Cornell Statistical Consulting Unit

Data Visualization: Choosing a Figure Type

Erika Mudrak

1 Introduction

Data visualization plays an important role in the exploratory phase and in reporting the results of a statistical analysis. It seeks to speed the process of insight as it is easier to communicate a large amount of data via images rather than words or numbers. This newsletter provides a general overview of aspects to consider and available choices when designing graphs to convey research results to a scientific or lay audience. Graphs can describe the distribution of a single variable (univariate), or can show the relationship between two (bivariate) or more (multivariate) variables. In general the type of data (continuous or categorical) guides the choice of graph but within each of these categories many options are possible.

2 Univariate Plots

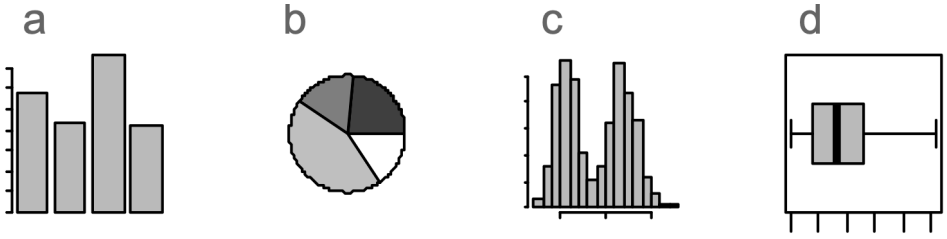

Figure 2.1: Commonly used univariate plots: (a) bar chart; (b) pie chart; (c) histogram; (d) box plot

2.1 Categorical variables

The frequency of each class can easily be shown in a bar chart (Figure 2.1(a)), making it easier to see the difference in number of observed variables in each group. Pie charts (Figure 2.1(b)) are popular, but should be used only to compare a slice to the rest of the pie (not slice to slice), and are best used with fewer than six categories.

2.2 Continuous variables

The distribution of continuous values can be easily visualized with a histogram (Figure 2.1(c)). The information can be further summarized with a box plot (Figure 2.1(d)) which shows the median, 1st and 3rd quartiles, and extreme values of the data.

3 Bivariate plots

It is standard practice to show the dependent variable (DV) on the y-axis and the independent variables (IV) on the x-axis.

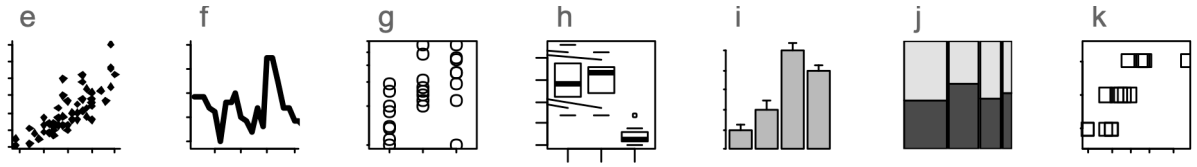

Figure 3.1: Commonly used bivariate plots: (e) scatter plot; (f) line graph; (g) dot plot; (h) box plots; (i) bar charts; (j) mosaic plot; (k) dot plot with horizontal orientation.

3.1 Continuous DV and Continuous IV

If observations are independent from each other, try a scatter plot (Figure 3.1(e)), with each observation represented as a point. If the IVs are sequential in some way (time, growth stage), you can connect each point with a line to make a line graph (Figure 3.1(f)).

3.2 Continuous DV and Categorical IV

If the data set is small, you can show all the observations grouped by category in a dot plot (Figure 3.1(g)). For larger data sets that often require summary statistics to show meaningful patterns, available possibilities include box plots (Figure 3.1(h)) that show the distribution of data through quartiles, and bar charts (Figure 3.1(i)) that show group means along with error bars. Bar charts (with or without error bars) are intended to compare magnitude of values, so it is important to always start the y-axis scale at zero. Box plots on the other hand are better for comparing data spread, so choose y-axis limits to maximize the space the box plots take up on the graph.

3.3 Categorical DV and Categorical IV

When the dependent variable is categorical, visualization is less straightforward. The relationship between two categorical variables is often analyzed via cross tabulation. A way to visualize such a table is through a mosaic plot (Figure 3.1(j)).

3.4 Categorical DV and Continuous IV

An appropriate choice is a dot plot with a horizontal orientation (Figure 3.1(k)).

4 Multivariate Plots

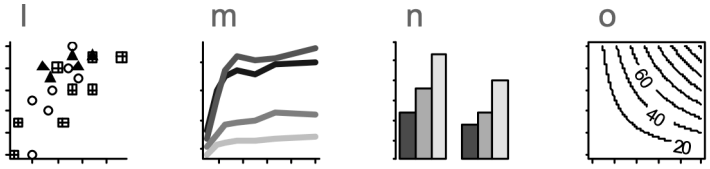

Figure 4.1: Commonly used multivariate plots: (l) scatter plot with different symbols or colors; (m) line graph with multiple lines; (n) grouped bar chart; (o) contour plot.

4.1 Continuous DV, Continuous and Categorical IVs

Add an additional categorical IV to a scatter plot with different symbols or colors (Figure 4.1(l)), and in line graphs by showing multiple lines (Figure 4.1(m)) with different colors or styles.

4.2 Continuous DV, 2 Categorical IVs

Show the outcome of more than one categorical variable by nesting groups, such as in a grouped bar chart (Figure 4.1(n)), which shows two types of categorical variables by using labeling and color.

4.3 Continuous DV, 2 Continuous IVs

The relationship between three continuous variables can be done nicely with a contour plot (Figure 4.1(o)), where the two IVs are the X and Y axes, and the DV is shown as contours similar to elevation in a topographic map. This 2D representation of three dimensions of data avoids the perspective issues inherent in 3D plotting.

4.4 Any DV, 3+ IVs

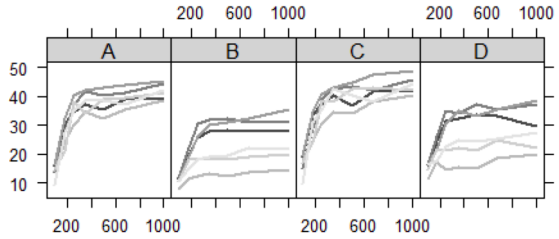

Figure 4.2: Trellis chart with four line graphs, one for each value of a categorical variable.

If you have many IVs, it can easily get very confusing to try to display all the data in a single graph like those we have shown above. Trellis charts (Figure 4.2, also known as lattice charts, panel charts and small multiples, are a good solution. Trellis charts repeat simpler charts over a series of one or two conditions in a grid layout. The use of consistent scales allows for easy comparisons between conditions.

5 References

Wilkinson, Leland. The Grammar of Graphics (Statistics and Computing). Springer-Verlag. Berlin. 2005.

Tufte, Edward R., 1942-. The Visual Display of Quantitative Information. Cheshire, Conn. :Graphics Press, 2001.

Created Fall 2014. Last updated April 2022.