Qualitative Data in Surveys: Lessons from *The Black Swan*

By Rajan Sambandam

White Paper Series

TRC
Consumer Insights.
Market Innovations.

This essay is about qualitative data obtained from quantitative studies and how to analyze them. The thesis here is that the framework used to analyze such data is different from that used for directly obtained qualitative data from methods such as IDIs and focus groups. Understanding the difference between quantitative and qualitative frameworks for data analysis (and in particular, the difference between statistical and managerial outliers) can help in deriving more value when the qualitative data are collected in a regular survey.

But first, let's take a detour through a recently published popular book that provides an analogy that helps in understanding our problem and the potential solution.

A Brief Tour of *The Black Swan*

In the informative (and entertaining) book *The Black Swan*, Nassim Nicholas Taleb argues that real data are either distributed normally (from "mediocristan") or not (from "extremistan"). The former are characterized by data that follow the traditional normal distribution (or bell curve). The majority of the distribution is near the middle surrounding the average and as we venture further out the number of observations becomes increasingly scarce. It is a distribution that defines many phenomena in the natural world. In fact, basic statistics shows that with a reasonable number of observations most distributions start approximating the normal.

Taleb says that the second (extreme) distribution is quite unlike the normal even though it can appear to be similar at first sight. Consider two examples: weight and assets of people. At first they appear to be similar and capable of being represented by normal distributions. But in reality weight is from a normal distribution and assets are from an extreme distribution. When we look at the weights of a few hundred people, most will be near the middle and a few will be outliers. But biology prevents the outliers from straying too far. Removing the heaviest person (a statistical outlier) does not change the mean of the distribution in any meaningful way. In other words, a normal distribution is well suited to represent this data.

Now consider the other variable, assets. It would appear at first to be well represented by a normal distribution. But what if one of the people included is Bill Gates or Warren Buffett? Removing that one outlier can have a significant impact on the mean of the distribution, although from a managerial perspective you don't want to remove that outlier because of its obvious importance. Given that, is the mean a proper way of summarizing the distribution? No, and in fact, in practice we use the median to get around this problem. But that solution will only go so far. Taleb suggests that when people mistake the extreme for normal, they use normal assumptions when inappropriate and therefore fail to consider the likelihood and impact of extreme events. For him, variables like assets should not be modeled as normal.

The extreme distribution is very susceptible to outliers and can be simply explained by the 80/20 rule. For example, 80% of a company's revenue comes from 20% of the customers. Then within the 20%, there is another 80/20 split and so on to the point where one customer could have a huge influence on the entire distribution. In the case of assets of a certain group of people, that person could be Bill Gates. Taleb's point is that current models of finance have mistaken the extreme for the normal distribution and have hence severely underestimated the built-in risk. Scenarios that were seen as several standard deviations away from the mean and hence seen as extremely unlikely ("black swan") have happened and caused havoc in the economy. If it were seen as an extreme distribution (as he did before the current crisis) then predictions of doom would have been loud and clear.

Back to Survey Data

What does all this have to do with quantitative and qualitative (open ended) survey data? Scaled quantitative survey data form distributions that are at least approximately normal. On a scale of 1-10 there are really no outliers in the conventional sense. Scores of 1 (or especially 10) are not particularly remarkable or unpredictable. More importantly, with reasonable sample sizes the mean of the distribution provides a very good approximation of the meaning of the distribution. Generally speaking, a company with a mean satisfaction score of 6 is going to have more dissatisfied customers than one with a mean satisfaction score of 8. Not considering the extremes of the distribution is generally not a problem as single observations there are unlikely to change anything. That is, statistical outliers (those that are not of use to a manager) can be ignored without consequence.

The story, however, gets more interesting with responses to open ended questions. No scales are being used here to artificially create end points. People are free to say what they want both in terms of quantity and quality. But can their responses really be described as normal even if a large number were used? What would such an assumption imply? That with proper coding of responses the

"mean" of the distribution adequately represents its essence? What do we miss when we do that? Well, if the distribution is at least somewhat extreme then we could be missing the outlier that can have a huge impact.

Let's use an example to think through this. Say that 500 open ended responses were collected in response to a question about possible ways of improving the product. As it happens often in practice, there will be something like the 80/20 rule in that fewer people will have a lot to say and many will have little to say. The common way of summarizing this information (since one doesn't want to present the 500 responses verbatim) is to code them based on similarity of responses. This would lead to categories of answers revolving around concepts such as price, quality, service etc. In turn, this will be presented as 30% feel that price could be improved, 25% feel that service is a problem and so on. Using the black swan analogy, what has happened here is that an open-ended possibly "extreme" distribution has been force fitted into the "normal" paradigm. What are the consequences?

Any responses that are unique and potentially very valuable will now either be subsumed into a larger dull category ("need better service") or worse yet, sent to the "Other" category never to be heard from again. So the one person out of 500 who provided a brilliant suggestion that could potentially be a blockbuster (or more modestly a simple cost saver) for the company may never be heard from. It happens because the qualitative data with inherently different properties were treated like quantitative data that can be described by a few summary measures. In other words, managerial outliers (those that are of great use to a manager) when ignored can be a significant loss to the company.

So how does one overcome this problem? Reading a sample of the responses will not solve the problem. Even text analysis software may not provide a proper solution as they tend to be summary focused. The most straightforward answer is for a person knowledgeable about the research problem (and perhaps with an understanding of the quantitative results) to read every open-ended response. On the one hand this is more daunting than it sounds because not just reading but contemplation is involved in identifying the interesting outlier. On the other hand it is less daunting than it sounds because one could quite quickly rip through the majority of responses that are not substantive. Heuristics will need to be developed that allow the practical researcher to strike the optimal trade-off between speed and care.

In Conclusion

More generally, what we see here is the attention paid to the kind of data being collected and what it says about insight that can be gathered. Scaled data are essentially pre-defined. They serve a very specific purpose: quantifying the proportion of people who agree or disagree with something. They will not directly provide any information outside of that. Hence the opportunity to discover something new is almost never there. The open-ended data on the other hand are free to roam where they want. They always have the potential (repeat, potential) to provide something new. So the way to analyze them should be different in order to allow outliers to properly reveal themselves. Outliers can be in two categories: statistical and managerial. Quantitative data have statistical outliers and usually they are not of much interest. Qualitative data can have managerial outliers which may be all important.
Of course, every study is not going to produce that one comment that can change a company's fortune. To help increase the chance of getting such feedback Smart Incentives could be useful (see References). But there are benefits even when brilliant outliers aren't to be found. Treating qualitative data in their own right rather than by rote analysis, allows for better and proper integration with quantitative data and therefore superior insights overall.

References

Taleb, Nassim Nicholas, *The Black Swan: The Impact of the Highly Improbable*, Random House, 2007

Sambandam, Rajan (2005), "You May Get More Than You Pay For", *Quirks Marketing Research Review*, Vol 19, page