Paper 085-2012

Adding Count as a Data Label in a Scatter Plot

Suneetha Puttabasavaiah, Mayo Clinic, Rochester, MN

ABSTRACT

When there are too many data points in a scatter plot, it is useful to have a count label instead of showing all of the overlapping data points. Here are the simple steps to add a statistic as a data label to a graph:

1. Get the required statistic values for variable 1 and 2.

2. Output the required statistic values to a data set.

3. Sort the original data set by variable 1 and 2.

4. Merge the statistic values with the original data set.

5. Use the Datalabel option in SGPlot to show the statistic as a label.

INTRODUCTION

SAS® 9.2 introduced the SGPlot procedure as a way to produce a variety of graphics efficiently. Adding data labels makes a graph more effective as they provide more information about the data. This paper discusses enhancing the basic scatter plot by adding count and percentages as data labels using the ‘datalabel’ option in SGPlot procedure.

The basic scatter plot inserts symbols in the body of the graph where two plotted variables intersect. If there are two or more sets of matching data pairs, the graphical representation of these data does not show the multiple instances. Consider a case where two variables, ‘xvar’ and ‘yvar’, from the dataset ‘dat’ are plotted on a scatter plot. In Figure 1, each marker on the graph, represents multiple overlapping data points, thus it does not provide a complete picture of the data since most of the data are hidden from view. The graph can be enhanced by adding counts and percentages as data labels to the markers representing the data points.

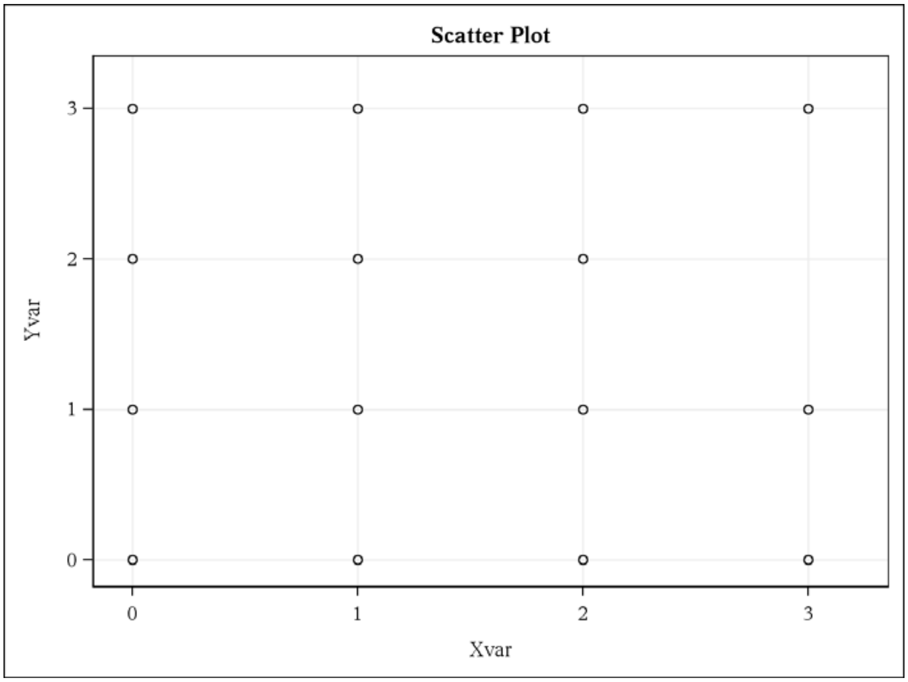

Figure1. Scatter plot of xvar and yvar without data labels.

EXAMPLE 1: ADDING COUNT AS A DATA LABEL

SAS code may be used to add count as a data label to Figure1, above.

Step 1: Get the count value for the label using a procedure such as proc freq and output to a dataset.

```
proc freq data=dat;
  tables xvar * yvar / out=dat_cnt;
  run;
```

Step 2: Sort datasets and add the count values to the ‘dat’ dataset.

```
proc sort data=dat_cnt;
  by xvar yvar;
  run;

proc sort data=dat out=dat_sort;
  by xvar yvar;
  run;

data mrg;
  merge dat_sort dat_cnt;
  by xvar yvar;
  run;
```

Step 3: Use the ‘datalabel’ option in SGPlot to add count as a data label for the markers.

```
title 'Scatter Plot With Count As A Data Label For The Markers’;
proc sgplot data=mrg;
  scatter x=xvar y=yvar / datalabel=count;
  yaxis label="Yvar" grid values=(0 to 3 by 1) offsetmin=0.05 offsetmax=0.05;
  xaxis label="Xvar" grid values=(0 to 3 by 1) offsetmin=0.05 offsetmax=0.05;
  run;
```

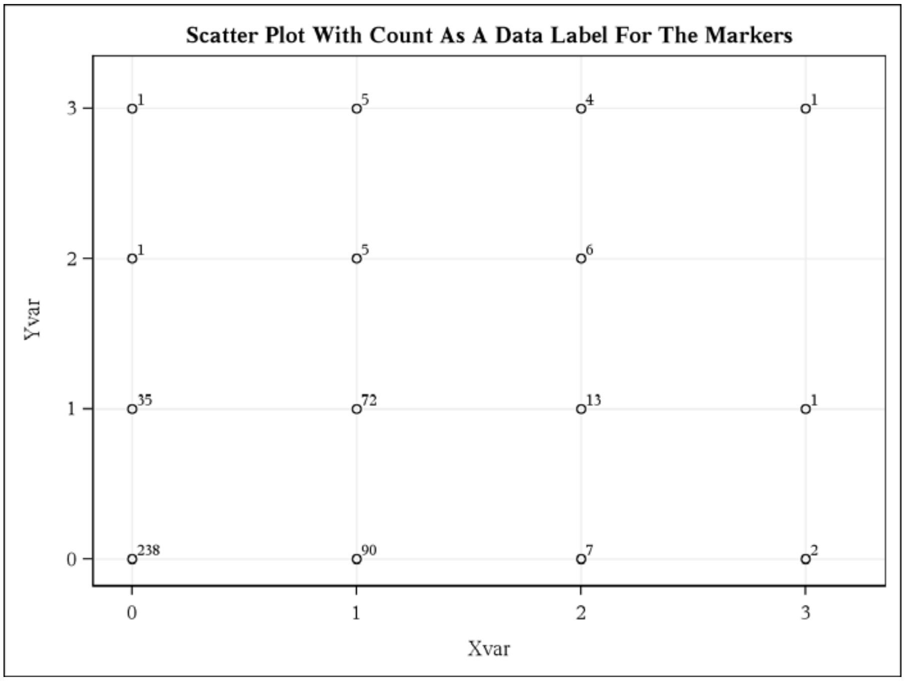

Figure2. Scatter plot of xvar and yvar with count as a data label.

Figure2 provides a more accurate representation of the data as it indicates the number of observations at each marker. For example, we can easily tell there are 238 cases where both xvar and yvar values are 0.

EXAMPLE 2: ADDING COUNT AND PERCENTAGE AS A DATA LABEL

Figure2 can be enhanced by adding percentages. Without the total number of observations (N), percentage information is incomplete. We will calculate total number of observations (N) and add that to the title of the graph.

Step 1: Get the count value for the label using a procedure such as proc freq and output to a dataset.

```
proc freq data=dat;
  tables xvar * yvar / out=dat_cnt;
  run;
```

Note: Step 2 mentioned in Example 1 is optional. This scatter plot can be plotted from the frequency output dataset.

Step 2: Create a dataset with a variable for the data labels by concatenating the count and percentage values. Calculate the total number of observations (N), and store it in a macro variable so it can be inserted in the title.

```
data dat_pct ;
  set dat_cnt nobs=last;    *** Get last observation number **;
  by xvar yvar;

  *** Concatenate count and percentage values **;
  length cnpct $ 15;
  cnpct=cat('  (',count,', ',round(percent,.1),'%',')');

  *** Create a Macro Variable for N  **;
  if _n_=1 then tot=0;
  tot+count;
  if _n_=last then call symput('N',put(tot,3.));
  run;
```

Step 3: Use the 'datalabel' option in SGPlot to add count and percentage as a data label for the markers.

```
Title1 'Scatter Plot With Count And Percentage As A Data Label For The
Markers';
Title2 'Total Number Of Observations (N) = &N';

proc sgplot data=dat_pct;
  scatter x=xvar y=yvar / datalabel=cnpct;
  yaxis label="Yvar" grid values=(0 to 3 by 1) offsetmin=0.05 offsetmax=0.05;
  xaxis label="Xvar" grid values=(0 to 3 by 1) offsetmin=0.05 offsetmax=0.05;
  run;
```

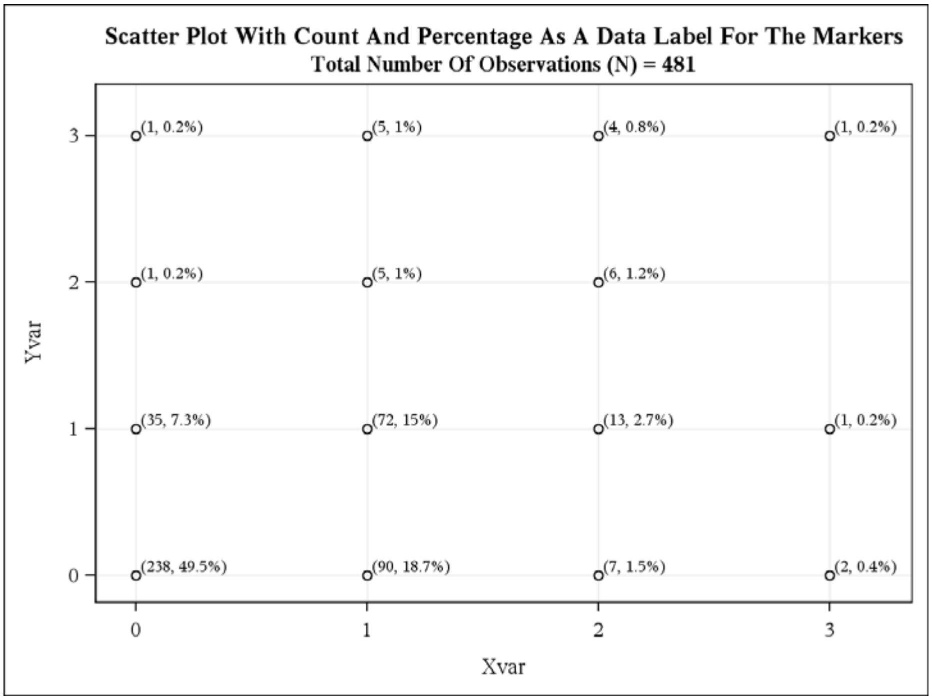

Figure3. Scatter plot of xvar and yvar with count and percentage as data labels.

CONCLUSION

These simple steps discussed above enhance the scatter plot, provide an alternative to data alteration (i.e. jittering) and provide more information to help understand the data. This idea can be used to add any other type of statistic as data labels to a scatterplot.

ACKNOWLEDGMENTS

I would like to thank Pamela Atherton, Paul Novotny and Daniel Satele for their valuable input to this paper and for their review time.

CONTACT INFORMATION

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

Name: Suneetha Puttabasavaiah
Enterprise: Mayo Clinic
Address: 200 First Street SW
City, State ZIP: Rochester, MN 55905
Work Phone: 507-266-2722
E-mail: puttabasavaiah.suneetha@mayo.edu