Nuts and Bolts of Data Mining: The Histogram

By Tim Graettinger

Practitioners, business users, developers, and academics love new data mining tools and methods. And yet, successful data mining requires much more than powerful tools. For all the strides that data mining tools have made during my 20-year career, using them well and interpreting their results still requires hard work and serious, critical thought. Remember, "A fool with a tool is still a fool[1]."

That's why I'm writing a series of articles on the nuts and bolts of data mining - starting with this one. In this series, you will learn what it takes to be successful with data mining, what the common pitfalls are, how to avoid or remedy problems, and how to interpret results.

We begin with a real workhorse for data mining and analysis, the histogram. Histograms are bar charts that display the frequency distribution of a numeric quantity, like home value or income. The most famous frequency distribution is the classic bell-shaped curve, also known as the "normal" distribution[2]. Although the bell-shaped histogram is well-known and well understood mathematically, it does not occur that often in actual real-world practice. Among the histograms encountered most frequently in practice are the following: "money", "count", and "outlier". We will look at each one of them in turn.

The Money Histogram

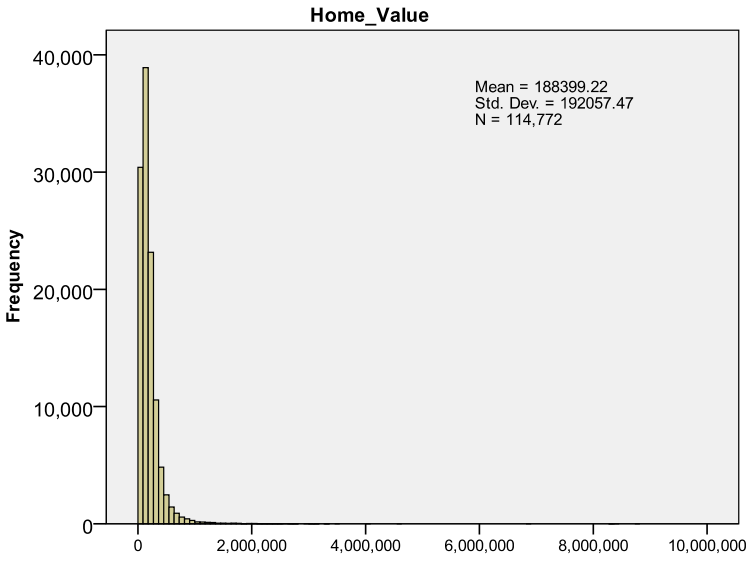

Figure 1 - Histogram of Home Values

"Money" histograms arise in practice when financial data are plotted. The data are usually transaction amounts - home values, salaries, prices paid for products, gift amounts donated to a charity – that are always positive. Figure 1 displays a sample of home values. There is a left-hand "wall" at 0, and the data pushes out to the right to higher and higher positive values. Notice that there are so few of the very high values

that they don't even show up on the chart.

Since this is NOT a bell-shaped curve, there is no "center" of the curve. As a result, for such money data, the standard statistics like the mean (average) and standard deviation are not "mean"-ingful. The median value is somewhat more useful, since it represents the "middle" value.

But neither of these measures, mean or median, tells the whole story of a money data element. In fact, it is the logarithm[3] of money data that is often distributed according to a bell-shaped curve. In Figure 2, I plot the histogram of the base-10 logarithms of the same home values from Figure 1. Using the values in Figure 2 for reference, you can see that the average of the log-value is 5.24. In "regular" (non-log) numbers, this indicates a home value of $172,000 – that is, $10^{5.24}$.

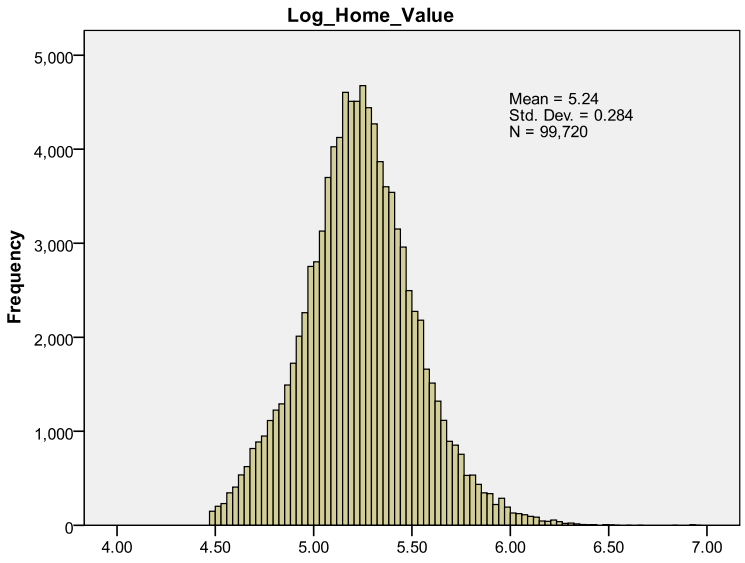

Figure 2 - Histogram of Logarithm Home Values

You can also see from Figure 2 that the standard deviation in log terms is 0.284. And $10^{0.284}$ is 1.92, or almost 2. In plain English, then, one standard deviation above or below the log mean roughly doubles or halves the home value, respectively. For instance, on the logarithm chart, one standard deviation above the mean is a log-value of 5.524 (5.24+0.284), which translates to a home value of $334,000, or about double the average home value of $172,000. One standard deviation below is a log-value of 4.956 (5.24-0.284), which amounts to a home value of $90,000, roughly half of the average value of $172,000.

One more comment: taking full unit steps in the logarithm world, say from 5.0 to 6.0, amounts to taking order-of-magnitude[4] steps in the regular world. Let that sink in for a minute. You don't usually see order-of-magnitude differences in people's heights, or weights, or other physical attributes. But it's not that unusual for one person's money market balance to be 10 or 100 or 1000 times greater than someone else's. This is a key point about financial data – and one you'll want to make often when presenting it. In terms of their financial data, some people can be

more than a little different from the average – they can be a lot different. And they need to be acknowledged and treated as a lot different. By using the logarithm of money data elements, you gain clarity and insight. In my work, I use this technique constantly.

The Count Histogram

A "count" histogram, like the one shown in Figure 3, is also commonly encountered in real-world practice. Here, the horizontal axis consists of integer values (0, 1, 2, 3, …), and the heights are the frequency counts. A good real-world example of a "count" data element is the number of donations a constituent has made to a non-profit organization[5]. In my experience, most constituents have made exactly one donation. The number of constituents who have made 2, 3, 4, or more donations falls off quickly – sometimes even geometrically.

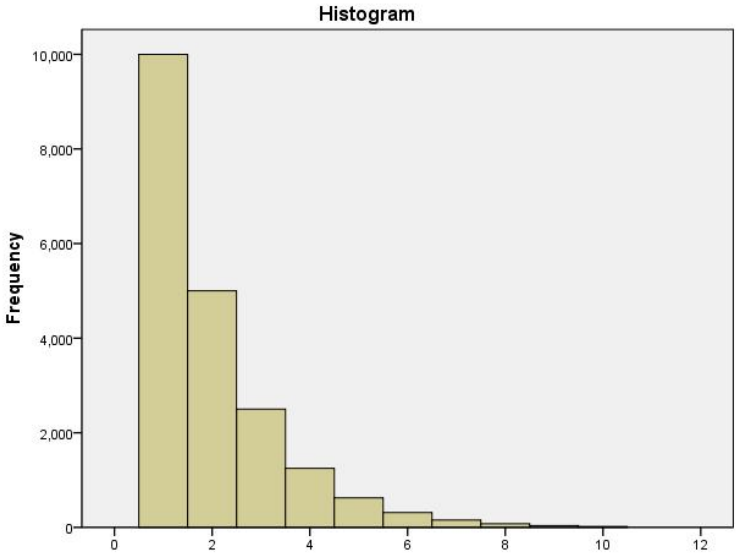

Figure 3 - Histogram of the Number of Donations

To check if the fall-off is geometric, calculate the ratio of the heights from one bar to the next. For instance, suppose that 10,000 constituents made one donation, 5,000 made two donations, 2,500 made three donations, and so on. The ratio of two-time donors to one-time donors is 5,000/10,000 or ½, and the ratio of three-time donors to two-time donors is 2,500/5,000 which is also ½. In this perfectly geometric situation, the number of one-time donors is equal to the sum of all the multiple donors – combined[6]!

This is a key and compelling point for you to make in a presentation. It underscores the importance of converting one-contribution donors into multi-contribution donors. If half of your constituents only make one donation, you are constantly in the mode of acquiring a large number of new donors every year just to sustain the organization. Growing the organization requires even more acquisition. It's an old adage because it's true – the cost and effort to retain a donor (or customer) is much less

than that required to bring a new one in the door. And the count histogram provides empirical support for that argument.

The Outlier Histogram

The "outlier" histogram is the last of the commonly-occurring histograms that you really need to know. Visually, it's incredibly uninteresting – that's right, UN-interesting. It just looks like one big bar – and nothing else – as in Figure 4. But its looks are very deceiving. This can be the signature of real trouble, if unnoticed and left untreated.

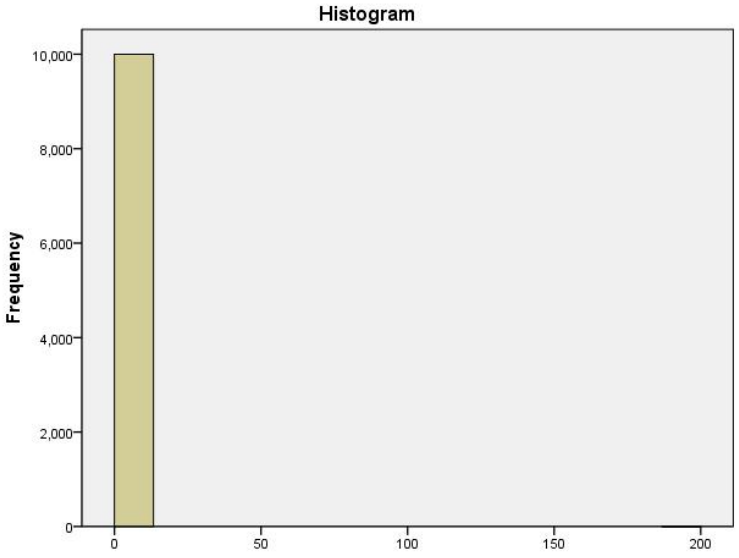

Figure 4 - Histogram of Wealth Deciles with an Outlier

What's really going on here, and why is it so troublesome? Consider again a non-profit organization. Suppose that one data element is the estimated wealth decile for the donors. Decile values should be in the range of 1 to 10, decile 1 being the least wealthy donors and decile 10 being the most wealthy. Suppose that, through a glitch in creating an extract for analysis, one row of data is misaligned. As a result, a value of 199 from another field is placed in the slot normally occupied by the wealth decile.

When this data is plotted as a histogram in Figure 4, it appears as a very tall bar on the left-hand side of the graph. The single, misaligned high value stretches the horizontal axis of the graph far to the right. However, the single high value is insufficient to generate a visible bar on the histogram.

Outliers, or values much different than the norm, are an important subject that deserves much deeper treatment than I can provide in this article. But, you might find a couple of recommendations and comments to be helpful.

First, it is critical to detect and identify outliers. There are reasonable numerical methods[7] that can detect some outliers automatically. They can be very useful if you have many columns of data to analyze. If the number of columns in a data set is relatively small, I still find visual scanning of the histograms to be very quick and very effective. Why is it so critical to detect the unusual, outlying values? Undetected and untreated, they skew most standard statistical analyses; they can lead to erroneous conclusions; and as a result, they can yield poor, costly decisions. In the example, without outlier detection, further analysis would conclude that the wealth decile was unimportant for predicting future giving. However, when the outlier is detected and repaired – resulting in the histogram shown in Figure 5 - more analysis reveals that the wealth decile is actually quite valuable for predicting future donations.

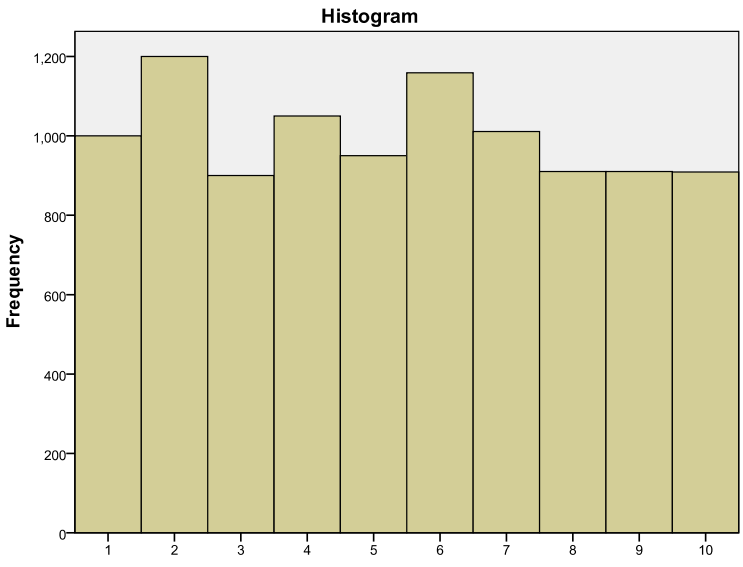

Figure 5 - Wealth Deciles with Outlier Repaired

Second, you need to understand why the outliers are present – and not just delete them out-of-hand. Are they glitches from an extract process, like the example above? Are they simple data-entry errors? Or are they true, actual values that are very unique, rare, and valuable – like a $10 million dollar donation to a charity where the average gift is $20? This last category – the unusual, but true – is frequently quite interesting and often tricky to deal with. But the key lesson is that outliers should not be ignored, and you can easily detect their presence with a histogram.

Wrap-Up

With this article, we begin a new series on the nuts and bolts of data mining. Through the series, you will gain a new-and-improved understanding of the data mining trade: how to use tools and techniques efficiently and effectively. Here we looked at a workhorse in the arsenal – the histogram. It can tell you a lot about where problems lie (like outliers), where opportunities lie (as in retaining donors), and where the important

differences lie (as in home values). The histogram has been around for a long time - with good reason. Use it well.

Tim Graettinger, Ph.D., is the President of Discovery Corps, Inc. (http://www.discoverycorpsinc.com), a Pittsburgh-area company specializing in data mining, visualization, and predictive analytics.

Your comments and questions about this article are welcome. Please contact Tim at (724)-743-3642 or tgraettinger@discoverycorpsinc.com

---

[1] I found this quote attributed to Grady Booch, but I suspect it is dates to the invention of the first tool.
[2] The normal distribution is also sometimes known as the Gaussian distribution, named for Carl Friedrich Gauss.
[3] You can think of a logarithm as an exponent. For instance, $10^2 = 100$, where 10 is called the base, and 2 is the logarithm of 100 using that base.
[4] "Order of magnitude" is another way of saying a factor of ten times.
[5] Or think about this example as the number of purchases a customer has made.
[6] Work it out – it's true!
[7] See "Robust Regression and Outlier Detection" by P. Rousseuw and A. Leroy for a useful introduction.