Observations from the Winners of the 2013 Statistics Poster Competition – Praise and Future Improvements

Jürgen Symanzik*, Utah State University, Department of Mathematics and Statistics, Logan, UT 84322-3900, symanzik@math.usu.edu

Naomi B. Robbins*, NBR Graphs, Wayne, NJ 07470-6523, naomi@nbr-graphs.com

Richard M. Heiberger*, Temple University, Department of Statistics, Philadelphia, PA 19122-6083, rmh@temple.edu

*Member of the Special Task Force of the Statistical Graphics Section of the American Statistical Association (ASA)

1. Introduction

The goal of this article is to make readers aware of some of the excellent plots produced by the winners of the 2013 Statistics Poster Competition, accessible at http://www.amstat.org/education/posterprojects/2013posters.cfm. This annual competition is jointly organized by the American Statistical Association (ASA) and the National Council of Teachers of Mathematics (NCTM). We were impressed by the variety and quality of statistical graphics produced by these young researchers. In some rare cases, however, we thought that some alternative representation would have helped to further improve the overall message communicated by these posters. In the next section, we will outline what we really liked – and how some of the charts easily could be transformed into even more meaningful charts.

2. From Good Graphics to Even Better Graphics

We generally commend the creators of these figures for adding titles, labels, and legends to their charts. Very well done!

2.1 Pie Charts

Pie charts can be found almost everywhere – on the web, in newspapers, and in several of the winning posters. However, there exist a few rules that should be followed when creating pie charts.

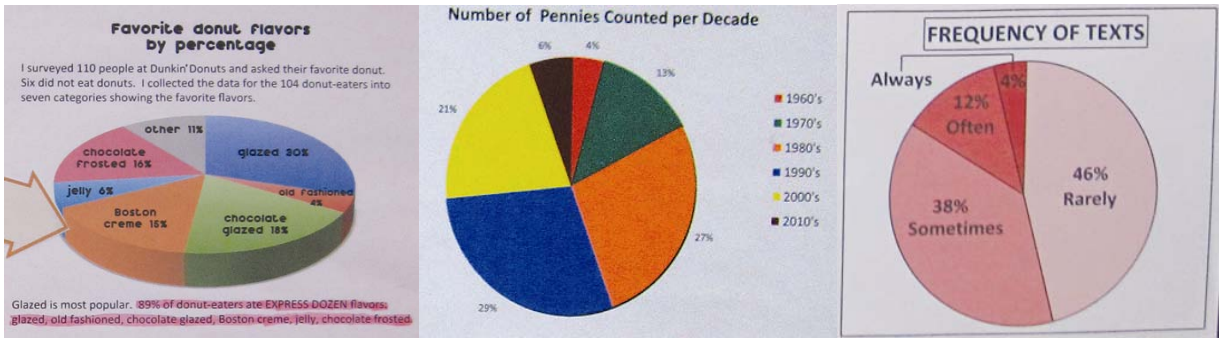

Figure 1: Three different pie charts: (left) unsorted 3D pie chart, (center) unsorted 2D pie chart, and (right) sorted 2D pie chart.

In general, 3-dimensional (3D) plots are very misleading as areas in the front appear to be much larger than areas in the back of the chart, in particular if a rim is plotted as in Figure 1 (left). At first glance, the green area in the front right appears to be much bigger than the blue area in the rear right of Figure 1 (left). Therefore, 2-dimensional (2D) pie charts are preferred over 3D pie charts. In addition, the order in which the individual categories are arranged in the slices of the pie chart is important.

An ordered pie chart such as in Figure 1 (right) makes it much easier to compare categories than an unordered pie chart such as in Figure 1 (center). There exist a few different rules how best to order a pie chart. One of those rules has been followed in Figure 1 (right): Start on top (at the noon position) and fill in the categories in a clockwise direction, starting with the largest percentage and ending with the smallest percentage. In case of an "other" category, this is often filled in last. This makes it relatively easy to see which categories jointly make up 25%, 50%, and 75% of the data. Moreover, due to the sorting, it is immediately clear which category contains a (slightly) larger percentage of the data. This is not always obvious in an unsorted pie chart such as the one in Figure 1 (center) where it is not clear whether the orange or the blue area is bigger.

2.2 (Stacked) Bar Charts

Bar charts and stacked bar charts are another type of widely used charts. However, there exist a few recommendations how to make bar charts more effective.

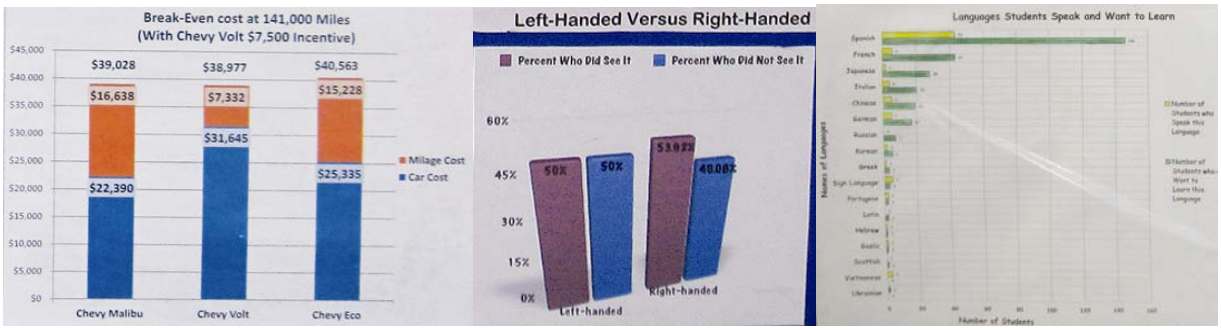

Figure 2: Three different bar charts: (left) stacked bar chart, (center) 3D side-by-side bar chart, and (right) sorted side-by-side bar chart.

Stacked bar charts can be useful when comparing percentages. However, they are difficult to interpret for human readers if we look at totals, such as counts or a monetary value such as in Figure 2 (left). In this chart, it is obvious that the rightmost stacked bar is higher than the leftmost stacked bar. However, is this only due to the larger amount of the blue area at the bottom? What can we say about the orange area? Is this also larger in the rightmost bar (compared to the leftmost bar), about the same, or even smaller?

Instead of using stacked bar charts, in many situations side-by-side bar charts are preferred. However, similar to 3D pie charts, we should also avoid 3D bar charts, such as the one shown in Figure 2 (center). Here, the lack of a common baseline makes it difficult to visually compare the lengths of the four bars. Typically, we only use the lengths of the bars (and not their area) when we compare multiple bars in a bar chart. This is difficult when the bars do not start on a common horizontal baseline. About how much smaller is the rightmost blue bar in this chart, compared to the leftmost purple bar in this chart? The 3D effect makes it impossible to answer this question visually (without looking at the printed numbers).

Often, 2D side-by-side bar charts are most effective, such as the one in Figure 2 (right). Moreover, such side-by-side charts are even more effective when the bars are sorted from highest to lowest (or vice versa) by the most important variable. Here, this seems to be the language students want to learn (shown in green). Due to the sorting, it is immediately evident that about three times as many students want to learn Spanish, compared to the second-ranked language which is French. Side-by-side bar charts also allow us to spot unusual patterns relatively quickly: More students know sign language (five), compared to only three students who want to learn it.

It does not hurt to remind readers that we technically have to compare the size of the areas in bar charts. But, as mentioned above, this is simplified to comparing the lengths of the bars (given that they are all equally wide). Therefore, it is a must that all bars start at zero since all

lengths start at zero. Everything else would be some major kind of cheating with charts. This rule was generally followed in the posters.

We do not want to go into an in-depth discussion of the use of colors here. However, the reader should be reminded that colors may appear differently under different lighting and shading conditions, when projected onto a wall or seen on a computer screen, or when drawn on different types of paper. Therefore, some readers may have problems seeing the yellow bars in Figure 2 (right) while for other readers, these bars are easily discernible.

2.3 Graphs for Comparisons

It is often necessary to compare data from different groups or from different years. A few suitable charts should be mentioned here.

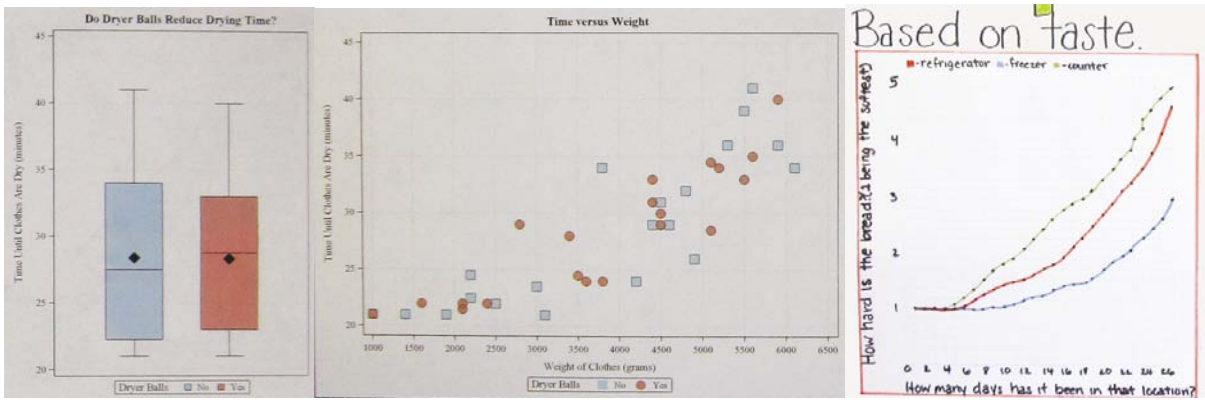

Figure 3: Excellent charts for comparison: (left) side-by-side boxplots, (center) scatterplot with an additional categorical variable, and (right) line graph (time series plot) of three levels of a categorical variable.

In Figure 3 (left), side-by-side boxplots of a quantitative variable are shown for two levels of a categorical variable. This makes it easy to compare the medians, first and third quartiles, minimum, maximum, and possible outliers (if any) for the different levels of a categorical variable. This could be gender, different cities or countries, or different age groups. It is important that the boxplots are drawn using the same scale (which is done here).

Figure 3 (center) shows a scatterplot of two quantitative variables. The different levels of a categorical variable (representing “yes” and “no”) are shown via different colors and plotting symbols (“red circles” and “blue squares”, respectively). Overall, an exponential increase can be seen in this chart. This scatterplot suggests that it makes no major difference whether dryer balls are used or are not used. A fourth categorical variable could be added via different plotting symbols (such as x, +, or o), while a fourth quantitative variable could be added via

different sizes of the plotting symbols (where the area of the plotting symbol, and not its diameter, should be proportional to the numeric value).

Figure 3 (right) shows a line graph (time series plot) for three levels of a categorical variable for 27 time points. The data have been standardized to 1 at day 0. Typically, we standardize to 1 or 100%. While there is hardly any difference among the hardness of bread for the different factor levels during the first three days, the three factor levels result in different increases, starting with day 4. While two of these increases could be described as quadratic or exponential (for the red and blue factor levels), the third factor level (green) shows a fairly linear increase of hardness.

2.4 Other Noteworthy Charts

There are a few other (almost) unique charts that should be mentioned here.

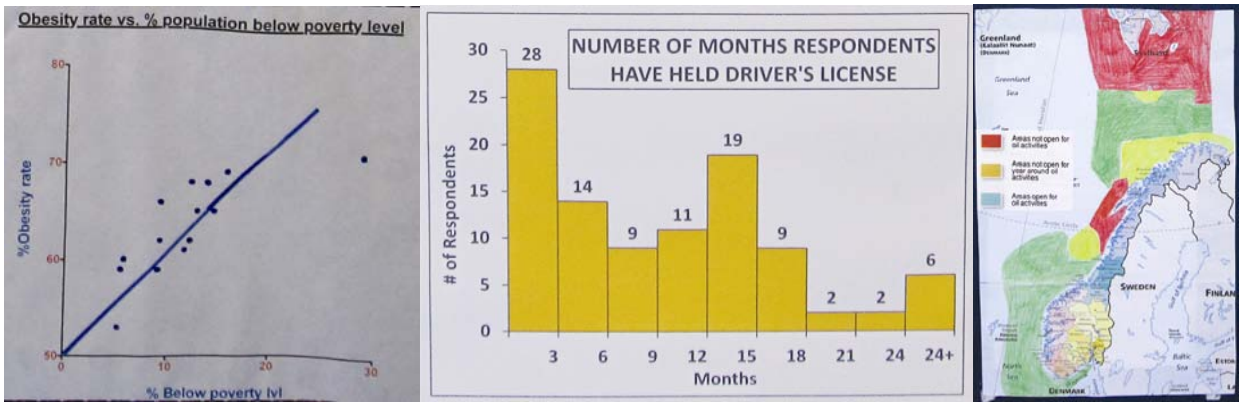

Figure 4: (left) Scatterplot with regression line, (center) histogram, and (right) choropleth map.

When we want to describe the relationship between an explanatory variable (shown on the horizontal axis) and a response variable (shown on the vertical axis), we frequently add some smooth line to a scatterplot. This can be a least-squares regression line as shown in the scatterplot in Figure 4 (left), a polynomial, exponential, or logarithmic fit to the data, or even a smoothed curve if no other relationship between the two variables can be easily found.

Histograms, such as the one shown in Figure 4 (center) are an effective way to summarize one quantitative variable. It is important that not too many and not too few classes are used. Several recommendations exist in the literature how many classes should be used. Moreover, it is recommended that all classes are equally wide. If this is not possible, it is necessary to translate the data into a density scale where the vertical axis represents the count (or percentage) per unit on the horizontal axis.

This histogram shows some interesting bimodal distribution with peaks at 0-3 months and at 12-15 months the reader may not have expected in advance. The fact that the population in this study consists only of high-school students at the junior and senior level may serve as a possible explanation for this bimodality.

Whenever data have a geographic (spatial) component, a map is of high importance, such as the regions of the Atlantic Ocean shown to the west and north of Norway in the choropleth map in Figure 4 (right). A choropleth map is a map where the statistical information for a geographic region (such as countries or administrative districts) is shown via color (or shading). Choropleth maps are frequently used to show income in different sub-regions, election outcomes, or climatic data (such as temperatures and precipitation). It is much easier to comprehend geographic (spatial) information when shown on a map, compared to only a textual description of the geographic locations.

3. Conclusion

A variety of excellent charts have been used in the winning posters of the 2013 Statistics Poster Competition in all age groups. However, with respect to pie charts and bar charts, some versions of these charts are preferred to other versions of the same charts. We would like to congratulate the authors of these posters once more for their excellent statistical contributions.

For More Information: Books and Articles

- Figures from http://www.amstat.org/education/posterprojects/2013posters.cfm (see also Amstat News, August 2013, pp. 26 – 33, http://magazine.amstat.org/wp-content/uploads/2013an/August2013.pdf). Photos are courtesy of the American Statistical Association (ASA).
- Robbins, Naomi B. 2013. *Creating More Effective Graphs*, Chart House, Ramsey, NJ (reprinted from Wiley 2005)
- Su, Yu-Sung. 2008. "It's easy to produce chartjunk using Microsoft®Excel 2007 but hard to make good graphs*," Computational Statistics and Data Analysis 52,* pp. 4594 – 4601.
- Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*, 2nd edition, Graphics Press, Cheshire, CT. (First edition 1983)
- Wainer, Howard. 2000. *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. Psychology Press, London. (Reprinted from Copernicus 1997)

For More Information: Blogs

Camoes, Jorge. *ExcelCharts*, http://www.excelcharts.com/blog.

Peltier, Jon. *Peltier Tech Blog – Peltier Tech Excel Charts and Programming,* http://www.peltiertech.com/WordPress.

Robbins, Naomi B. *Effective Graphs*. http://www.forbes.com/sites/naomirobbins/.