

More Informative Scatter Plots -- Adding a Third Dimension with Bubbles

Frank E. Kucera, Wonderlic Personnel Test, Inc., Libertyville, IL

ABSTRACT

This paper offers a very simple technique for adding a third dimension to conventional scatter plots produced by the Gplot procedure and the PLOT statement in SAS/GRAPH[®]. A conventional scatter plot may be adequate when you are plotting a limited number of data points. But when you are plotting a large number of data points, there may be many hidden points. These hidden points obscure data distribution within the scatter plot. You can use the variable COUNT (from the optional output data set of the FREQ procedure) in the BUBBLE statement of PROC PLOT to add a frequency dimension where there are hidden data points.

INTRODUCTION

A scatter plot is a good starting point for the analysis of the relationship between two variables in a SAS[®] data set. Scatter plots are suitable for limited data. But my data sets often contain thousands of observations to plot, and multiple points with the same value obscure data concentration within the scatter plot.

This paper looks at the relationship between cognitive ability as measured by the Wonderlic Personnel Test (WPT) and basic verbal and quantitative skills development as measured by the Wonderlic Basic Skills Test (WBST). The WPT is a 50 item test administered in 12 minutes, and scored based on the total number of correct responses. The WBST has two components; 1) the Test of Verbal Skills that has 50 items measuring word knowledge, sentence construction and information retrieval skills, and 2) the Test of Quantitative Skills that has 45 items measuring explicit, applied, and interpretive problem solving skills. Both WBST subtests are administered in 20 minutes, and scored based on item response theory. (IRT) The WBST Composite Score is the simple average of the total Verbal and Quantitative test scores and the score range is 0 to 500. The sample consists of 3,061 examinees 18 years of age or older with 12 years or less of education. Test scores are in data set TESTS, which contains the variables SKILL (WBST score) and ABILITY (WPT score), as well as other demographic information about each examinee. We will graphically evaluate the relationship between cognitive ability and basic verbal and quantitative skills.

CONVENTIONAL SCATTER PLOT

The following SAS/GRAPH code created the conventional scatter plot in Figure 1:

```
proc gplot data=TESTS;  
  plot SKILL*ABILITY;
```

This scatter plot has 3,061 points. Many points represent several occurrences of the same SKILL and ABILITY. The plot indicates that SKILL increases with ABILITY. The data cloud, however, is

very wide. Are the data points distributed evenly within the cloud or are they concentrated in one area?

ALTERNATIVE BUBBLE PLOT

The solution is Figure 2, a bubble plot that is created simply by adding minimal code as follows:

```
proc freq data=TESTS;  
  tables SKILL*ABILITY  
  / noprint out=BUBS (keep=SKILL ABILITY COUNT);
```

```
proc gplot data=BUBS;  
  bubble SKILL*ABILITY=COUNT / bscale=area bsize=10;
```

PROC FREQ creates the optional output data set BUBS containing the variable COUNT, which is the frequency of occurrence of each combination of SKILL and ABILITY. In the BUBBLE statement, COUNT controls the size of a bubble for each combination of SKILL and ABILITY. The BSCALE= option determines whether the radius of the bubble or the area of the bubble will control the bubble scaling proportion. The default value is area. The BSIZE= option specifies an overall scaling factor for the bubbles, indicating the size of the largest bubble. The default value is 5. (SAS/GRAPH Software: Usage, Version 6, First Edition.)

In this example, the bubble plot provides the following additional information about the relationship between SKILL and ABILITY:

- Most ABILITY scores are between 10 and 25.
- SKILL scores are distributed with a central tendency for ABILITY scores within this range.
- For nearly every SKILL level, there is a wide range of ABILITY, but ABILITY has a central tendency as well.

CONCLUSION

For scatter plots with a large number of points, a bubble plot is well worth the minimal amount of additional programming required considering the magnitude of additional information provided.

REFERENCES

SAS Institute Inc. (1991), SAS/GRAPH[®] Software: Usage,, Version 6, First Edition, Cary, NC: SAS Institute Inc.

SAS and SAS/GRAPH are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Frank E. Kucera
Wonderlic Personnel Test Inc.
1509 N. Milwaukee Ave.
Libertyville, IL 60048
(847) 247-2417 , e-mail address: reserach@wonderlic.com

