Living in A Lognormal World

February 3rd, 2010
Nina Zumel
www.win-vector.com

Recently, we had a client come to us with (among other things) the following question: Who is more valuable, Customer Type A, or Customer Type B?

This client already tracked the net profit and loss generated by every customer who used his services, and had begun to analyze his customers by group. He was especially interested in Customer Type A; his gut instinct told him that Type A customers were quite profitable compared to the others (Type B) and he wanted to back up this feeling with numbers.

He found that, on average, Type A customers generate about $92 profit per month, and Type B customers average about $115 per month (The data and figures that we are using in this discussion aren't actual client data, of course, but a notional example). He also found that while Type A customers make up about 4% of the customer base, they generate less than 4% of the net profit per month. So Type A customers actually seem to be less profitable than Type B customers. Apparently, our client was mistaken.

Or was he?

A little more elementary statistics revealed that the median profit generated by Type A customers is $65 — e.g., half the customers from group A generate more than $65 profit per month. The median for Type B customers is about $4.80 — so half the customers from group B generate less than five dollars profit every month. Maybe our client's instincts aren't completely off-base.

Let's look at the distribution of net profit across both customer populations:

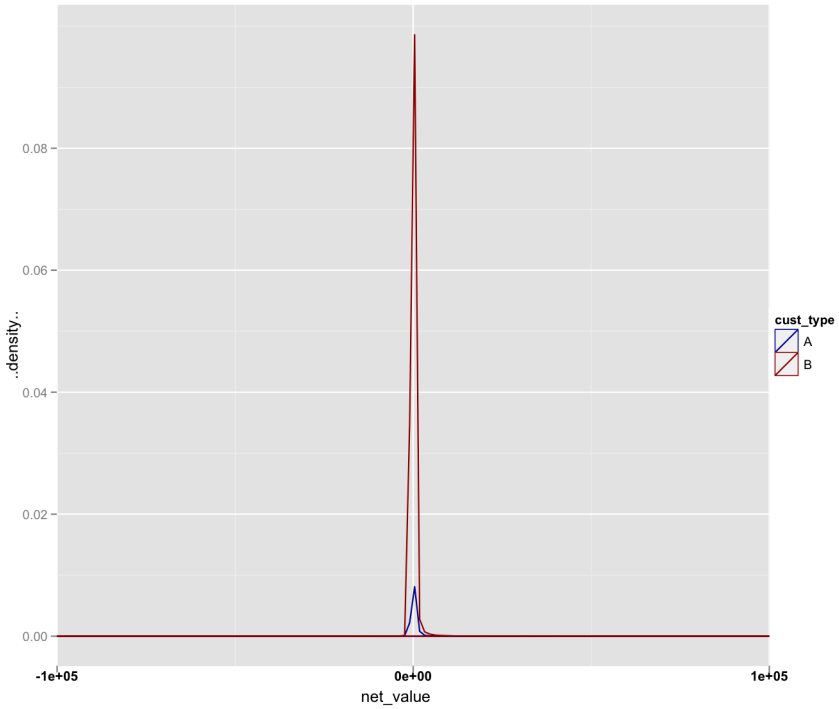

Figure 1: Distribution of net profit for Type A customers (blue) and Type B customers (red). The x-axis gives the net profit or loss, and the y-axis gives the fraction of the population that generates a given net profit.

This pattern is typical among the customers of many businesses. The majority of customers generate relatively moderate profit (or loss); but there is an important minority of large-profit and large-loss customers out on both tails. In this case, the monthly customer value actually ranges from losses in the tens of thousands to profits of several hundred thousands (I clipped the graph, for "clarity").

I hesitate to call these large magnitude customers "outliers" because that term implies anomalous, possibly erroneous, data. In this case, the "outliers" are relatively rare, but important, customers who can potentially make the difference between a company that is in the black or in the red. Still, they are the exception and their behavior doesn't necessarily tell you anything about the behavior of your typical customer. Knowing the mean profitability of a given customer group is important, of course, but the estimate will be dominated by your exceptionally profitable or lossy customers in that group, and as we've seen, that hides information about the majority of your customers.

You might remember from our Good Graphs article[1] that if you have positive skewed data with a wide dynamic range, graphing the data on a log scale helps you see phenomena across the entire range of data that you might miss on the ordinary graph. Unfortunately, we have data here in the positive and negative range. So let's split the customers into three groups: profitable, unprofitable, and break-even. About 5-6% of the customer base is break-even, roughly the same proportion in Groups A and B; we'll ignore them for now, and look at the profitable customers first (over 80% of the customers, in both groups).

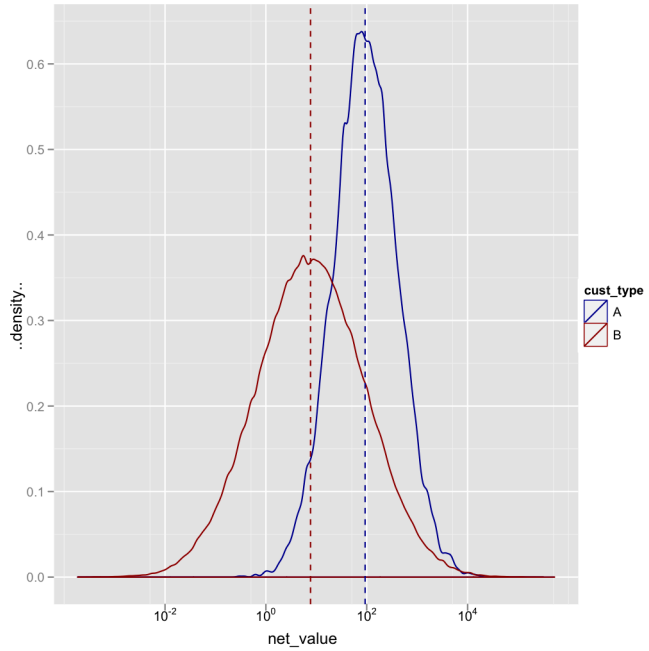

Figure 2: Distribution of profit from profitable Type A customers (blue) and Type B customers (red). The x-axis gives net profit on a log 10 scale, so every labelled tick corresponds to a change by a factor of 100 (eg. 10^0 = $1, 10^2 = $100, and so on). The y-axis represents the fraction of the profitable customer base that generates a given profit.

Now we can clearly see that (among profitable customers) the typical Type A customer is in fact more profitable than the typical Type B customer. The mean profit from profitable Type A customers is about

[1] http://www.win-vector.com/blog/2009/08/good-graphs-graphical-perception-and-data-visualization/

$227, and the median profit is about $93 (shown by the dashed blue line). About 2/3 of the profitable Type A customers generate between $21 and $400 in profit, and over 95% of them generate between $5 and $1721 in profit. We can call that 95% the set of "typical" profitable Type A customers. That's not a standard definition, but it's an intuitive one, and useful for this discussion.

Approximately 2.5% of Type A customers generate profits greater than $1721; let's call them the Type A "best-customers," some of whom generate profits in the tens of thousands. They are responsible for 30% of the profit that comes from profitable Type A customers, and 3% of the profit that comes from all profitable customers (even though they only make up 0.2% of that population).

Profitable Type B customers generate $148 mean profit, and about $7.67 median profit (the red dashed line). A typical profitable Type B customer generates between six cents and $1031 in profit — a lower range than what the typical Type A customer generates, although the very highest-performing Type B customers are competitive with the highest-performing Type A customers (about 130 Type B customers outperform all the Type A customers).

Unfortunately, when Type A customers are unprofitable, they are typically more unprofitable than those of Type B. This is another reason why the mean profit from Type A customers overall was so low. Our client correctly perceived that Type A customers are typically quite profitable, but there is a small population of real clunkers in the group, too.

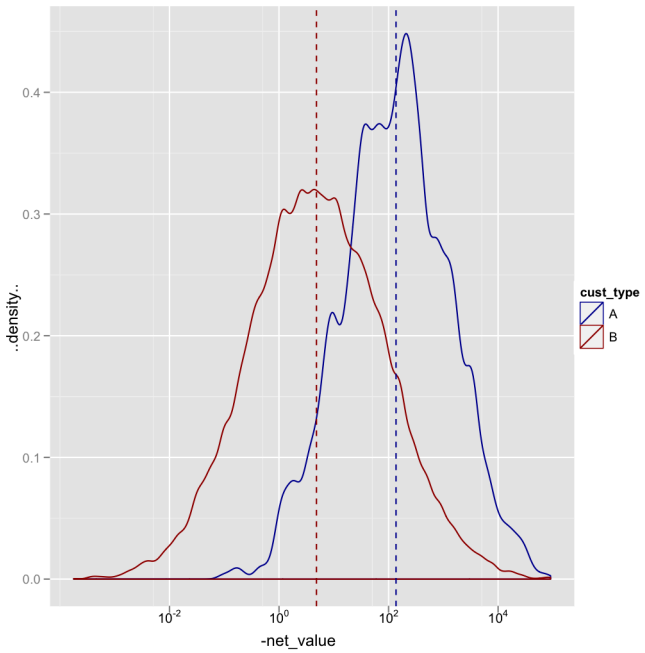

Figure 3: Distribution of loss from unprofitable Type A customers (blue) and Type B customers (red). The x-axis gives loss on a log 10 scale; further to the right on the graph means a larger loss. An unprofitable Type A customer loses a median of $137 a month, and a mean of $1180. Unprofitable Type B customers lose a median of $4.80, and a mean of $210.

We can do a similar analysis for the entire base of profitable customers. We would find that the typical profitable customer generates between six cents and $1200 in profit every month (median $8.65, mean $153), and that the 2.5% of best-customers generate over 60% of the profits.

The Lognormal Distribution

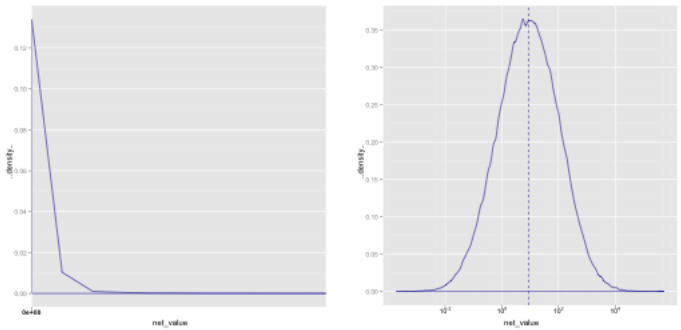

Figure 4: (Left) Distribution of profitable customers (graph clipped at $10,000). The x-axis gives the net profit, and the y-axis gives the fraction of the population that generates a given net profit. (Right) Distribution of profitable customers plotted on a log scale.

The distribution of highly skewed positive data, like the value of profitable customers, incomes, sales, or stock prices, can often be modelled as a lognormal distribution[2]: that is, the log of the data is distributed in a bell-shaped curve centered (in log space) at the median of the data (remember, for a normal curve, the median and the mean are the same). In our case, both the profits (seen above, in Figure 4) and the losses are distributed approximately lognormally. For lognormal populations, the mean is generally much higher than the median, and the bulk of the contribution towards the mean will be made by a small population of highest-valued data points. *If you use the mean as a stand-in for value, you will overstate the value of most of your customers.*

If your customer value data is distributed approximately lognormally, then you can quickly estimate the range of values that 95% of your customers will fall into. About 95% of normally distributed data will fall within plus/minus two standard deviations of the mean, and taking logarithms converts multiplication into addition. So:

If

- *sd* is the standard deviation of the natural log of your customer value data,

[2] http://en.wikipedia.org/wiki/Log-normal_distribution

- *M* is the median profit, and
- *k* = exp(*sd*),

Then

- 95% of customer value is in the range (*M/(k*k)*, *M*k*k*).

The 2.5% of customers who generate more than *M*k*k* profit are your best-customers, who often drive a majority of your profit.

Long Tail Theory

The distribution of customers above sounds a lot like Chris Anderson's Long Tail Theory[3] of consumer goods. Most of the revenue of (for example) a bookseller or a music store comes from a few "hits", or blockbusters, with the rest of the merchant's inventory out along the tail of Figure 5, moving a relatively small volume per title.

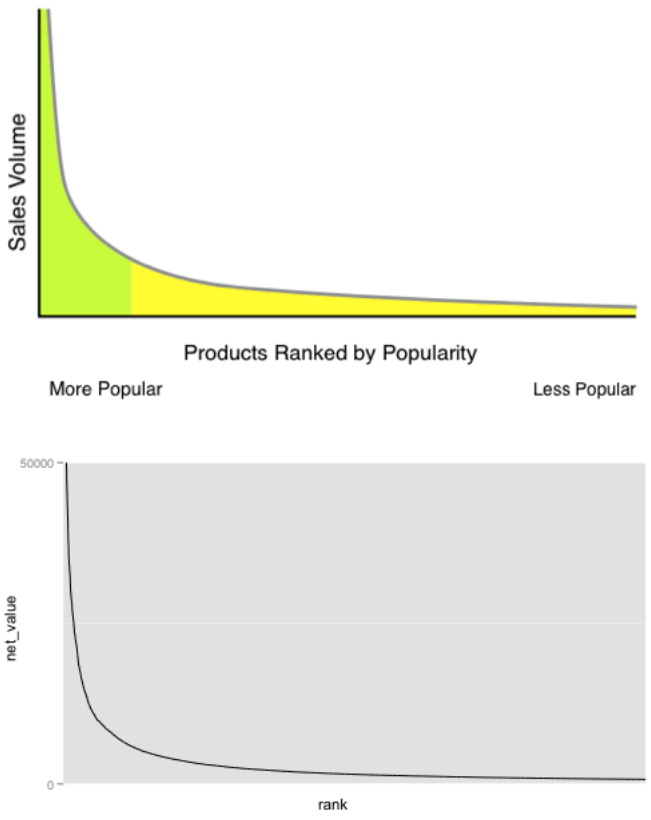

Figure 5: (Top) A notional long tail curve. The y-axis represents sales volume, and the x-axis represents goods ranked from most to least popular. The highest selling goods are to the left. Note that this figure represents the sales curve differently from how the distribution of customer value is represented on the left side of Figure 4.
(Bottom) The customer value data (top 10,000 customers) from Figure 4, plotted in the style above. The y-axis has been limited to $50,000 for clarity.

[3] http://www.wired.com/wired/archive/12.10/tail.html

Anderson generally assumes that sales of such goods are distributed as a power law distribution, rather than a lognormal; the log of power law data isn't distributed symmetrically, but actually has a longer tail to the right. This means that even for the log of the data, the mean is higher than the median. In fact, in some cases, the mean of a power law distribution can be infinite. If sales volume is power law distributed, then top-selling hits are responsible for an even larger percentage of total sales volume than would be the case with a lognormal.

The Pareto Distribution[4], which is one form of a power distribution, has been proposed as an alternative to the lognormal for modelling income distribution and other similar phenomena. Researchers have debated whether lognormal or Pareto is a better model for income distribution since at least the 1950s. Qualitatively, the two distributions have similar behavior. There are certain estimation and forecasting tasks where it does make a difference if your data follows a power law rather than a lognormal, but for the purposes of this discussion, it doesn't really matter. For those who are interested, Michael Mitzenmacher has a fairly approachable discussion[5] about the difference between power laws and lognormal distributions.

Back to Long Tail Theory. Historically, merchants tend to concentrate on high-volume items, due to space limitations and the cost of holding inventory. Overall, however, the sum total of tail-product sales will add up to a respectable volume, especially for web retailers who have unlimited "floor space" — or so the Long Tail theory goes. A retailer must then decide whether to follow the traditional "hits-oriented" strategy, or a more "tail-oriented" strategy that caters to the numerous niche markets.

If we draw an analogy with customer value, then best-customers are "hits." Obviously, our client would like to "fire" his unprofitable customers while retaining his best-performing customers, and even attract more customers like them. But what about his little customers

[4] http://en.wikipedia.org/wiki/Pareto_distribution

[5] http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.im/1089229510

— the 95% of customers in the typical range? If his retention and growth strategy focuses primarily on attracting and retaining big customers, he is following a hits-oriented strategy. If his campaign also includes reaching out to little customers, then he is following something analogous to a tail strategy.

Not all business works like a music or book seller; the appropriate strategy will vary. Still, we can think of a few reasons why keeping little customers happy is a good idea.

For one thing, big customers are not only rare, but they are the ones that your competitors covet the most. Little customers, meanwhile, can still add up to a respectable chunk of change (close to 40% of net profit in our example above). A solid cushion of smaller customers may soften the blow to your profit margin, should a few of your bigger customers defect.

Consider computer sales. Microsoft and Dell serve both the corporate and consumer markets. To judge from their past marketing practices, they consider business customers to be the more valuable segment. But business IT sales have declined in the current moribund economic climate; analysts attribute the growth in computer sales for the last quarter of 2009 primarily to consumer spending[6]. Dell's market growth for that last quarter was much lower than that of HP, Acer, and Apple, which are more consumer-oriented companies. It's also worth noting that Microsoft saw a 14% decline in revenue[7] for the quarter ending September 30, 2009, compared to the year-ago quarter (and their earnings were in large part due to sales of the Xbox, a consumer product), while at the same time, consumer-oriented Apple saw a 24%

[6] http://www.cultofmac.com/apple-saw-24-growth-in-q4-2009-as-computer-market-bounces-back/26184
[7] http://www.neowin.net/news/main/09/10/23/windows-and-xbox-help-microsoft-earnings-beat-predictions

increase in revenue[8] from its year-ago quarter.

Your pool of little customers is also a pool of potential future best-customers. And you can't always guess which ones[9]. So a wise strategy might be to allocate part of your retention and growth campaign to providing loyalty incentives to smaller customers, and educating them about how your higher-end services or products might benefit them. Those little customers who have the means or opportunity to move on to the next level might very well appreciate your efforts, and stay with you, rather than defecting to a competitor.

Optimizing Sales vs. Optimizing Customers

One last thought about retail hits and high-value customers. McPhee's Theory of Exposure, which is cited by Anita Elberse in her Harvard Business Review article "Should You Invest in the Long Tail?"[10], states that the popularity of music, film, TV or books is largely driven by "marginal audience participants" — the casual, or light, consumer. Casual consumers gravitate to already popular products because they have limited exposure to alternatives, and hence limited knowledge of them. Consumers of more obscure products, on the other hand, tend to be heavy (and knowledgable) consumers: voracious readers, dedicated music or film buffs, or enthusiasts of specific genres, like science-fiction or horror.

McPhee's research was done in 1963, using subjects who had a fairly small range of choices, compared to internet scale. Elberse found,

---

[8] http://www.cultofmac.com/apple-saw-24-growth-in-q4-2009-as-computer-market-bounces-back/26184

[9] http://insight.kellogg.northwestern.edu/index.php/Kellogg/article/predicting_customer_lifetimevalue

[10] http://hbr.org/2008/07/should-you-invest-in-the-long-tail/ar/1

however, that the phenomena McPhee described still held for the internet merchants that she studied. She uses this observation (along with McPhee's companion theory of Double Jeopardy[11]) to argue that retailers should not substantially alter their traditional hits-based strategies. There is an alternative interpretation:

If your business follows McPhee's theory, then hit products disproportionately attract low-value (low-volume) customers, and vice-versa.

So an overly hits-oriented strategy will skew you towards a base of low-value customers. Indeed, Seth Godin argues[12] that iTunes and Amazon, who are in a better position to implement a more tail-oriented strategy, are thriving at the expense of physical stores exactly because they have been able to steal the quality (high-volume) customers away.

The moral is that both sales and customer value live in a lognormal world, where blockbuster products are marketed to a large cloud of low revenue customers, and high revenue best-customers are supported by large catalogues of low volume products. Fail to serve one side of this relationship, and you risk losing the other side.

[11] http://en.wikipedia.org/wiki/Double_jeopardy_(marketing)
[12] http://sethgodin.typepad.com/seths_blog/2009/12/its-not-the-rats-you-need-to-worry-about.html