

Using Statistical Graphics to Understand Your Data (Not Just Present Results)

David J. Pasta, Ovation Research Group, San Francisco, CA

ABSTRACT

Figures and graphs are often used to present the results from statistical analyses. As business graphics become increasingly common, there is an expectation that much (if not most!) numerical data will be presented in graphical form. This is, generally, a good thing (although more attention to the principles of good graphical presentation would be welcome). But statistical analysts seem to have lost sight of the value of using graphical displays to better understand the data they are analyzing. With the advent of ODS Graphics in Version 9, SAS® has made it easier to get the "basic" analytical graphics. This paper gives examples of graphical displays that lead to insights into the data (as opposed to displaying something that is already understood). Several examples are given where graphs were used as an analytical tool.

INTRODUCTION

Business graphics are in increasingly widespread use. It seems fair to characterize the purpose of business graphics as the communication of information to the reader/viewer. This is an important role for graphical displays, and one that can benefit from continued attention. There is even an entire section of this conference devoted to "Data Presentation." But this is the "Data Analysis and Statistics" section – why are we discussing statistical graphics? We are discussing a different use for statistical graphics, one that seems to get less and less attention as time goes on: the use of graphical displays to better understand the data.

Many discussions of graphical displays in the last twenty years begin in the same place: with Ed Tufte's magnum opus, *The Visual Display of Quantitative Information* (1983). This was not the first work on the principles of graphical displays, but it was a major landmark. Tufte has written two additional books, *Envisioning Information* (1990), and *Visual Explanations* (1997). Time spent reading (or re-reading) these books will be well spent, indeed. Tufte lays out very sound principles of graphical displays and gives many examples, both good and bad. He is a gifted expositor, and the books sometimes read like novels.

Another good source of information on and examples of excellent statistical graphics is Michael Friendly's book, *Visualizing Categorical Data* (2000). Not only does Friendly provide very good advice on many aspects of graphing categorical data, but also he provides SAS macros for producing a wide array of sophisticated graphics.

Both Tufte's work and Friendly's provide important guidelines for creating good statistical graphics that are apropos whether the purpose of the graphic is to communicate information or to better understand the data. In fact, there really isn't much distinction between the two, is there? Isn't it true that a good graphical display is a good graphical display?

Well, yes. But in the practical world (which is where I spend most of my time), it's sometimes "not worth the trouble" to produce a graph for every little relationship you see, for every variable's distribution, for every point estimate and associated confidence interval. But it only takes a few seconds to scan a large number of well-designed graphs to see if there are any surprises, and if in fact there **are** any surprises, the time will have been very well spent. One of the things we need to do, as data analysts, is convince ourselves that it **is** "worth the trouble" – or at least worth the trouble more often than we've been doing it.

The advent of ODS Graphics in Version 9 marks a recognition by SAS that the statistical graphic is a fundamental part of statistical analysis and should be easy to obtain. SAS also understood that there would be the need to customize those graphics to reflect different tastes and purposes, so there is a mechanism for doing so. It is modeled on the corresponding mechanism for ODS tables, using templates and principles of object-oriented programming to allow the inheritance of attributes. Let's hope that ODS Graphics lowers the barrier for graphics to the point that it's almost always "worth the trouble" to get basic analytical graphics. They are very useful tools for understanding data.

The purpose of this presentation is to provide some examples where the right graph can provide insights into data analysis that would be hard to obtain another way. Although the examples often use artificial data, the examples are based on situations that arose during analysis of real data. The real data are not always available because they are proprietary or difficult to access or both. The examples are mostly of scatterplots rather than charts. Charts are very common for communicating information but I generally don't find them as useful for providing insight into data. There are exceptions, however.

We begin with some basic ideas that have general applicability for nearly everyone doing data analysis and move through more unusual examples to some examples that might be characterized as esoteric. The specifics of those last examples may be less important than the idea that there might be a similar insight hiding in **your** data if you try to think about it in a different way.

WHAT PLOTTING SYMBOLS SHOULD I USE?

Let's start with something pretty simple. You're doing a scatter plot of some data – what plotting symbol should you use? The default is the "plus" sign, and that works pretty well for a lot of purposes. For display, you might consider a large circle or dot (a filled-in circle) or maybe an asterisk to get attention. Or maybe something interesting – a heart, maybe, or a little palm tree? SAS provides a set of special symbols, and there are some good symbols among them. You can also specify another font and choose a character that way – that's where you'd find the palm tree (the cartographic font, letter L). For basic plotting, the symbol I like best is the X.

In my data, there are often ties in the x-values, or in the y-values, or both. When you have a lot of ties in one of the dimensions, if you use the + symbol you get less ink when the points are nearby in one direction and tied in the other. This has a tendency to mislead your eye into seeing fewer data points than actually are there. The dot is particularly bad if you have a lot of data near each other, because you end up with just a solid mass of ink even at relatively low density, making it harder to distinguish the medium- and high-density areas from each other. The circle (and its kin, the triangle, square, and diamond) is nice and has its uses, but it is hard to see the center and therefore hard sometimes to see whether a point lies above or below (or left or right) of a reference value. The point is usable but sometimes rather hard to see, even with a large size specified.

With the X, you can clearly see the location of the point and nearby values are usually not in exactly the position to have the arms of the symbol overlap, so each separate point is usually visible unless the data are very dense indeed.

When there are several values to distinguish on a graph, I usually make use of the circle, square, and triangle first and the diamond next. The star (which is obtained by specifying VALUE==, one of my favorite bits of SAS syntax) is a good next symbol in this sequence. I find those easy to distinguish visually and often end up making a reasonable display for presentation purposes, too. Other times, it is natural to use open circles and filled circles (dots), and open squares and filled squares, to distinguish two factors (circles vs. squares; filled vs. open) in an easy-to-understand way.

Finally, I sometimes use different symbols to represent higher levels of a third variable. There the idea is to increase the amount of ink associated with the symbol as the intensity of the background variable increases. A natural progression is from a hyphen (or vertical bar) to a plus (or X) to the special symbol called a star (but is really an asterisk), and then a hash (pound sign). This is a progression from one line to two to three to four. One can also do something similar with the size of the symbol – say, a star or a dot. If the symbol is open, the larger intensity may not get additional attention.

How much difference does the plotting symbol make? Maybe not a huge amount, but enough to make some difference and spend some time choosing carefully. I recommend you start with "X marks the spot"

BUT THERE'S TOO MUCH DATA TO LOOK AT

So you do a nice scatterplot of your data, and pretty much all you can see is a mess. There's just too much data to look at. What can you do?

First, you can change your plotting symbol to something smaller: if the X is not working for you, you can move to the point and that might do the trick. Second, you might try plotting just a subset of the data. Often the interesting parts of the data are the highest values, or the lowest, or the ones in the middle, or the ones in the upper right, or ... you get the idea. Or maybe you want to look at the deviations from a model. Maybe the data fall very close to a straight line, but it's hard to see the vertical detail on a scale that accommodates the full range of the line. Third, it may be that a transformation of the data will provide a more even spread of the data. The most common form of transformation is the logarithmic transform, to produce a log-linear or log-log graph.

Producing a pretty logarithmic graph is not easy with SAS/Graph because of the difficulty of controlling the scale and tick mark labels. (The Annotate facility provides a mechanism for solving that problem.) But if you're interested primarily for analytical purposes, beauty is less important than getting a good look at your data. You can live with the defaults you get when you specify log scaling to PROC GPLOT, or you can "cheat" and actually transform the data in advance. If you do transform the data, I would recommend using common logs (log base 10) rather than natural logs for most situations. It's easier to remember the order of magnitude with common logs. If you do end up using natural logs, make use of the fact that it's approximately true that "e-cubed is 20". It's actually about 20.086, so that's not far off. This also works to get big numbers

figured out, if you remember your powers of 2. For example, e^{21} is approximately equal to 20^7 , which is 2^7 times 10^7 or about 128×10^7 . The actual value is about 132×10^7 , only off by about 3%.

Another way to look at large amounts of data is to create a contour plot using PROC GCONTOUR. These are not usually very pretty, but with judicious use of shading you can often get a sense of where the heavy parts of the data are. My experience is that these are rarely effective in publications – there's a lot of explaining and the contours often do not reproduce very well – but are good at providing you with insight into data that may be hard to visualize in other ways.

But what if there is **still** too much data to look at? Then look at it in pieces – maybe arbitrary groups of a few percent of the data at a time if need be. It's debatable whether it's most important to look at graphs of the data when you have a little or a lot, but when you have a lot of data it's not feasible to look at tables, so often graphs are the best way to get a handle on the data.

MULTIPLE GROUPS AND MULTIPLE COLORS

Often we want to graph data that represent multiple groups. Maybe we have both males and females, or three different race-ethnicity groups. It's very common to plot all the groups together on the same graph and distinguish them by symbol. That's fine and even good. As someone with diminished color vision (not color-blind, but color-impaired), I tend not to use color to distinguish groups. Or if I do use color, I use colors that are easy to distinguish even for individuals with red-green color-blindness, by far the most common kind. An easy way to check whether the colors will work for the colorblind is to print the graph and then copy it on a black-and-white copier. If it's easy to distinguish the shades of gray that represent the different colors, it will be just fine.

As mentioned in the discussion of symbols, I like to use different symbols to distinguish different groups. If the groups represent a natural ordering, it makes sense for the symbols to reflect that ordering. It's also reasonable to reinforce the differences by having both different symbols and different colors, but you should be leery of trying to use both in the same graph to distinguish two factors: it often doesn't work well. Better, in my experience, is to use filled versus open symbols.

But in addition to plotting all the groups together on the same graph, I strongly encourage you to graph each separately (but on the same scale). It makes the most sense to do this using the symbols that are used when you combine the groups into a single graph, so that you get used to "men are squares, women are circles" or whatever scheme is being used. It's striking how often you will notice that stray from one group buried in with another group once you see the group in isolation.

PLOTTING THE RESULTS OF COMPLICATED MODELS

One of the important uses of statistical graphics to help you understand your data is to plot fitted models. It has become increasingly easy to estimate complicated models, both linear models (such as from REG, GLM, or MIXED) and generalized linear models (such as from GENMOD, LOGISTIC, or GLIMMIX). It can be hard to get a good sense of what the fitted values are from a complicated model, especially one with lots of interactions. It's especially hard to visualize the fitted model with interactions involving more than two factors. But even with models limited to two-factor interactions, with more than a few interactions it's hard to understand them, too.

A couple of years ago, Stefanie Silva presented a paper at WUSS on just this subject, "How to Get a Graph from a Complex Linear Model" (Silva, 2003). She shows step by step how to generate fitted values and plot them to provide a visual display of the fitted model. It's noteworthy how often even an experienced statistician will see something in the plotted model that was unnoticed in the tabular results. Often the graph shows that a different parameterization could lead to a simpler model. For example, it might not be obvious from the parameter estimates that the white males, white females, and nonwhite males are all essentially similar and that only the nonwhite females are markedly different. Or that the fitted lines for four different groups all intersect at approximately the same point far from the origin.

BAR CHARTS FOR UNDERSTANDING

The bar chart seems to be the standard business graphic – when in doubt, put it in a bar chart. At least that seems to be the case for those who don't seem to be in love with pizzas – the dreaded pie charts. (There are uses for pie charts, but not many. Tufte is eloquent on the subject.) And generally bar charts don't provide a lot of insight (although they are often quite effective at communicating information). But there are some times when bar charts are very useful indeed at helping understand data, not just displaying information almost as easily grasped in tabular form.

One simple example of using a bar chart for understanding is to reorder values on the "midpoints" axis (the x-axis if the bars are vertical). For example, if a survey asked questions about how much people agreed or disagreed with a statement, it is natural to present the questions in the order they were in the survey. But a better order might be from the statements with the

greatest agreement on average to those with the greatest disagreement. This is easily accomplished with some Data step manipulation to associate the variable name or label with the mean value from an ordinal scale. (This also requires having the nerve to calculate the mean of an ordinal variable, something that people seem much too reluctant to so.) Then just reading the labels in order provides much of the information in the bar graph and makes it easy to pick out the top and bottom few choices.

CUMULATIVE STACKED BAR CHARTS AND CUMULATIVE GRAPHS

The stacked bar chart is an underutilized business graphic. It does take a little bit of effort to understand the first time, but the effort pays dividends very quickly. When doing graphical displays to try to understand your data better, it can be invaluable. The stacked bar chart is especially useful to display cumulatives that occur in a natural order. For example, suppose you are considering health care utilization in several populations. In each time period, you know how many encounters each patient has had of various types: hospitalizations, emergency room visits, unscheduled office visits, and scheduled office visits. You'd like to compare the groups on each outcome, but it's of special interest to know if the visits are less resource-intensive in some groups – that instead of ER visits they are using unscheduled office visits. One way to approach this would be to create a stacked bar, where the bottom of the bar is the count of hospitalizations, then the next part of the bar is the count of ER visits, then the count of urgent care visits, and so on. Then the height of the part of the bar up to a given type of visit includes the count of all the more resource-intensive visits. For example, the top of the part of the bar that represents the unscheduled office visits shows the number of hospitalizations plus ER visits plus unscheduled office visits.

A similar approach can be used to categorize individuals according to their most intensive health care utilization, so what appears in the graph is not a number of visits (or average number per patient) but the proportion of patients with that as their most resource-intensive visit type during the time period in question.

Another application of the cumulative graph arose when considering insurance claims for a rather vague medical condition ("pelvic pain"). Subsequent analysis revealed later diagnoses that probably were the root cause of the pelvic pain, such as cancer or endometriosis. By placing the later diagnoses into a hierarchy of most to least definitive (and serious), we were able to plot over time the proportion of patients who had ever had a given diagnosis (but never one of the more serious diagnoses). This led to a simple visual of the trends over time in a single graphic. Narrowing bands represented an increase in more serious diagnoses that were greater proportionally than for the less serious diagnoses. Many other features could be picked out of a similar graph of cumulative categorization over time.

CONNECTING THE DOTS A DIFFERENT WAY

Some of the most striking insights I have gotten from statistical graphics have arisen from taking a graph and changing the way the dots are connected or reconceptualizing the axes. The examples I present are idiosyncratic to specific situations I have encountered, but I hope they may inspire you to try to think "outside the box" in your own work.

Both of the examples are related to husband-wife agreement (or disagreement) related to childbearing. In this study, each spouse was asked about their own feelings, but they were also asked about what they thought their spouse would say. For a given question, then, we have answers we refer to as husband actual (HA), wife actual (WA), the husband as perceived by the wife (HP), and the wife as perceived by the husband (WP). The study is longitudinal, so we have these values over a period of up to seven years at intervals of approximately 1 to 1.5 years. There are lots of ways of looking at this data, and they lead to different sorts of insights.

FIGURES AND TABLES

One of the most effective techniques for displaying information that may also be a good way to gain insights is to combine tabular information with a figure or graphical information with a table. Among my colleagues, some like to call these "figures" and some like to call them "tables."

One common application for these combined tables-figures is when there are multiple factors predicting an outcome and you want to express the magnitude of the effect and its statistical significance in tabular form (for detail) but also in graphical form (to provide an overall impression). By presenting the point estimates and confidence bounds in graphical form, a table that would be unlikely to be read turns into a figure that provides insight – and with the associated tabular information handy for reference.

With this example we may have slipped over the boundary from graphics to understand data and into graphics for communicating information. It's always a blurry line, and maybe that's a reminder that a good graphical is a good graphical display.

CONCLUSION

Good statistical graphics are an important means to communicate information, but it's important not to lose sight of the fact they can be very useful in helping you understand your data. I have described some ways to make statistical graphics more useful as analytic tools and given some examples of situations where graphical displays can lead to insights.

1. Plotting symbols: X marks the spot, open versus filled shapes, natural sequences of shapes
2. Lots of data: small plotting symbol, look at part of the data, residuals from a fitted model, log transformations, contour
3. Multiple groups: multiple symbols (not just colors), check colors in black and white, graph groups separately
4. Plotting complicated models: generate fitted values and plot them to show the model
5. Bar charts: reorder the bars to reflect a natural ordering
6. Cumulative bar charts and graphs: display a hierarchy as a stacked bar chart or a graph of "most extreme" outcomes
7. Connect the dots a different way: consider redefining the axes or literally connecting the dots differently
8. Figures and fables: create composite tables/figures to take advantage of the best of each for displaying information

It is my hope that you may be inspired to decide "it's worth the trouble" to do more statistical graphics to help understand your data better. With ODS Graphics, SAS is trying to make it easier for you.

REFERENCES

Friendly, Michael (2000), *Visualizing Categorical Data*, Cary, NC: SAS Institute, Inc.

Silva, Stefanie (2003), "How to Get a Graph from a Complex Linear Model," Proceedings of the Western Users of SAS Software Eleventh Annual Conference.

Tufte, Edward R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press

Tufte, Edward R. (1990), *Envisioning Information*, Cheshire, CT: Graphics Press

Tufte, Edward R. (1983), *Visual Explanations: Images and Quantities, Evidence and Narrative*, Cheshire, CT: Graphics Press

CONTACT INFORMATION

The author welcomes questions and comments. Please direct inquiries to:

David J. Pasta
Vice President, Statistics & Data Operations
Ovation Research Group
120 Howard St., Ste. 600
San Francisco, CA 94105
dpasta@ovation.org
(415) 371-2111

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.