ i = 0, \dots, 10, and\ h\ is\ a\ smoothing\ factor$$

IMPLEMENTATION

To illustrate the implementation, we generated 100,000 data points following Chisquare (df=10). Here are the 3 steps:

1) Use SAS PROC RANK to generate the deciles$\{\boldsymbol{min = x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} = max}\}$, (Figure 1)

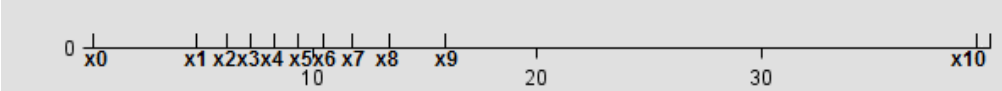

Figure 1: Partitions of the Sample data with Deciles

2) Calculate the relative densities at $\{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$ based on the sample data using central-difference estimator. Draw the Needle Plot with the heights as the estimated relative densities (Figure 2).

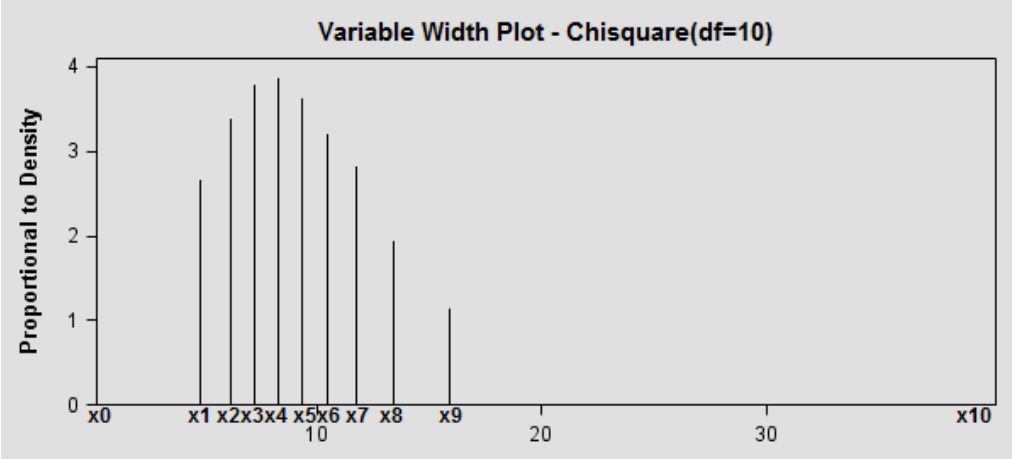

Figure 2: Needle Plot of the Densities

3) Connect the top of the needles, the final VMP shown as following (Figure 3):

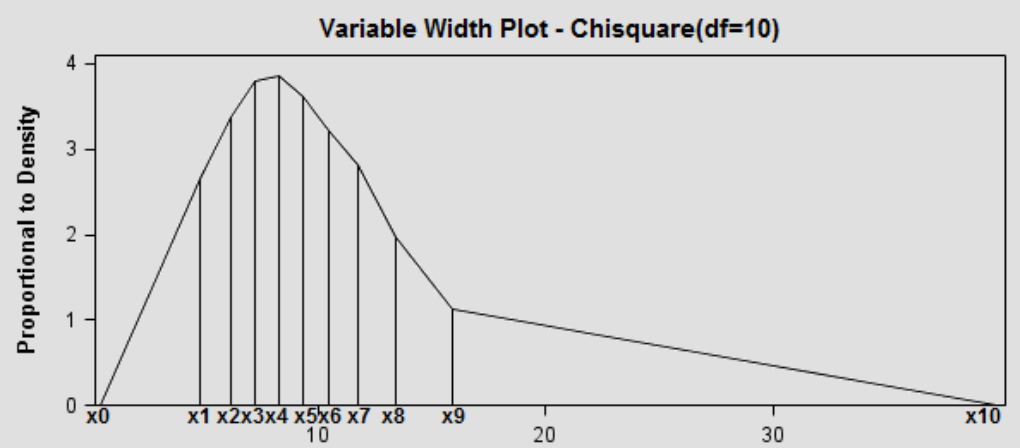

Figure 3: Final Plot of VMP

PROPERTIES

VMP combines 3 important properties related to the sample data distribution into one plot.

Property 1: the percentiles of the sample data distribution are presented at locations of the partitions.

Property 2: the sample data densities (or proportion to the densities) at the location of the partitions are represented by the heights of the needles.

Property 3: the densities estimated are based on the sample data with your choice of density estimator.

Compared to the histogram plot, the VMP contains more information about the sample data distribution with percentiles and relative densities. It has the variable width of bins. For histograms, the midpoints must be evenly spaced. The midpoints are not necessarily located at the percentiles. The percentages are for evenly-spaced areas centered at the midpoints. As shown the following 2 plots (Figure 4), the first one is the VMP for the Chisquare (df=10). The second plot is from the combination of histogram plot, density plot, fringe plot, and boxplot. As we observe, the VMP contains more information related to the densities and the percentiles (min, max, median, and the

deciles). From the VMP, the highest data density is observed around 30 percentile (x3=7.3, if the minor tick marks on x-axis were presented). There are about 40% of the data packed between x2 and x6. The VMP also shows the relative locations of mean and median, which could be very useful to study the data skewness. On the other hand, the histogram plot along can't easily present all these information related to the sample data distribution. The similar level of information is only obtained from combining multiple plots shown in the second plot.

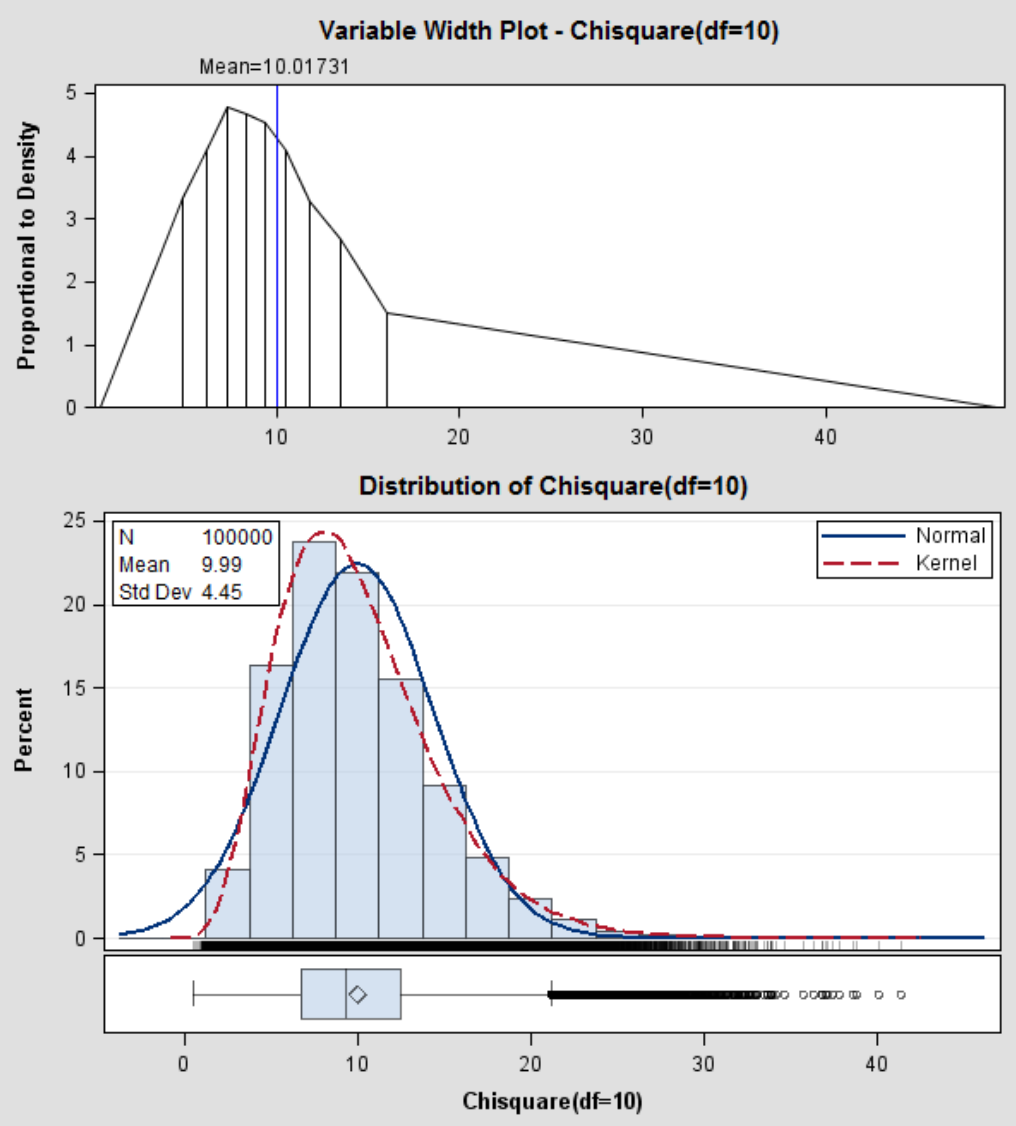

Figure 4: Comparison of VMP and Other Plots

A REAL EXAMPLE

We want to compare three methods for computing confidence intervals (CI): standard Wald CI, Exact CI, and profile likelihood CI. Simulations have been carried out for all the methods. The coverages of 95% CI have been calculated for 10,000 times for each method. We expect that the 95% CI should cover 95% of the simulated data points. The simulation results for the CI coverages are shown using the VMP (Figure 5). From the graphs, we observe that the Profile method performs the best. There are about 80% of CIs with coverages of 95% or more of simulated data points. 20% of CIs have coverages of less than 95% of simulated data points. Most of the CIs (the vertical lines) are tightly around 95% reference line (red), which is highly desirable for the 95% CI. The Wald method performs a little bit worse than the Profile method with the vertical lines pushed slightly away from 95%, which means these CIs are over estimated or too wide. And as expected, the exact method performs worst with even wider CIs.

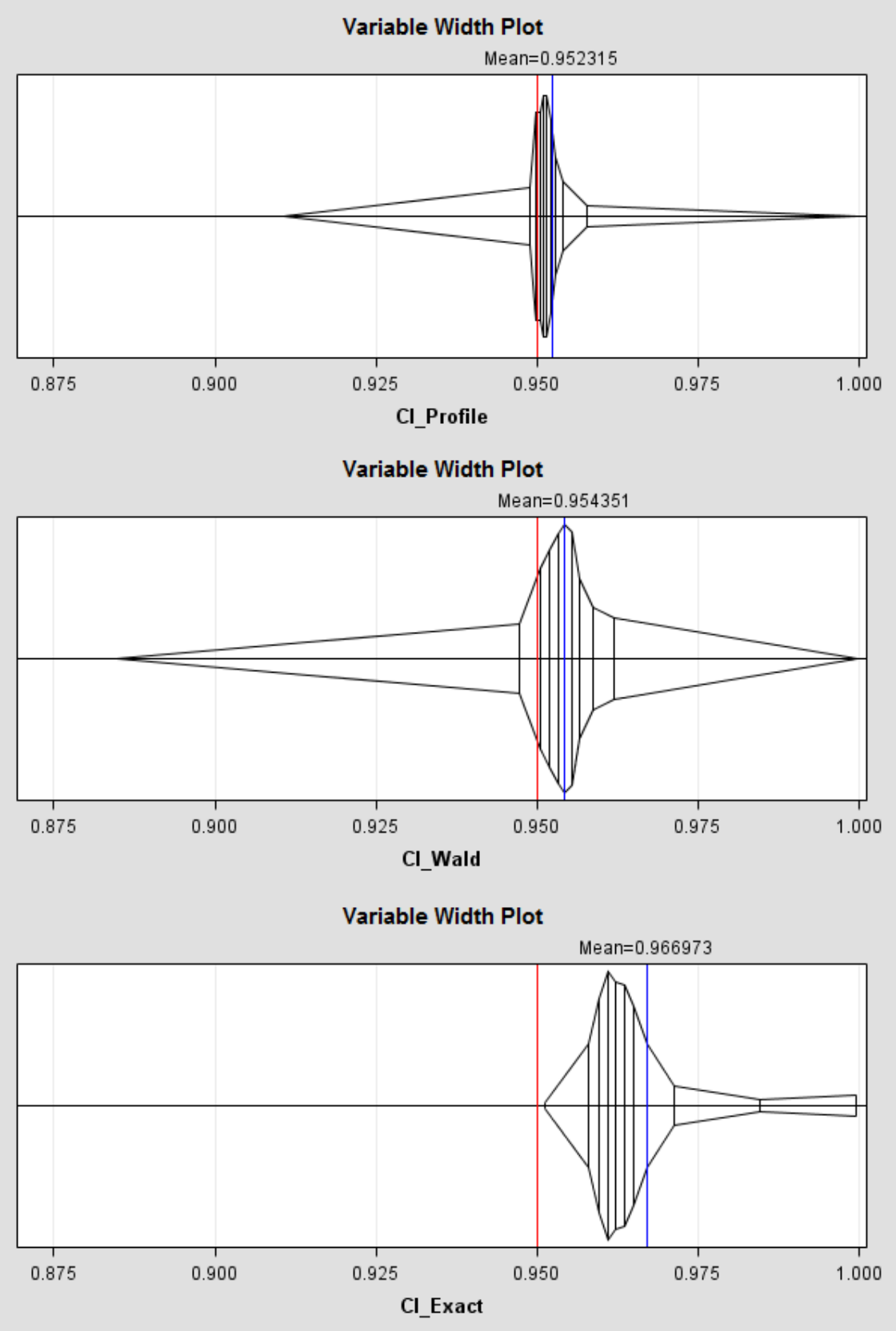

Figure 5: CI Methods comparison using VMP

SAS CODE/MACRO

```
%macro VWPlot(inds=, var=, numPart=10, smooth=0.1);
proc rank data=&inds.(keep=&var.) groups=&numPart. out=_VMPlot1_;
	var &var.;
	ranks __RK__;
run;
proc freq data=_VMPlot1_;tables __RK__;run;

proc sql;
	select (max(&var.)-min(&var.))/&numPart. into :__range__
	from &inds.;
run;
quit;
proc sql;
	select min(&var.) into :__minval__
	from &inds.;
run;
quit;
proc sql;
	select mean(&var.) into :__mnval__
	from &inds.;
run;
quit;
proc univariate data=&inds.;
	var &var.;
	output out=_VMPmedian_ median=__median__;
run;
data _null_;
	set _VMPmedian_;
	call symput("__mdval__",__median__);
run;

data _VMPlot2_;
	set _VMPlot1_;
	__cat__=ceil((&var.-&__minval__.)/(&__range__.*&smooth.));
run;
proc sort data=_VMPlot2_;
	by &var. __cat__;
run;
proc freq data=_VMPlot2_;
	tables __cat__ /out=_VMPlot3_;
run;
data _VMPlot4_;
	merge _VMPlot2_ _VMPlot3_;
	by __cat__;
	__allID__=1;
run;
proc sort data=_VMPlot4_;
	by __allID__ __RK__ &var. ;
run;
data _VMPlot5_;
	set _VMPlot4_;
	by __allID__ __RK__ &var.;
	if first.__RK__ or last.__allID__;
run;
data _VMPlot51_;
	set _VMPlot5_;
	__group__=1;
run;
data _VMPlot52_;
	set _VMPlot5_;
```

```
        percent=-percent;
        __group__=2;
run;
data _VMPlot6_;
        set _VMPlot51_ _VMPlot52_;
run;
proc template;
        define statgraph VWPlot;
                begingraph / designwidth=6in designheight=3in;
                        entrytitle "Variable Width Plot";
                        layout overlay /
                                xaxisopts=(griddisplay=auto_on linearopts=(thresholdmin=0 thresholdmax=0
                                viewmin=0.8 viewmax=1))
                                yaxisopts=(display=none);
                        seriesplot x=&var. y=percent / group=__group__ lineattrs=(color=black PATTERN=1);
                                needleplot x=&var. y=percent / group=__group__ lineattrs=(color=black
                                PATTERN=1);
                                referenceline x=&__mnval__. /lineattrs=(color=blue)
                                curvelabel="Mean=&__mnval__.";
                                referenceline x=0.95 /lineattrs=(color=red);
                        endlayout;
                endgraph;
        end;
run;
proc sgrender data=_VMPlot6_ template=VWPlot ;
run;
%mend VWPlot;
```

CONCLUSION

The variable-width plot is not available in SAS. It can be very useful for data analyses, especially for understanding the data distributions. It can help overcome certain limitations of some existing graphic tools. It conceptually is simple. The implementation using SAS GTL gives SAS user one additional graphical presentation tool for data analysis.

REFERENCES

[1] Lee, J. Jack and Tu, Z. Nora. “A Versatile One-Dimensional Distribution Plot: The BLiP Plot”. *The American Statistician*, 51.4, 353-358, 1997
[2] Warren W. Esty and Jeffery D. Banfield (2003). “The Box-Percentile Pot”. *Journal of Statistical Software*, 8(17), 2003.
[3] Rosenblatt, M. (1956). “Remarks on some nonparametric estimates of a density function”. *The Annals of Mathematical Statistics*, 27 832-837, 1956.

CONTACT INFORMATION

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

Name: Songtao Jiang
Enterprise: Boston Scientific Corporation
Address: 50 Boston Scientific Way
City, State ZIP: Marlborough, MA 01752
Work Phone: 508-683-4432
Fax: 508-683-5642
E-mail: JIANGS@BSCI.com
Web: http://www.bostonscientific.com/us/index.html