A General Family of Related Scatterplots including the Bubble Plot Frequency Matrix with an R Function for their Automatic Generation via Adaptive Graphics

David W. Gerbing
School of Business Administration
Portland State University

Abstract
A family of related 1- or 2-dimensional scatterplots and related statistical analyses result from any combination of continuous or categorical variables: the traditional scatterplot of two continuous variables, a bubble (balloon) scatter plot from two categorical variables, a means plot from a categorical variable paired with a continuous variable, and a Cleveland dot plot as a scatterplot that pairs a continuous variable with a categorical variable. For multiple plots on the same graph, specify multiple x-variables or y-variables. Summarize univariate distributions with either a 1-dimensional scatter plot of a continuous variable, or with a 1-dimensional bubble plot for a categorical variable as a more compact replacement of the traditional bar chart. From the specification of multiple categorical x-variables that share the same response scale, generalize the later to a matrix of 1-dimensional bubble plots here called the bubble plot frequency matrix. Implement this conceptualization with an R function that invokes what is called here adaptive graphics to select the appropriate analysis and display of the appropriate scatter plot from the identical simple syntax, for variables X and Y, ScatterPlot(X) and ScatterPlot(X,Y), or with abbreviations sp(X) and sp(X,Y).

Introduced by Francis Galton in the 19th century (Friendly & Denis, 2005), the scatter plot is one of the premier techniques for the visualization of the relationships between continuous (numeric) variables. Here the concept of the scatter plot is generalized to family of scatter plots that includes any combination of continuous and categorical variables in 1 or 2 dimensions. Most of these different plots are well known, such as the bubble (balloon) plot in which the size of each plotted point represents a characteristic such as its corresponding joint frequency. What is provided here, however, in addition to some new forms of the scatter plot, is the conceptualization of these various scatter plots as systematic expressions of a general scatter plot syntax for implementation.

To implement, the general scatter plot function introduced here, ScatterPlot, abbre-
viated \textit{sp}, is part of the lessR contributed package (Gerbing, 2016) to the free, cross-platform R system for data analysis. The purpose of lessR is to simplify the procedures of data analysis from basic bar charts through least-squares and logistic regression analyses. The operative concept introduced here is adaptive graphics, perhaps more generally expressed as adaptive analyses, in which the underlying software analyzes the structure of the data submitted as input and then chooses the specific appropriate analysis consistent with that structure. Obtain any instance of the general family of scatter plots with the same function call, for variables X and Y, with either \texttt{sp(X)} for a 1-dimensional scatter plot, or \texttt{sp(X,Y)} for the 2-dimensional version. The full name \texttt{ScatterPlot} can also be invoked but here the abbreviation is used.

The primary aspect of the data that structures the analysis is continuous versus categorical variable types. A categorical variable may be formally defined as an R factor with specific categories called levels. Or, invoking a lessR extension, integer valued variables may also be analyzed as categorical, depending on the specified number of unique values of the variable.

Read the data. Following is a systematic description of the family of related scatter plots, each illustrated with an example, so reading the data precedes each analysis. With lessR a data table is read from an external file into an R data table with the \texttt{Read} function. Usually read into a data table with the name of mydata, the default data table name for the lessR functions.

To browse for a data file on the local computer system:

\begin{verbatim}
> mydata <- Read()
\end{verbatim}

Or a data file can be identified by its path name or, if on the web, by its URL:

\begin{verbatim}
> mydata <- Read("http:\lessrstats.com\data\Employee.csv")
\end{verbatim}

To maintain simplicity, invoking the same \texttt{Read} syntax reads data files formatted as csv text files, tab-delimited text files, Excel files, SPSS files, SAS files, and native R files. Again, the software adapts to the situation, choosing which internal read function to invoke, freeing the analyst from these non-substantive decisions.

Traditional 2-dimensional scatter plot. The traditional scatter plot applies to a pair of continuous variables, with each point the physical mapping of the two corresponding data values for a specific case, or row of data. By default lessR plots the points with transparency of the fill color to accommodate limited over-plotting.

\begin{verbatim}
> sp(Years, Salary)
\end{verbatim}

The inclusion of the 0.95 data ellipse and the best-fit \texttt{loess} line is straightforward, illustrated in Figure 2.

\begin{verbatim}
> sp(Years, Salary, ellipse=TRUE, fit.line=TRUE)
\end{verbatim}

Other values for the data ellipse, including multiple values, can be specified, such as \texttt{ellipse=c(.7,.8,.9)} to simultaneously provide the 0.7, 0.8 and 0.9 data ellipses. To provide the least-squares fit line, invoke \texttt{fit.line=ls}.

The corresponding statistical analysis is for the correlation coefficient.
Figure 1. Traditional 2-D scatter plot with two continuous variables.

Figure 2. Traditional 2-D scatter plot with two continuous variables.

>>> Pearson's product-moment correlation

Years: Time of company employment
Salary: Annual salary

Number of paired values with neither missing, n = 36

Sample Correlation of Years and Salary: r = 0.852

Alternative Hypothesis: True correlation is not equal to 0
  t-value: 9.501, df: 34, p-value: 0.000
95% Confidence Interval of Population Correlation
Lower Bound: 0.727   Upper Bound: 0.923
> sp(Years, Salary, kind="bubble", bubble.size=.1, bubble.counts=FALSE)

![2-D scatter plot with two continuous variables with points plotted as bubbles.](image)

Figure 3. 2-D scatter plot with two continuous variables with points plotted as bubbles.

Categorical X-variable, Continuous Y-variable. The scatterplot that results from the first variable listed as categorical and the second variable continuous is a means plot. The characteristic of this scatterplot is that the plotted points are aligned in vertical columns, one column for each value of the categorical variable, illustrated in Figure 4.

> sp(Dept, Salary)

![Means plot version of a 2-D scatter plot.](image)

Figure 4. Means plot version of a 2-D scatter plot.

Salary: Annual salary
- by levels of -
Dept: Department employed
### Continuous X-variable and Categorical Y-variable.

The result is the Cleveland dot plot (Cleveland, 1993).

```r
> mydata <- mydata[1:15,]
> nm <- row.names(mydata)
> sp(Salary, nm)
```

![Cleveland dot plot](image)

\textit{Figure 5}. Cleveland dot plot.

2-D scatterplot for categorical data. When both variables are categorical, there is the typical problem of over-plotting. For example, a scatterplot of the data values for two 5-pt Likert scales results in only $5 \times 5 = 25$ possibilities, so most if not all paired data values occur multiple times for a reasonably large sample size. The more appropriate plot in this circumstance is the bubble plot, shown in Figure 6. The corresponding frequency for each set of paired values is also displayed for the larger, more meaningful frequencies, that is, those that can be displayed within the bubble.

The illustration is of 6-pt Likert scales, scored 0 to 5, for responses from the Christie and Geis (1970) Mach IV scale from Hunter, Gerbing, and Boster (1982).

```r
> sp(m06, m07)
```

The resulting statistical output is analysis of the two-way cross-tabulation table.

m06: Honesty is the best policy in all cases
ADAPTIVE GRAPHICS FOR SCATTER PLOTS

Figure 6. Bubble plot with two categorical variables.

- by levels of -
  m07: There is no excuse for lying to someone else

Joint and Marginal Frequencies
-----------------------------

<table>
<thead>
<tr>
<th></th>
<th>m06</th>
<th>m07</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>24</td>
<td>7</td>
<td>6</td>
<td>18</td>
<td>2</td>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>14</td>
<td>30</td>
<td>13</td>
<td>24</td>
<td>2</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>16</td>
<td>12</td>
<td>2</td>
<td>43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>5</td>
<td>56</td>
<td>16</td>
<td>93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>34</td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>18</td>
<td>47</td>
<td>63</td>
<td>44</td>
<td>121</td>
<td>58</td>
<td>351</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cramer’s V: 0.380

Chi-square Test: Chisq = 253.103, df = 25, p-value = 0.000

>>> Low cell expected frequencies, so chi-squared approximation may not be accurate

The default value of n.cat for the ScatterPlot function is 10.

This version of the scatterplot graph can also be enhanced with value labels, here also illustrated with the least squares regression line.

> LikertCats <- c("Strongly Disagree", "Disagree", "Slightly Disagree", "Slightly Agree", "Agree", "Strongly Agree")
> sp(m06, m07, value.labels=LikertCats, fit.line="ls")
Strongly Disagree  Disagree  Slightly Disagree  Slightly Agree  Agree  Strongly Agree

Figure 7. Bubble plot with two categorical variables and value labels.

Continuous X-variable 1-D Scatterplot.

> sp(Salary)

Figure 8. 1-dimensional scatterplot.

Categorical X-variable 1-D Scatterplot. As with the 2-dimensional application, the scatterplot of a single categorical variable is subject to over-plotting of the comparatively few response categories. As such, the scatterplot is presented as a bubble plot.

> sp(m07)

--- m07: There is no excuse for lying to someone else ---

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>64</td>
<td>87</td>
<td>43</td>
<td>93</td>
<td>47</td>
<td>351</td>
</tr>
<tr>
<td>0.048</td>
<td>0.182</td>
<td>0.248</td>
<td>0.123</td>
<td>0.265</td>
<td>0.134</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Chi-squared test of null hypothesis of equal probabilities

Chisq = 70.556, df = 5, p-value = 0.000
Bubble Plot Frequency Matrix. The bubble plot frequency matrix (BPFM) is a generalization of the 1-D scatter plot of a categorical variable, applied to a range of categorical variables all with the same set of response categories, such as a common Likert scale. R provides several methods by which to specify a range of variables.

> sp(c(m06,m07,m09,m10))

### Frequencies of Responses by Variable

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>m06</td>
<td>18</td>
<td>47</td>
<td>63</td>
<td>44</td>
<td>121</td>
<td>58</td>
<td>351</td>
</tr>
<tr>
<td>m07</td>
<td>17</td>
<td>64</td>
<td>87</td>
<td>43</td>
<td>93</td>
<td>47</td>
<td>351</td>
</tr>
<tr>
<td>m09</td>
<td>6</td>
<td>10</td>
<td>16</td>
<td>35</td>
<td>84</td>
<td>200</td>
<td>351</td>
</tr>
<tr>
<td>-----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>m10</td>
<td>4</td>
<td>7</td>
<td>32</td>
<td>50</td>
<td>110</td>
<td>148</td>
<td>351</td>
</tr>
</tbody>
</table>
References