Proceedings of the 2006 Crystal Ball User Conference

HOW TO LIE WITH STATISTICAL GRAPHICS

Andy Sleeper
Successful Statistics LLC
1502 Mathews Street
Fort Collins, CO 80524
Andy@SuccessfulStatistics.com
970-420-0243

ABSTRACT

All statistical graphics are lies, because each graph must filter and modify the data to create a useful presentation. In fact, a completely truthful graph is neither possible or desirable. Good graphs, as opposed to evil graphs, have graphical integrity, when the conclusions reached by a reasonable viewer are consistent with the process that produced the data. This paper examines the impact of choices in filtering, aggregation, and design on graphical integrity. Examples include applications to Crystal Ball® simulations and to current events.

1 TRUTH, LIES, AND STATISTICAL GRAPHICS

Images cause viewers to experience thoughts and feelings which they would not experience without the image. For example, consider Figure 1, which is a reproduction of one of the earliest examples of Impressionist art. A viewer may think about a sunrise, or water, or a sky, or boats, or many other things, but none of these things are real.

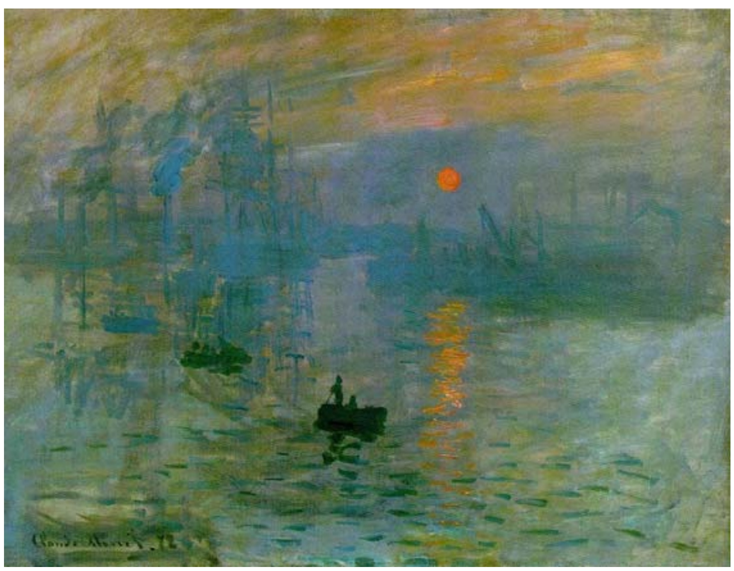

Figure 1: "Impression: Sunrise" by Claude Monet, 1873

The truth about this image is that it is a region of colored dots on this piece of paper or on a computer screen, nothing more. Any other thoughts or feelings evoked by this image are the result of a cunning lie. Every statistical graphic is also a lie. Consider Example 2 below.

In the summer of 2005, my wife was alarmed because of our large water bills. To investigate this phenomenon, I gathered data from our water bills. For each month, I divided the gallons used by the days in the billing period, arriving at an estimate for the average daily water usage. Figure 2 is an area graph of these estimates.

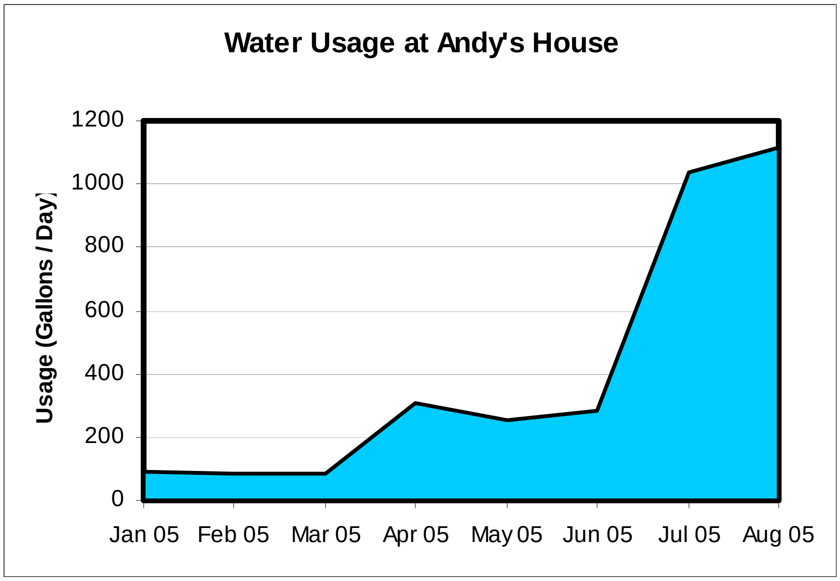

Figure 2: Average Daily Water Usage at My House

Although I have truthfully described the process I followed to make Figure 2, this graph is a lie in many ways:

- The data is filtered to only show usage between January and August, 2005. A graph over multiple years would help to put this data in context. For example, in the summer of 2002, we used over twice as much water per day as we did in the summer of 2005. By only showing 2005, Figure 2 shows this data out of context.
- Although the graph is labeled "Water Usage," the graph does not show water usage. True water usage would be zero most of the time, occasionally spiking to a high value every time we use water for any purpose. Such a graph would be an unreadable mess of vertical lines, which would conceal any monthly trends. Figure 2 shows month-to-month trends well, but hides within-month patterns.
- Billing periods do not have equal numbers of days, but each "month" on the graph receives equal visual weight.

Even though Figure 2 is a lie, it is useful, because it helped my wife and I to understand why our water bills were high, and what we must do to reduce the bills. Therefore, this lie is a good thing.

"Lie" is a strong word, loaded with negative connotations. Statistical graphics are powerful manipulators of perception, and this power may be used for good or for evil. I have used "lie" in this paper to create awareness of the strong emotional power of graphs. Scientists and engineers might deny the emotional content of graphs, but this denial only concedes to the graph creator subconscious control over the viewer's perceptions and conclusions.

2 GRAPHICAL INTEGRITY

The difference between a good graph and an evil graph is integrity. **A graph has integrity if the conclusions drawn by a reasonable viewer are consistent with the true cause and effect relationships in the physical system**. Figure 3 illustrates the relationships in this definition.

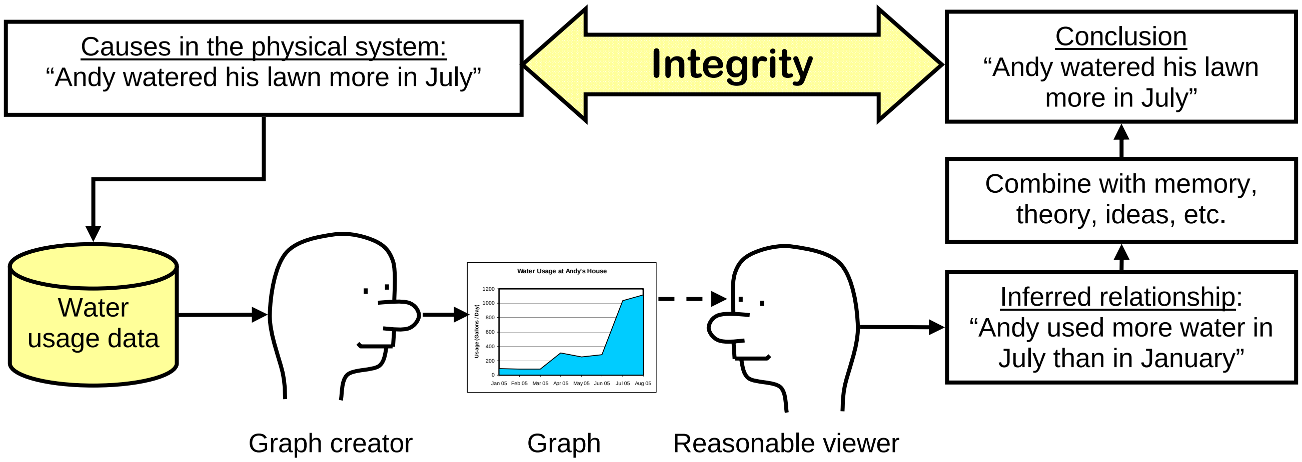

Figure 3: Graphical Integrity

To understand integrity, start with a graph and consider how a viewer perceives it. A viewer first looks at the variables represented by a graph and infers relationships between those variables. Using Figure 2 as an example, there are only two variables: water usage and time. A viewer might infer that I used more water in July than I did in January of 2005.

Data only tells what is, not why or how. In this example, the water data does not explain why I used more water. To arrive at a conclusion about cause and effect, the viewer must combine the data with other information, pulled from memory, theories or ideas. Then, the viewer might conclude that I watered my lawn in the summer months, and that this is the likely cause of my high water usage. Since this conclusion matches the real reason for my high water usage, the graph has integrity.

What if the viewer thinks I reside in Australia, where January is summer and July is winter? Would the viewer's conclusion be different? To reach the correct conclusion about the data, one must assume facts not expressed by the data, including that I live in the Northern Hemisphere, that I have a lawn, and so on.

It may be difficult to determine whether a graph has integrity. Here are some of the difficulties:

- The true causes behind the data may be unknown. This is always the case when doing exploratory data analysis. It is tricky to avoid fooling oneself. The antidote to this problem is to view many different kinds of graphs, looking for consistent and reasonable conclusions.
- The graph creator may have biases and opinions that influence the graph. When the graph creator is biased, the graph may promote a certain conclusion for design reasons, not because of the data itself. Graphs with a biased design are very common in subject areas where there is strong public disagreement.
- Reasonableness of the viewer is subjective. For example, a biased graph creator may think that only viewers that agree with his bias are reasonable, and others are not.

Graphical integrity is a product of many decisions made by the graph creator. These decisions fall into three broad categories: filtering, aggregation, and design. These are the topics of the next three sections.

3 FILTERING

Filtering is the decision to include or exclude portions of the data. Figure 4 shows five views of a time series, each view with successively more filtering. In the first graph, only noise is visible, without patterns. The second graph clearly shows a cyclical pattern. The third graph hints at cycles, but with only this graph, a viewer might not conclude that the process is cyclic. The fourth graph only shows an increasing trend, without cycles, and the fifth graph only shows meaningless short-term noise.

A filter is a window. Images much bigger than the window or much smaller than the window cannot be seen. A particular filter reveals only images that fit neatly in the window.

In almost every case, it is impossible to create a totally unfiltered graph. We cannot possess or process an infinite dataset. With real data, some filtering is unavoidable, but additional filtering is voluntary. When filtering causes patterns or trends in a process to be taken out of context by a reasonable viewer, the filtering is excessive.

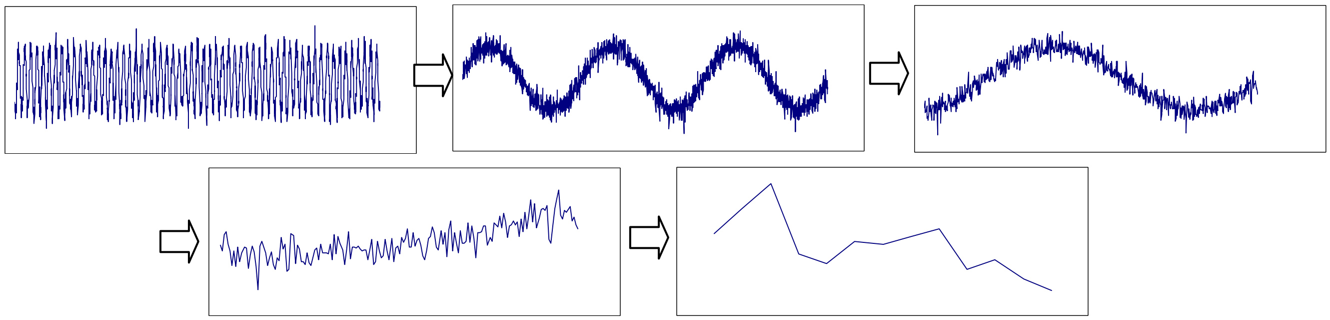

Figure 4 - Examples of Filtering

Crystal Ball automatically filters data when making histograms. Figure 5 is a forecast histogram using automatic scale limits. In this case, only 983 out of 1000 trials appear. There is no way to tell where the other 17 observations lie, and these might be very interesting.

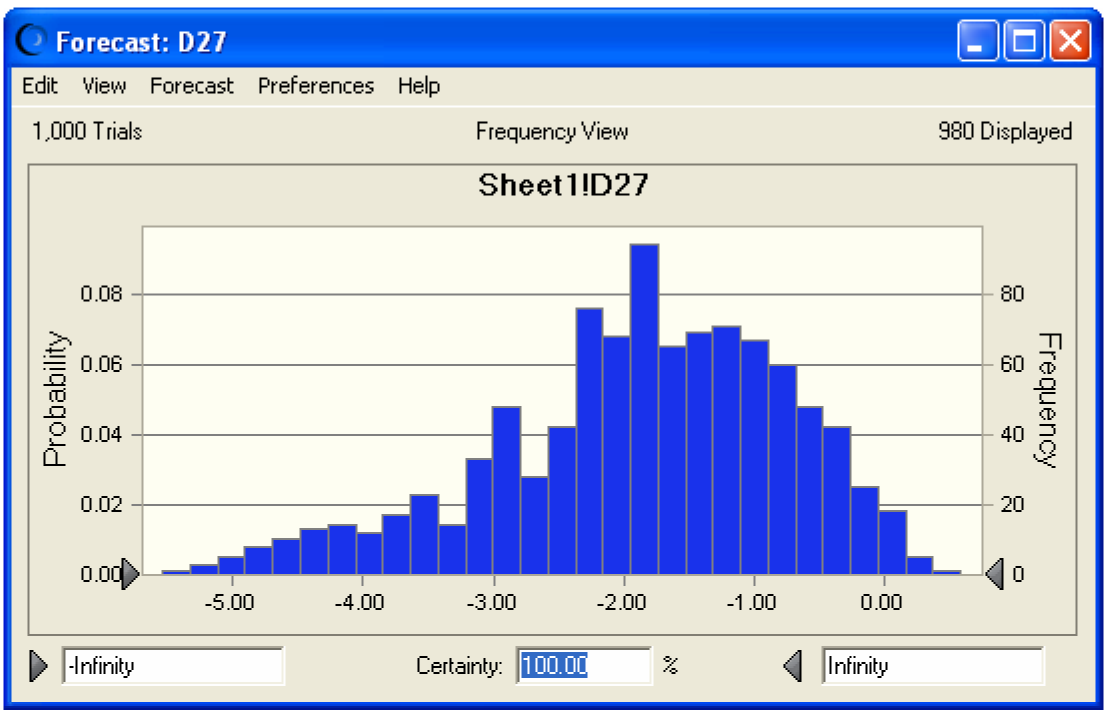

Figure 5: Crystal Ball Forecast Chart with Automatic Scale Limits

To see a histogram including all trials, click Preferences ⇨ Chart… from the Chart menu. Select the Axis tab. Select Scale Type to be Fixed, and leave Min and Max at –Infinity and +Infinity.

In my opinion, filtering ought to be a conscious choice, not a default decision made automatically by software.

4 AGGREGATION

In statistical graphics, aggregation is the process of putting data into groups and plotting groups instead of individual data symbols. Histograms such as Figure 5 require aggregation. Aggregation is an important graphical tool, because it allows the viewer to more easily see patterns between groups. Aggregation also hides patterns within groups.

With too little aggregation, the noise and chaos of individual data tends to swamp out important patterns. With too much aggregation, important patterns may be hidden by being inside groups.

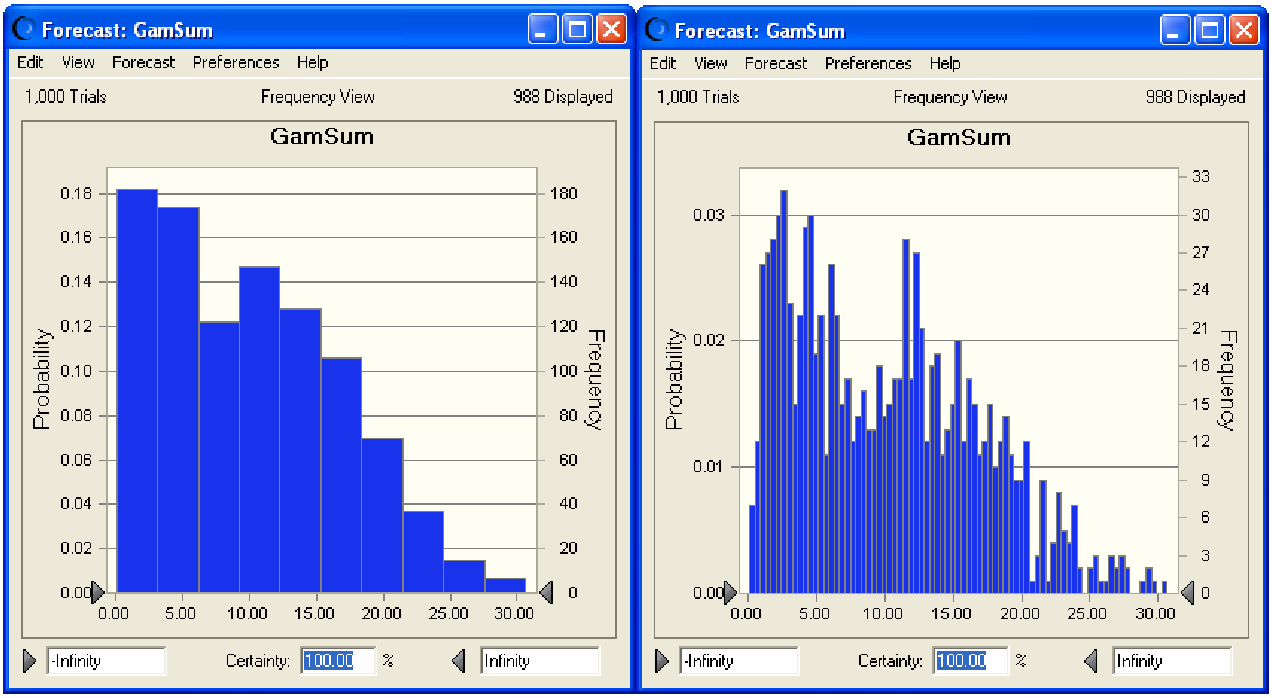

Figure 6: The Effect of Aggregation in Histograms

Figure 6 shows two Crystal Ball histograms of the same set of forecast data. The histogram on the left has a lot of aggregation (few bins), while the one on the right has little aggregation (many bins). Each graph reveals different patterns in the data. In addition to patterns revealed or concealed, consider the possible emotional impact of these graphs on a viewer. The graph on the right is more jarring and unpleasant than the graph on the left because of the mass of vertical lines. Which histogram is best depends on the objectives of the person doing the simulation. To explore Crystal Ball data, try many aggregations by using the ctrl-b hotkey. For presentation purposes, the smoothest graph that reveals the important patterns is best.

5 GRAPHICAL DESIGN

Graph creators have many design choices. Modern software presents a bewildering array of graph types and optional settings. These many decisions can be applied to enhance or obstruct graphical integrity. It is important to remember that just because an option is available does not make it desirable! The purpose of a graph is to fairly convey the story in the data. Every element in the graph should support that objective. Special effects, text boxes, and decorations detract from integrity if they interfere with the accurate perception of the data. Here are a list of general design guidelines for statistical graphics:

- Some graph styles connect the dots with lines, while others use separate symbols. When the story is the trend, connect the dots. When the story is the amounts represented by the symbols, leave them separate.
- Some graph styles shade the area between the data and the baseline while others do not. Shading should be used only if (a) the data has a natural baseline value, and (b) the scale of the graph includes the baseline value. Histograms should be shaded, because the counts of observations in each bin have a baseline value of zero, and the scale starts at zero. When shading is used, the viewer perceives the height of the shading as representing the data, so the height of the shaded visual element must be proportional to the data represented.
- Just say NO to 3-D graphs, like Figure 7. In a 3-D graph, the perceived size of small values is exaggerated, and it is much harder to read the values represented by the symbols. 3-D graphs are examples of gratuitous art, with no redeeming value to the viewer of the graph.
- Pie charts, bubble charts, and other novelty graphs make it harder to perceive the values represented, and should be avoided. Studies have shown that people can perceive the lengths of lines and bars more accurately than pie pieces or cute little people. (see references by Cleveland and Baird)

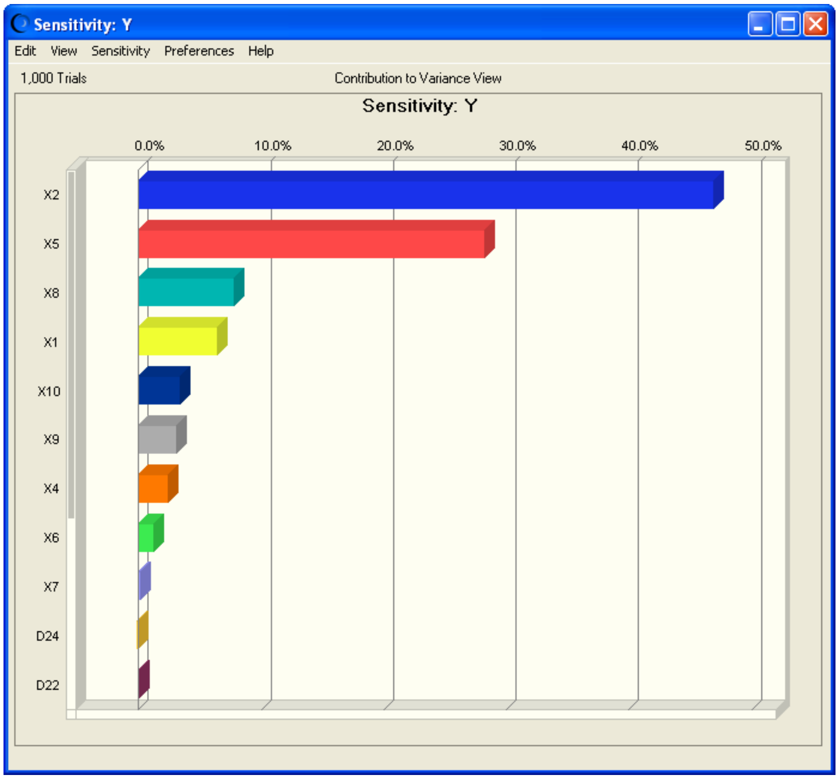

Figure 7: 3-D Graphs Distort Perceived Values

Colors may be the most important choices in the design of a statistical graphic, because of their powerful influence on the viewer's perceptions. One way to classify color is by the Hue – Saturation – Luminance (HSL) system. Each of these three dimensions of color has its own impact on perception:

- Hue defines where a color lies along a spectrum starting with red, proceeding through the colors of the rainbow, and ending up at red again. The "warm" hues (red, orange, yellow) are more associated with strong emotion than the "cool" hues (green, blue). Some specific colors have strong associations, depending on the culture of the viewer. Red is often associated with bad things, like blood, fire, losing money, etc. In a graph, red ought to be reserved for bad things, like boundary lines. When data is represented by red, the viewer's reaction to that data will likely be negative. Crystal Ball forecast graphs use red for data falling outside the certainty grabbers, and this color cannot be changed. Observe how swapping the blue and red in Figure 8 changes the perception of "good" and "bad" data in the graphs.
- Saturation is the vividness of a color. Fully saturated red has no trace of blue or green. As saturation decreases, the color is diluted by other primary colors, until at zero saturation, the color is gray. When a graph represents several data elements represented by different colors, the viewer should see the various colors as different, but equal in importance. Fully saturated colors are the most different from each other, but they also elicit the strongest emotional response. The colors in Figure 7 are nearly saturated. To lessen the potential emotional impact of colors, reduce the saturation somewhat.
- Luminance is the lightness or darkness of a color. Zero luminance is black. Full luminance is white. When luminance is zero or full, different hues and saturations cannot be seen. In a graph, luminance controls the perception of contrast. For example, data should strongly contrast with the background. With a white background, data colors should be dark, with low luminance. With a black background, data colors should be light, with high luminance. To compare data elements as equals, assign them colors with the same luminance.

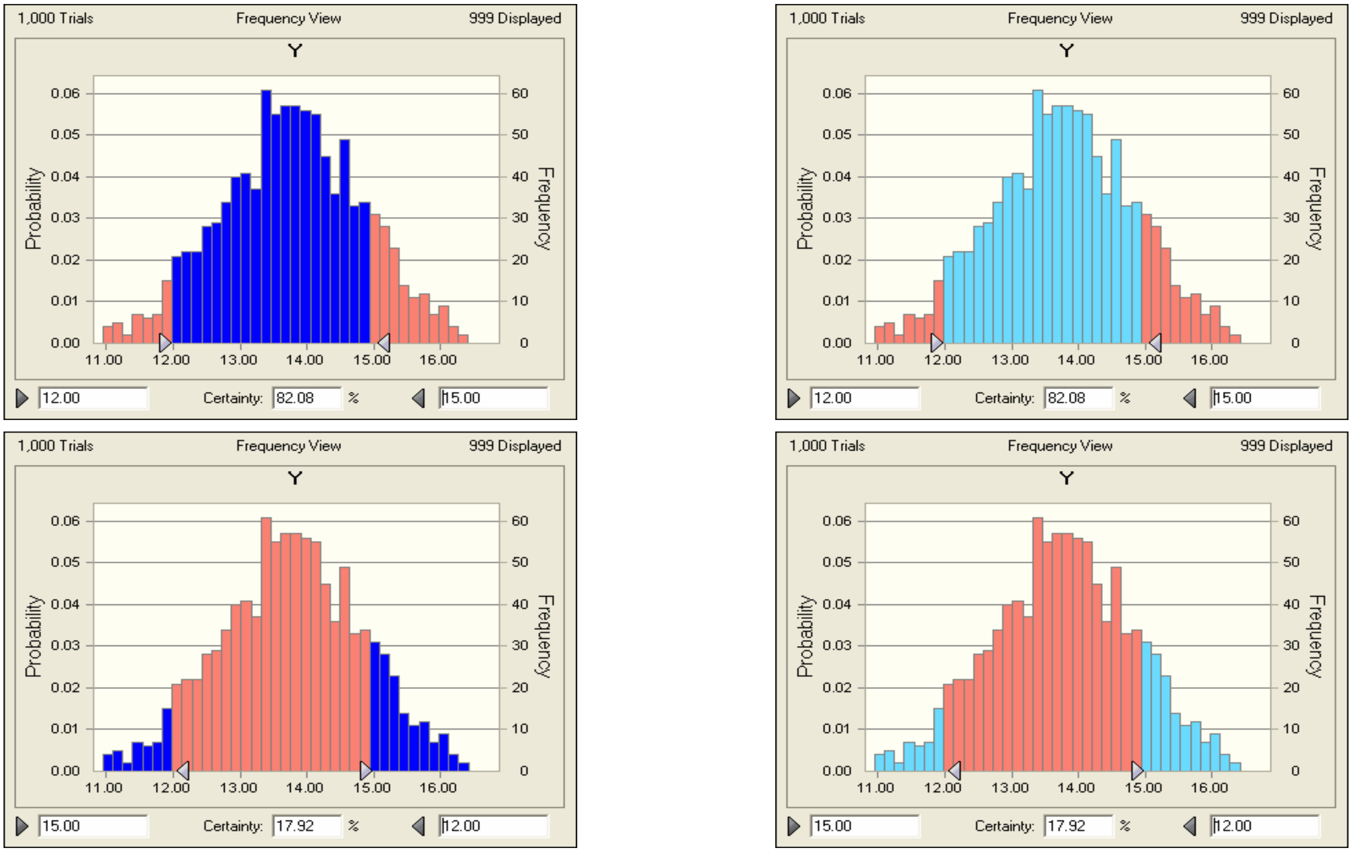

Figure 8: Color Choices Affect Perception

The graphs in the left column of Figure 8 represent default colors in Crystal Ball. The blue and red colors have very different hues and luminance. Because of the difference in luminance, the blue color appears stronger than the weak red. Also, the red is not as negative as it would be with a bright red, as in the second bar of Figure 7. In the right column of Figure 8, the blue and red colors have the same luminance. With the same luminance, this blue and red are more equal in perceived strength or importance.

Equiluminant colors, as in the right column of Figure 8, seem to be fair, but they can create problems for some viewers. Some colorblind people will not see any difference between them. Many people who are not colorblind will find this combination of colors unpleasant or disturbing. The reason for this is physiological. The eye and visual cortex have two sets of neurons. One set, connected to the cones in the retina, is very sensitive to color. The other set, connected to the rods in the retina, perceives motion, but not color. When looking at the right half of Figure 8, the viewer's brain both sees and doesn't see the contrast in colors. This confusion can be unpleasant.

Color is powerful, and there may be no perfect selection of colors for a graph. The best advice for graph creators is to be sensitive to how colors may be perceived. Colors allow complex datasets to be visualized in a single graph. The best colors provide the easiest perception without imposing judgments or biases on the data. The story is the data, not the graph. The story is best told by the data itself, using the graph as its voice.

6 FITTING DISTRIBUTIONS GRAPHICALLY

One task particularly important to users of Crystal Ball software is the fitting of distribution models to samples of observed or simulated data. When fitting distribution models to a dataset, Crystal Ball can provide both statistics and graphs to assess the lack of fit between the data and the distribution models. Figure 9 is an example of a Crystal Ball display with this information. Figure 9 includes two graphs, a histogram and a cumulative frequency plot. Both plots have an overlaid curve representing the best fitting Weibull distribution.

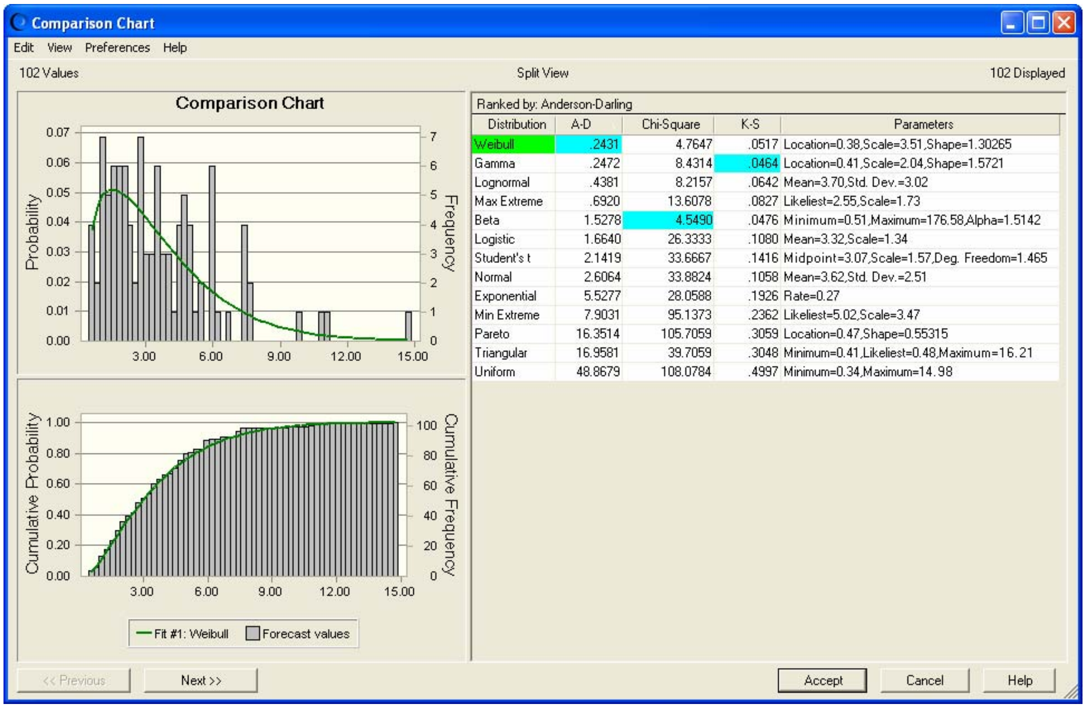

Figure 9: Crystal Ball Comparison Chart for Fitting Distribution Models

A probability plot is a different type of graph designed specifically for fitting distributions to observed data. Figure 10 is a set of probability plots made from the same data used for Figure 9. The four plots in Figure 10 each show how well one family of distribution models fits the data. Crystal Ball software does not produce probability plots, but many statistical packages do this. MINITAB® software was used to create Figure 10.

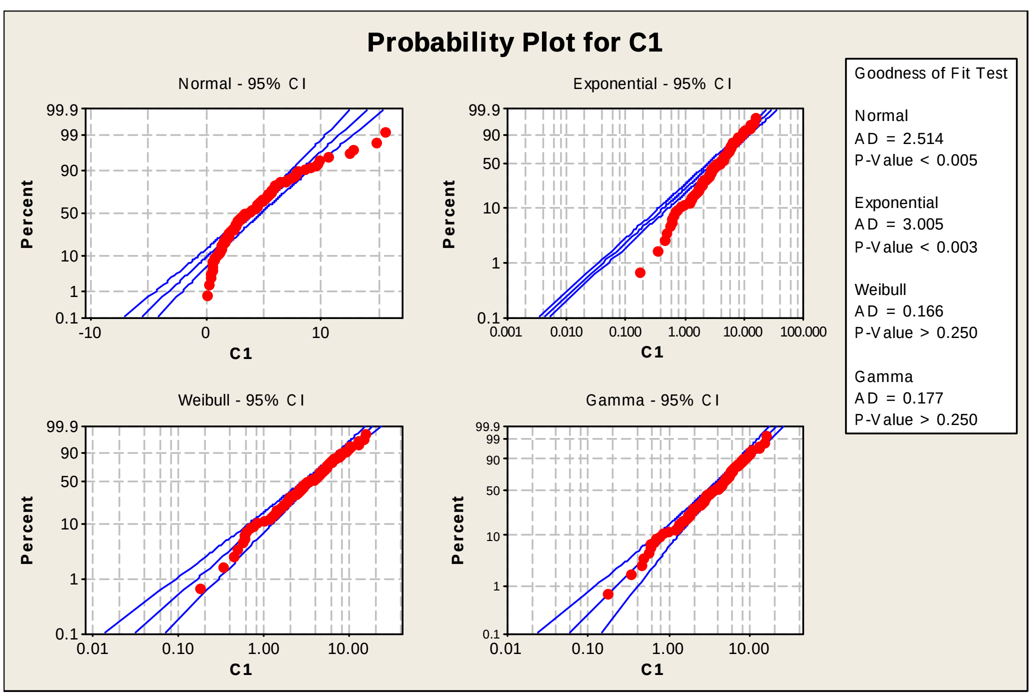

Figure 10: Probability Plots Produced by MINITAB

A probability plot is designed with a special grid, so that if the model fits the data, the dots will follow the diagonal line. Judging from the plots in Figure 10, which family of distributions fits this dataset best?

Probability plots are better than histograms for testing the fit of distribution models for these reasons:

- In a probability plot, one dot represents each observation, so all the observations receive equal visual weight. In a histogram, the bars represent different numbers of observations. Using a histogram to assess whether a distribution model fits gives equal visual weight to each bar, not to each observation. Although this is a matter of opinion, it is more appropriate to give each observation equal weight in the analysis.
- Human viewers are most sensitive to changes in slope for lines at a 45° angle. A probability plot takes advantage of this fact in the design of the plot.

7 GLOBAL WARMING

Many people have strong opinions about global warming, its causes, and its effects. Opinions on all sides of this issue have been shaped by graphs, as presented in hearings, publications, and on the Internet. In fact, my personal opinions about this issue motivated me to include it in this paper, and they shape the words I use to write about it. Even though I am writing this paper about statistical graphics for a prestigious conference, I cannot simply turn off my opinions about issues. Therefore, I challenge the reader to pay attention to the stories told by the data, and separate this from opinions expressed by me. By keeping data and opinion separate, readers can make their own informed, educated decisions about issues.

Figure 11 is a graph generated by the U.S. National Oceanographic and Atmospheric Administration, showing average annual surface temperature anomalies on this planet, from 1880 to 2002. Each anomaly plotted here is simply the difference between the measured average temperature and a baseline temperature, which is approximately 15°C. The graph shows three versions of the data, one for land temperatures, one for ocean temperatures, and one for overall temperatures.

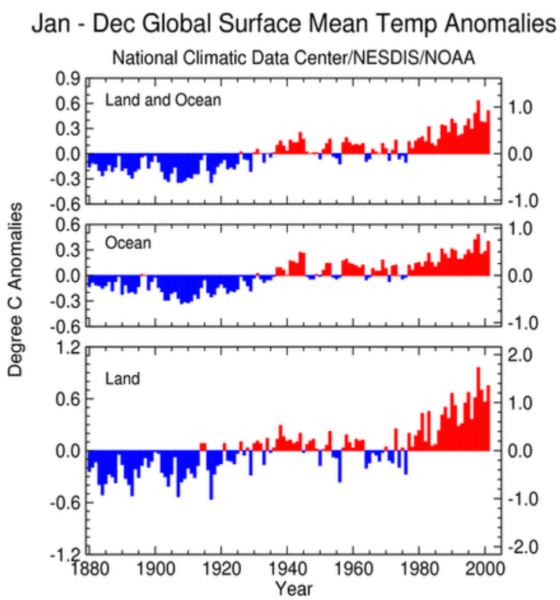

Figure 11 - Average Annual Surface Temperatures, 1880 – 2002

This graph lacks graphical integrity. Here are two of its flaws:

- Temperature data can never have a baseline, unless the baseline is absolute zero. The baseline chosen for this graph is an arbitrary, meaningless number. The shaded bars used to make this graph suggest that a year with a temperature of +1.0 is somehow "twice as warm" as a year with a temperature of +0.5, when the truth is that the +1.0 year is simply warmer by 0.5 degree. This choice of a graph style unfairly amplifies the perceived size of any trend in the mind of the viewer.

- The data changes color from a nice, cool blue to an alarming, scarlet red when it crosses the imaginary baseline. The data represents temperature, and the temperature is increasing. The data is neither good nor bad, but the choice of colors suggests a conclusion that "warm is bad."

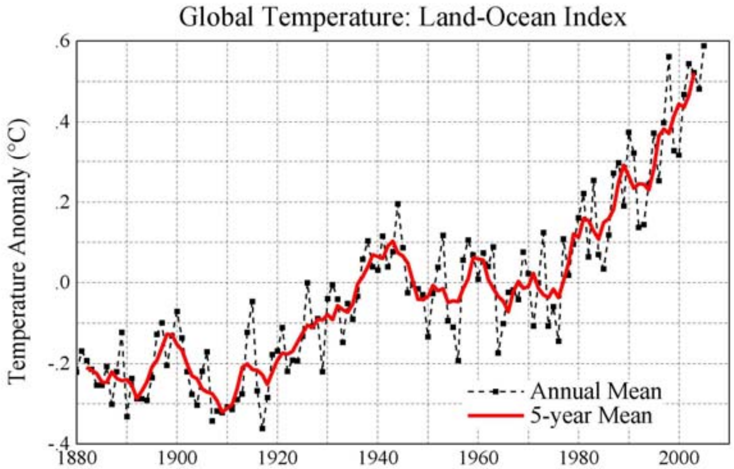

Figure 12: Average Annual Surface Temperatures with 5-Year Moving Average

Figure 12 is a better graph of the same data. In this graph, dots represent the data instead of bars. Without the imaginary baseline, the viewer can fairly assess the trends in the data. The red line represents a 5-year moving average, which is a useful form of aggregation. The moving average hides some of the yearly noise and helps the viewer see the longer term trends. While red is a loaded choice of color for the moving average line, at least the line remains the same color throughout the graph.

The May, 1998 issue of *National Geographic* contains a version of this same graph on page 45. In the *NG* graph, the line changes from yellow to red as temperatures become warmer. Adjacent to the graph is a picture of congested traffic on Los Angeles highways and this caption: "The U.S. is the major contributor of greenhouse gases." Note that the data says that the earth is warming, but it does not say why the earth is warming. The editorial placement of the traffic picture and caption next to the graph clearly expresses the opinion of the *National Geographic* about cause and effect. Cause and effect can never be concluded from observed data alone.

Figures 11 and 12 are both filtered, because the time represented is a very short period of this planet's history. Since reliable temperature measurements are unavailable for earlier times, Figures 11 and 12 show all the available data of this type. However, we have other ways of estimating historical temperatures. By examining core samples of Antarctic ice, scientists have estimated global temperature patterns going back more than 400,000 years (Figure 13).

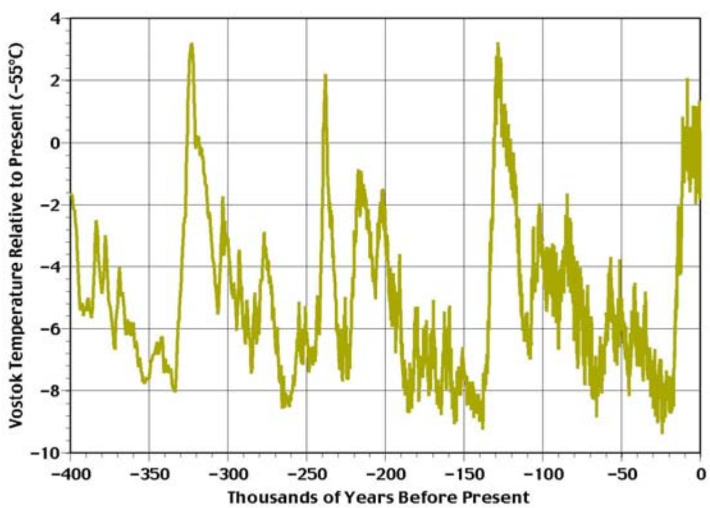

Figure 13: Global Temperature Patterns Estimated from Antarctic Ice

Figure 13 shows the cyclic patterns of ice ages and warmer periods during the planet's history. The last ice age ended approximately 15,000 years ago. Since then, temperatures have bounced within a window of 3°C. This pattern provides a context for evaluating the significance of recent warming. According to Figure 12, the Earth is 0.7°C warmer today than it was in 1880. Using the context of Figure 13, the 0.7° change in recent years is neither surprising nor alarming.

8 GUIDELINES FOR GRAPHICAL INTEGRITY

All graphs are lies. It is necessary to lie to make a useful graph. Good graphs have integrity, when the conclusions reached by a reasonable viewer are consistent with the causal relationships in the process that produced the data. The data itself ought to tell the story, not the design of the graph. These are a few guidelines for producing graphs with integrity:

- When the story is the trend or pattern over time, use graphs with connected symbols, such as Excel line or area graphs.
- When the data has a natural baseline value, and the baseline value can be included in the scale of the plot, use graphs which shade the area between the data and the baseline, like Excel bar, column or area graphs.
- Avoid the use of round or other non-rectangular shapes to represent the data, because it is hard to perceive the data values accurately. This includes pie charts, bubble charts, and many others.
- Avoid gratuitous art, which decorates but obscures clear perception of the data. This means no 3-D graphs!
- Avoid cluttering graphs. Any element in a graph which does not contribute to telling the story should not be there.
- Choose colors with care, considering possible emotional reactions of viewers. Reserve red for bad things.
- Filter the data enough to show important patterns; do not filter so much that the graph takes the data out of context.
- Aggregate the data enough to reveal patterns between groups; do not aggregate so much that important patterns are hidden within groups.
- Change the format or aspect ratio of the plot to bank lines to 45°. At this angle, viewers are the most sensitive to changes in line slope.

REFERENCES

Baird, J. C. and E. Noma (1978) *Fundamentals of Scaling and Psychophysics*, NY: Wiley

Cleveland, W. S. (1994) *The Elements of Graphing Data*, Summit, NJ: Hobart Press

Huff, D. and Geis, I. (1954) *How to Lie With Statistics*, New York, NY: Norton

Sleeper, A. D. (2006) *Design for Six Sigma Statistics: 59 Tools for Diagnosing and Solving Problems in DFSS Initiatives*, New York, NY: McGraw-Hill

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

Global warming graphs from `http://chemistry.beloit.edu/warming`

BIOGRAPHY

Andy Sleeper is President of Successful Statistics LLC in Fort Collins, Colorado. Andy provides customized Six Sigma and Design For Six Sigma (DFSS) training and consulting services to clients around the world. Before working as a Six Sigma consultant, Andy worked for Woodward Governor Company and Hewlett-Packard Company. He has a BS degree in Electrical Engineering (Rice, 1981) and a MS degree in Statistics (Colorado State, 1994). Andy is the author of *Design for Six Sigma Statistics: 59 Tools for Diagnosing and Solving Problems in DFSS Initiatives*, published by McGraw-Hill Professional Books. Contact Andy by e-mail at `Andy@SuccessfulStatistics.com` or by telephone at 970-420-0243. Visit Andy's web sites at `www.SuccessfulStatistics.com` or `www.DFSSstatistics.com`