perceptual edge

Boxes of Insight

Stephen Few
August 2005

This is the second in a series of columns that feature the winners of DM Review's 2005 data visualization competition. This month, I'm focusing on the second scenario of the competition, which presented the following data visualization challenge:

This scenario involves the display of employee salaries per salary grade, with a comparison of male versus female salaries and a comparison of actual salaries to the prescribed salary ranges per salary grade. The purpose is to detect possible inequities between males and females and to determine how closely the prescribed salary ranges are being observed.

Participants were provided with raw data for this scenario, which included the salaries of 100 employees spread across five salary grades. In order to bring the relevant characteristics to light, a solution would have to present male and female salaries per pay grade in a way that clearly supported the following two comparisons:

1. Male versus female salaries
2. Actual versus prescribed salaries

Examining how sets of values are distributed across their ranges can provide important insights. Distributions are ignored far too often by business analysts. In fact, few of the software products that are commonly used to graph quantitative data, including Microsoft Excel, provide the means to effectively display distributions such as the salaries in this scenario. We tend to compare distributions of data by reducing whole sets of values to a single number—an average—and assume this gives us all we need. A simple average, however, is rarely sufficient.

Figure 1 shows the male and female salaries expressed as averages in each of the salary ranges. Note that the averages reveal an apparent inequity between male and female salaries overall and in each of the salary grades, but they don't tell us much about the actual spread or shape of the salary distributions. The two horizontal black lines that intersect each set of bars mark the bottom and top of the prescribed salary ranges for each grade. Based on these graphs, can you tell if any of the salaries fall outside of the prescribed ranges? Not a chance! All you can tell is that the average of each set of salaries falls within the prescribed range, which one would expect.

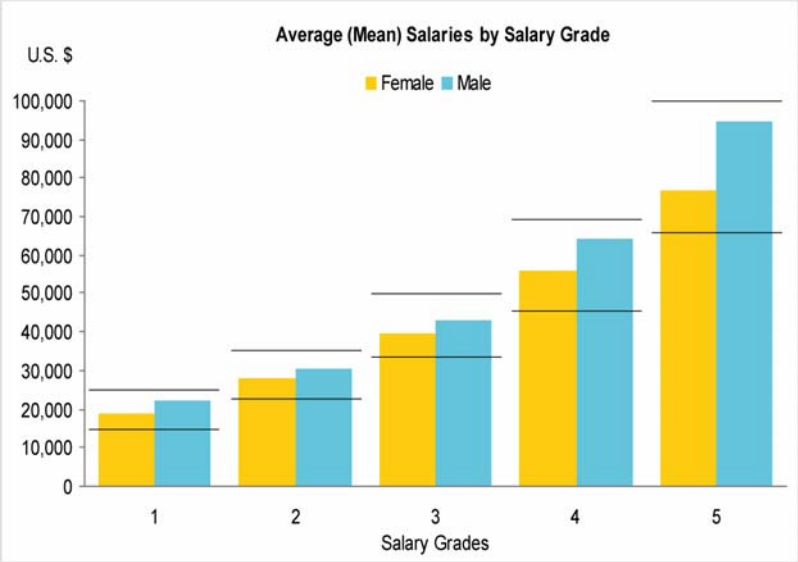

Figure 1: Example Solution

Now let's take a look at the winning solution in Figure 2, which was submitted by Christopher Hanes, an independent consultant. Christopher's solution makes use of a graph called a box plot or a box-and-whisker plot, which was first introduced in the 1970s by John Tukey, the father of exploratory data analysis. Box plots come in several minor variations, but they all work basically the same. The version that Christopher used is one that I like a lot, because it is easy for people to learn to interpret.

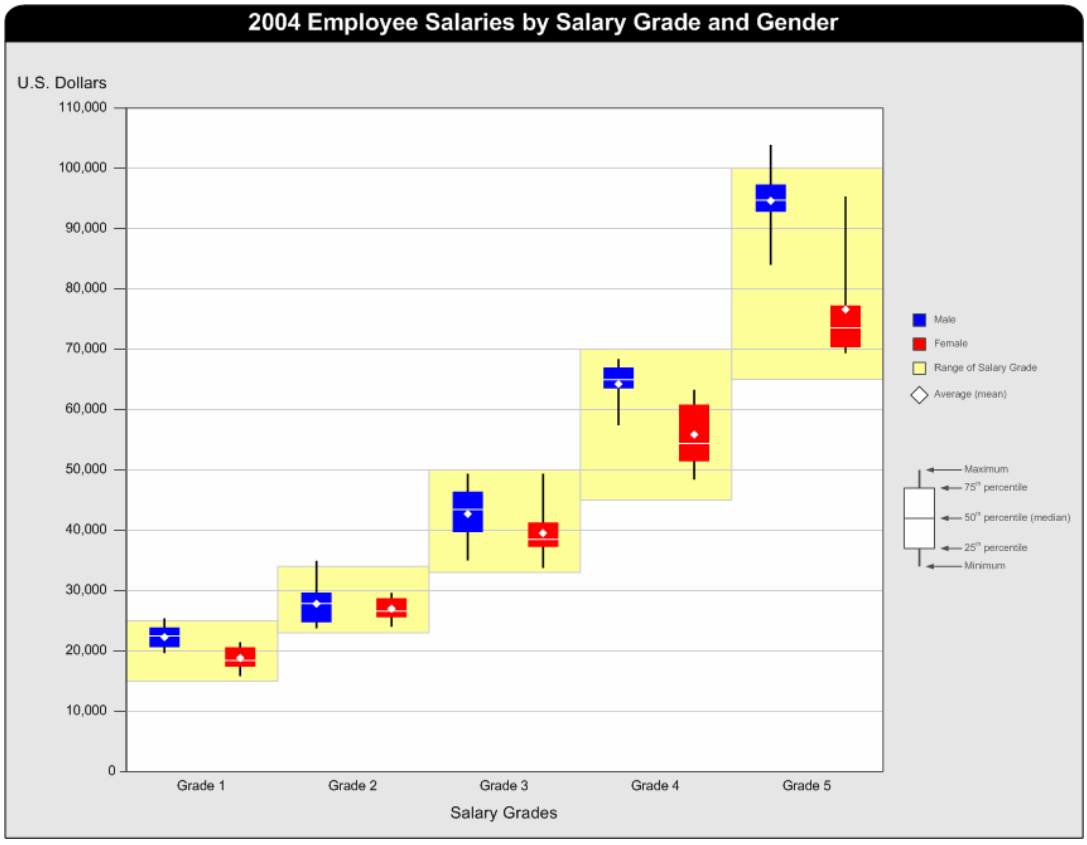

Figure 2: Box Plot Winning Solution by Christopher Hanes

A nice explanation for the box plot symbol is provided to the right of the plot area, which I've enlarged and extended slightly in Figure 3.

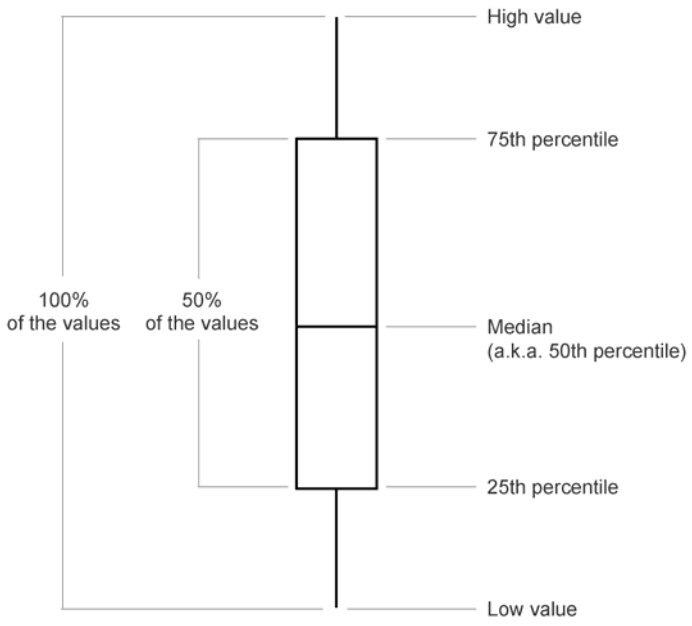

Figure 3: Box Plot Symbol Explanation

If box plots are foreign to you, and perhaps a bit intimidating, I guarantee that it will only take a moment to learn how to make sense of them. Given how much they can tell us about distributions of values, they are quite elegant yet simple in design. Here's a list of the separate facts that this box plot reveals:

- The highest value
- The lowest value
- The range of the values from the highest to the lowest (called the spread)
- The center of the full set of values, which reveals the point above and below which 50% of the values reside (called the median)
- The range of the middle 50% of the values (called the midspread)
- The point above which 25% and below which 75% of the values reside (called the 75th percentile)
- The point above which 75% and below which 25% of the values reside (called the 25th percentile)

Now let's look at a sample distribution displayed in Figure 4 to see what we can learn from it.

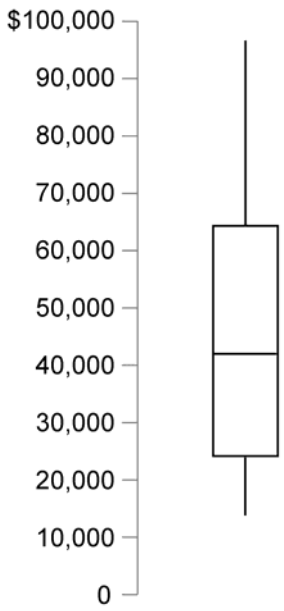

Figure 4: Sample Distribution

Assuming that this represents a distribution of salaries, the first thing this tells us is that the full range of salaries is quite large, extending from approximately $14,000 on the low end to approximately $97,000 on the high end. Secondly, we can see that more people earn salaries toward the lower rather than the higher end of the range. This is revealed by the fact that the median, encoded as the horizontal line in the middle of the rectangle (or box) at approximately $42,000, is closer to the bottom of the range than the top. Half of the employees earn between $25,000 and $65,000, which is definitely skewed toward the lower end of the overall range. The 25% of employees who earn the lowest salaries are grouped closely together across a relatively small $10,000 range of salaries. Notice how spread out the top 25% of employees are. This tells us that as we proceed up the salary scale there appear to be fewer and fewer people within each interval along the scale, such as from over $60,000 to $70,000, from over $70,000 to $80,000, and from above $90,000 to $100,000. In other words, salaries are not evenly spread across the entire range; they are tightly grouped near the lower end and spread more sparsely toward the upper end where the salaries are more extreme compared to the norm. This box plot offers a great deal more insight than a lone average, and even much more than an average complemented by the low and high salaries as well. Not bad for a simple box and three lines.

Given what you now know, imagine that you're the VP of human resources. Look once again at Christopher's solution in Figure 2. See what you can discover about male versus female salaries and how well the prescribed salary ranges for each grade are being observed. Here are the insights that Christopher reported:

- On average, women are paid less than men in all salary grades.
- The disparity in salaries between men and women becomes increasingly greater as one's salary increases.
- There are men in salary grades 2 and 5 who are paid more than their prescribed salary range.
- Salaries vary the most for women in the higher salary grades.

With a little training and practice, you too can learn to coax compelling stories like this from the numbers that measure what's going on at your own place of business. Without proper training, however, you can produce graphs until your fingers are numb and your PC grinds its final bit without ever gaining insight. Developing skill in data visualization is well worth the effort.

(This article was originally published in *DM Review*.)

About the Author

Stephen Few has worked for over 20 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the monthly *Visual Business Intelligence Newsletter*, speaks frequently at conferences, and teaches in the MBA program at the University of California, Berkeley. He is the author of two books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten* and *Information Dashboard Design: The Effective Visual Communication of Data*. You can learn more about Stephen's work and access an entire library of articles at www.perceptualedge.com. Between articles, you can read Stephen's thoughts on the industry in his blog.