

Quantitative vs. Categorical Data: A Difference Worth Knowing

Stephen Few

April 2005

When you create a graph, you step through a series of choices, including which type of graph you should use and several aspects of its appearance. Most people walk through these choices as if they were sleepwalking, with only a vague sense at best of what works, of why one choice is better than another. Without guiding principles rooted in a clear understanding of graph design, choices are arbitrary and the resulting communication fails in a way that can be costly to the business. To communicate effectively using graphs, you must understand the nature of the data, graphing conventions and a bit about visual perception—not only what works and what doesn't, but why.

This month's column focuses on the nature of quantitative information. Graphs display quantitative information: numbers that measure performance, predict the future and identify opportunities. The nature of quantitative information varies in some fundamental ways that tie directly to some of the choices you must make when graphing that information.

Quantitative information consists not only of numbers, but also of data that identifies what the numbers mean. If I walked up to you, looked you in the eye, and said, "The answer is 24,901," you would probably be confused, understandably suspicious that I had a few screws loose. By itself, a number means nothing. However, if I were to tell you that the circumference of the earth at the equator is 24,901 miles, that would mean something. To be complete and meaningful, quantitative information consists of both *quantitative data*—the numbers—and *categorical data*—the labels that tell us what the numbers measure. The graph in Figure 1 highlights this distinction by displaying the categorical data in black and the quantitative data in various other colors. (Note: The gray axes, excluding the red tick marks, are neither quantitative nor categorical data, and in fact are not data at all, but simply visual objects that support the graph by defining the plot area.)

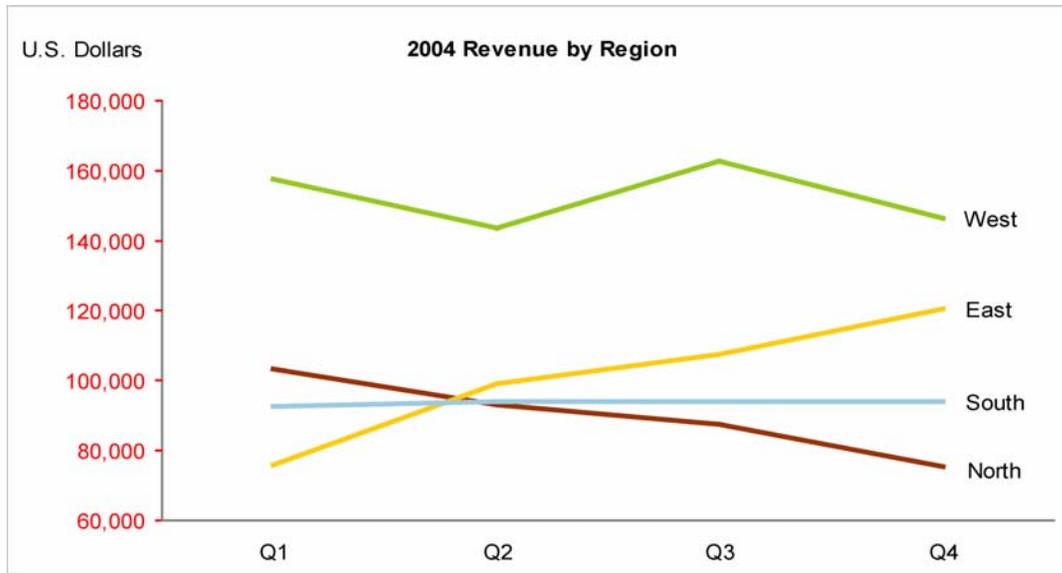


Figure 1: Illustration of the difference between quantitative data (red) and categorical data (black).

The graph in Figure 1 displays two scales: a quantitative scale along the vertical axis and a categorical scale along the horizontal axis. They differ in what they identify: quantitative values on the one hand and categorical items on the other. Most two-dimensional graphs consist of one quantitative scale and one categorical scale, although a familiar exception is the scatterplot, which has quantitative scales along both axes (see Figure 2). In a line graph, the categorical scale always appears on the horizontal axis. In a bar graph, the categorical scale can appear on either axis, with bars running horizontally or vertically. Data points—simple symbols such as dots, squares, triangles and so forth—are rarely used by themselves to encode values other than in scatterplots. Unlike bars and lines, data points can encode a quantitative value simultaneously along two scales.

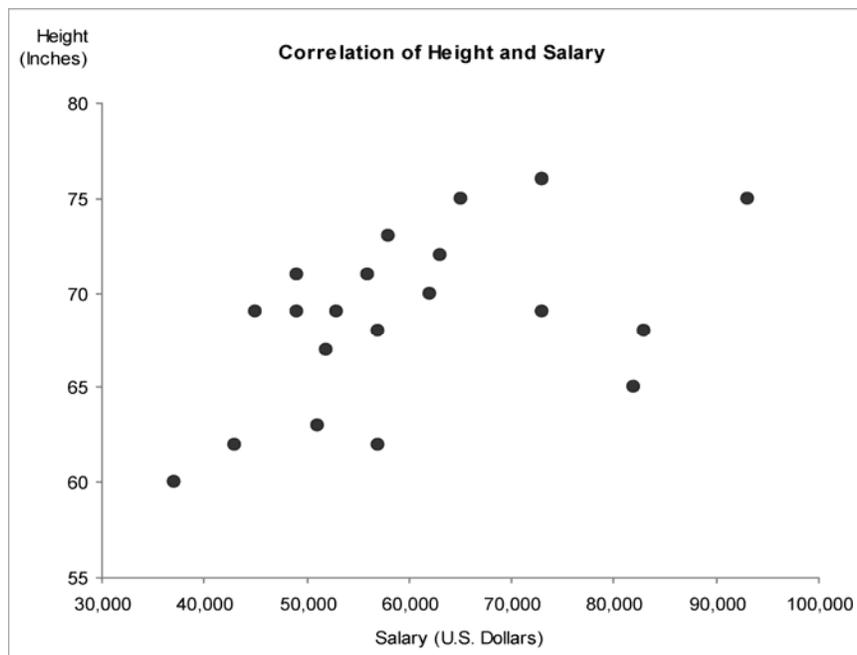


Figure 2: A scatterplot is the only commonly used 2-D graph that lacks a categorical scale along one of its two axes.

Three Types of Categorical Scales

When used in graphs, categorical scales come in three fundamental types: *nominal*, *ordinal* and *interval*.

Nominal scales consist of discrete items that belong to a common category, but really don't relate to one another in any particular way. They differ in name only (that is, nominally). The items in a nominal scale, in and of themselves, have no particular order and don't represent quantitative values. Typical examples include regions (e.g., The Americas, Asia and Europe) and departments (e.g., sales, marketing and finance).

Ordinal scales consist of items that have an intrinsic order, but like a nominal scale, the items in and of themselves do not represent quantitative values. Typical examples involve rankings, such as "A, B and C," "small, medium and large," and "poor, below average, average, above average and excellent."

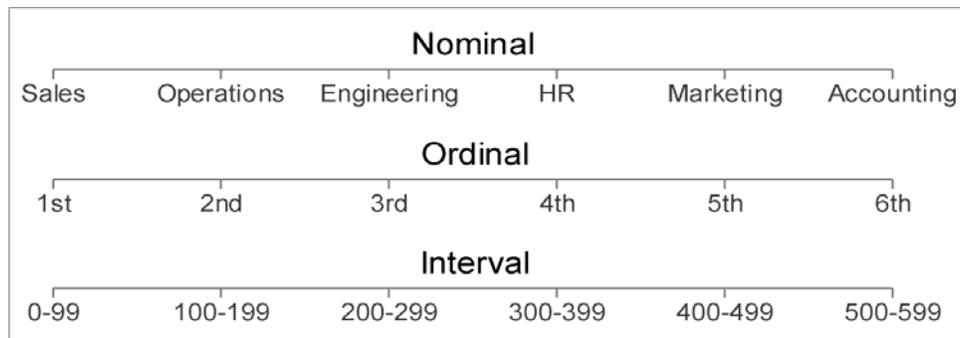


Figure 3: Examples of nominal, ordinal and interval scales.

Interval scales also consist of items that have an intrinsic order; but in this case, they represent quantitative values as well. An interval scale starts out as a quantitative scale that is then converted into a categorical scale by subdividing the range of values into a sequential series of smaller ranges of equal size and by giving each range a label. Consider the quantitative range that appears along the vertical scale in Figure 2. This range, from 55 to 80, could be converted into a categorical scale consisting of the following smaller ranges:

1. > 55 and $\frac{3}{4}$ 60,
2. > 60 and $\frac{3}{4}$ 65,
3. > 65 and $\frac{3}{4}$ 70,
4. > 70 and $\frac{3}{4}$ 75, and
5. > 75 and $\frac{3}{4}$ 80.

Here's a quick (and somewhat sneaky) test to see how well you've grasped these concepts. Can you identify the type of categorical scale that appears in Figure 4?

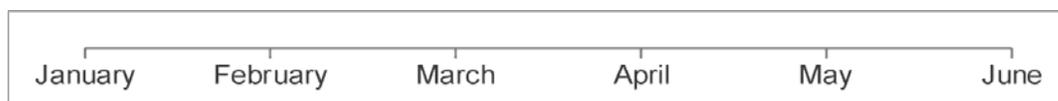


Figure 4: Example of a categorical scale that is commonly used in graphs. Can you determine which kind it is?

Months of the year obviously have an intrinsic order, which leaves the question: "Do the items correspond to quantitative values?" In fact, they do. Units of time such as years, quarters, months, weeks, days and hours are measures of quantity, and the individual items in any given unit of measure (e.g., years) represent equal intervals. Actually, months aren't exactly equal and even years vary in size occasionally due to leap years, but they are close enough in size to constitute an interval scale.

Categorical Scales and Graph Design

The primary graph design principle that applies to this distinction between nominal, ordinal and interval scales involves the use of lines to encode quantitative data. You should only use lines (as in a line graph) to encode data along an interval scale. In nominal and ordinal scales, the individual items are not related closely enough to be linked with lines, so you should use bars instead.

The strength of lines in a graph is their ability to reveal the trend of and patterns in the data. Lines suggest change from one item to the next, but change isn't happening if the items aren't closely related as sequential subdivisions of a continuous range of values. For instance, it is appropriate to use lines to display change from one day to the next or from one price range to the next, but not from one sales region to the next. (See Figure 5.)

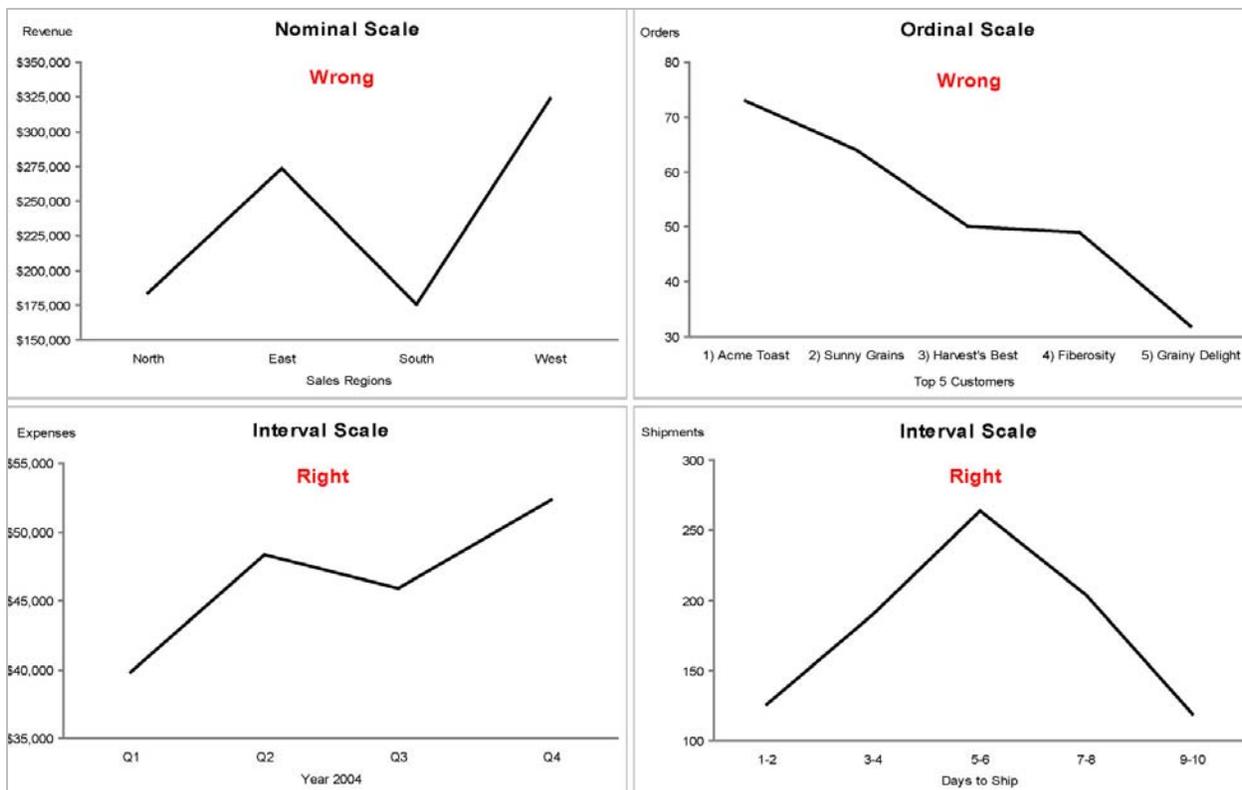


Figure 5: Examples of inappropriate and appropriate uses of lines in a graph.

With interval scales, you are *not* forced in all cases to use lines; you can use bars as well. If you want to emphasize the overall shape of the data or changes from one item to the next, lines work best. If, however, you want to emphasize individual items, such as individual

months, or to support discrete comparisons of multiple values at the same location along the interval scale, such as revenues and expenses for individual months, then bars work best.

These concepts are relevant to many issues and principles of graph design, so keep them handy as you continue to read this column in the future to learn more about data visualization.

(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.