Deceptive Data and Statistical Skullduggery

by Gordon H. Bell
published on AudienceDevelopment.com
March 10 and 17, 2008

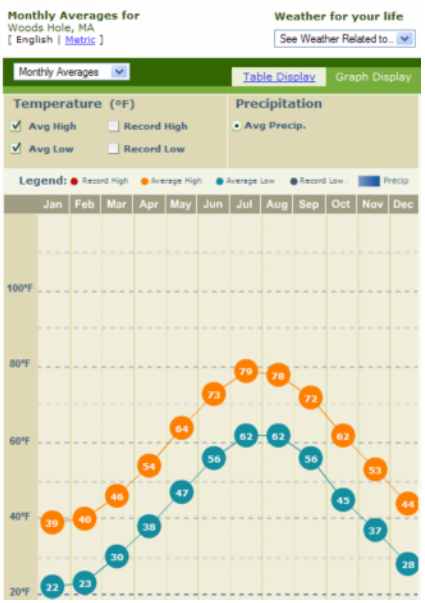
People have an innate trust of numbers. But like the recent results of political polls, market numbers must be viewed with healthy skepticism. On an industry survey, I may claim a 12.40% response rate, but if that is based on 1,000 of my best prospects, taken from the last shipping day before Christmas, or the gross response from a soft offer with a great premium, then that very precise number probably means little.

Specific, black-and-white numbers mask the colorful variation found in the natural world. Just as every fingerprint and snowflake is different, each marketing campaign produces different results. Market competition, customer behavior, and a host of uncontrollable factors work together in complex ways with immense uncertainty. Real-world data are seldom as accurate as they seem.

You should pay particular attention to four potential sources of error: ignoring variation in your data, using insufficient sample size, confusing correlation with causation, and assuming outcomes which have never been tested.

We're all doing fine… on average

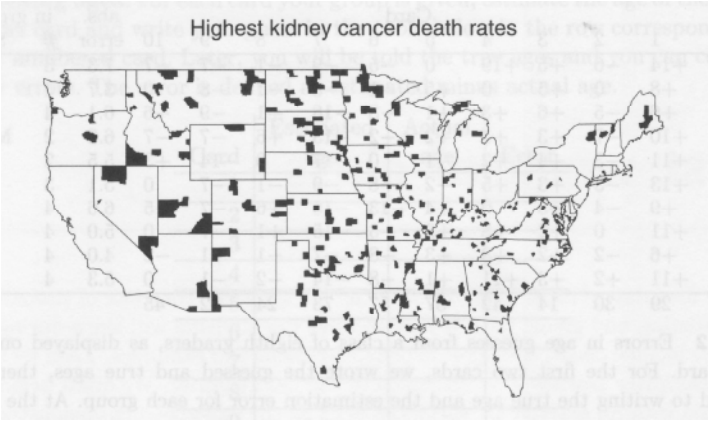
Let me start with a personal story. As we were planning a trip to Cape Cod last July, my family questioned my choice of a rental house with no air conditioning. I showed them the temperature chart from weather.com, below left, to alleviate their concerns. Even at the height of July, the temperature averages between a high of 79 and a low of 62. But after commenting that our family was far from average, my wife asked how high the temperature *could be*, so I showed her the chart at right. Adding the record highs and lows, you get a better sense of the variation around these averages. The second chart shows that the July temperature may very well reach the 90s (and in fact, it was a bit too hot to be without AC).

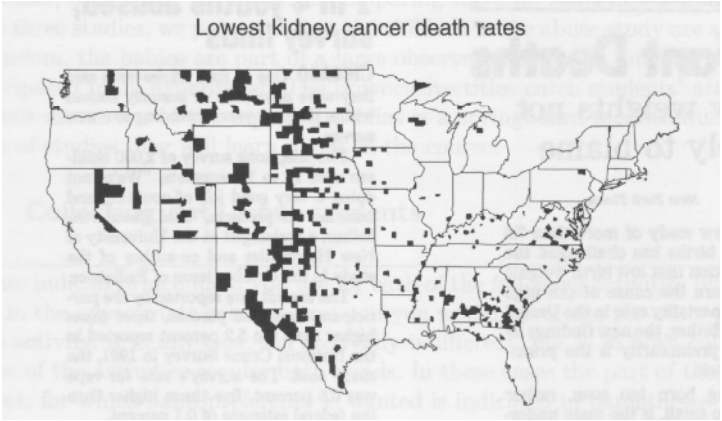


Averages show you where your data are centered, but overlook individual outcomes. About half of your marketing campaigns will perform above and half below average, so you are bound to be either pleasantly surprised or miserably dismayed with the result of your next campaign. Do not ignore variation, but incorporate it into your planning. Know your uncertainty.

You want more drama? Use less data!

The charts below highlight counties across the U.S. with the highest death rates due to kidney cancer (below left) and those counties with the lowest death rates (below right). What can you learn from these charts?





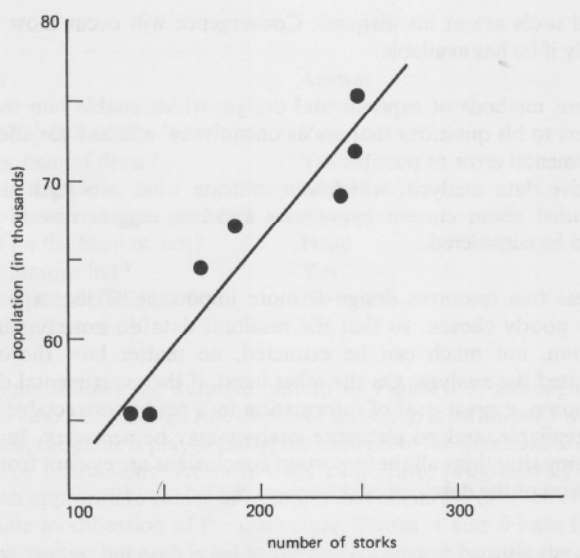
With the highest and lowest death rates concentrated in the mid-West and West, you may be safer living in California or the Northeast. But wait, how can certain counties in Montana and South Dakota be some of the best and worst places to live? It comes down to sample size. In a small county, no one may die from kidney cancer over a few years, leading to a 0% death rate. Alternatively, if one person dies in a small county, the death rate is surprisingly high. In larger counties with larger populations, these extreme values tend to average out. Relating this to marketing: if you send out 2 e-mails and one person subscribes, then you achieve a whopping 50% conversion rate! But send out 1 million e-mails and that rate may go down a bit.

Simply Deceptive Data

Simple data can be surprisingly deceptive. Lurking behind the mask of objectivity may be a great deal of natural variation, along with additional error from placing too much importance on too few numbers. If you understand the variation in your market data and gather enough information, you can unmask the uncertainty and find truth in the power of numbers.

Proof that storks carry babies

The old story that storks carry babies is, in fact, supported by objective data. The chart, at right, shows the population of the town of Oldenburg, Germany versus the number of storks sighted each year from 1930-1936. This relationship clearly shows that more babies are delivered (the population increases) when more storks are around to delivery them. This is real, accurate, indisputable data.



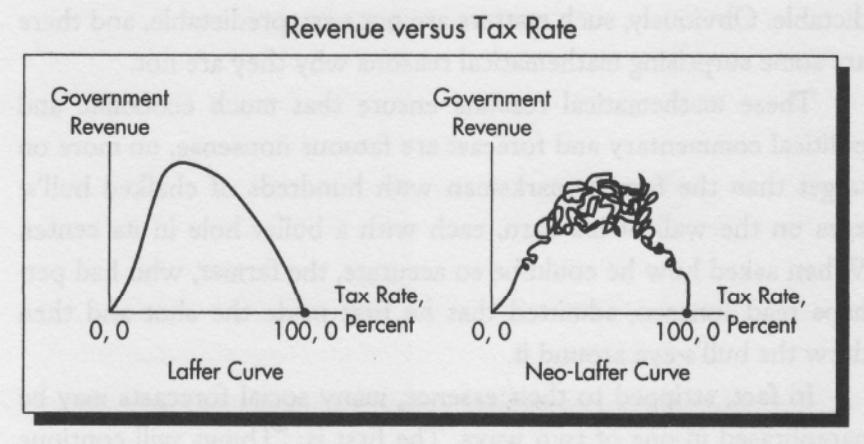
The problem is that correlation is not the same as causation. The town of Oldenburg may be an attractive home for both people and birds, but importing storks may not be the most effective path to urban renewal. Using even the most advanced statistical analyses, a correlation between variables does not prove they are related. Only a well-designed test can prove a cause-and-effect relationship.

Assuming the unknown

Testing is a combination of science and art—specific data translated into general knowledge. All you can prove is exactly what you test: at this point in time, the selected creative, price and offer mailed to a particular group of prospects produced the given response. The big challenge (and opportunity) is interpreting what the data mean as you rollout results in an ever-changing marketplace. For example, if you test a $10 and $20 subscription price, you might see a 5% and 3% net response rate for each. Does this mean that a $15 price would produce a 4% response, or a $30 price a 1% response? You may be able to make an educated guess between test levels, but extrapolating beyond test levels can be risky.

Assuming a simple relationship can, at times, lead to the wrong conclusion. An interesting example comes from the debate between tax rate and government revenue. The economist Arthur Laffer created the "Laffer Curve" starting with the simple truth that a 0% tax rate leads to no tax revenue and a 100% tax rate gives people no incentive work. He then assumes that increasing a low tax rate will lead to more revenue, but at some point people have a disincentive to work. On this downward portion of the curve, lowering the tax rate will increase revenue by stimulating the economy.

However, with so many variables affecting the economy, no one really knows what happens between a 0% and 100% tax rate, or where we are on this Laffer Curve (shown at left in the picture on the next page). Writing for the *Scientific American*, Martin Gardner argued for a more realistic "Neo-Laffer Curve" (the right side) showing the true uncertainty between the extremes. Charts can look so clean and neat, but you should know which data are real and which are assumed.



Finding clear answers

Data is power. And that power can be used to clarify or conceal the underlying reality of the marketplace and the world. You should understand the variation behind those precise market numbers. Gather enough information (sufficient sample size) to rise above the cloud of uncertainty. Make the transition from "I think" to "I know" by testing important variables to prove if correlations are truly cause-and-effect relationships. And do not assume that hard data can be stretched into universal truths. Keep these points in mind and you will avoid being duped by statistical skullduggery.

References

1) Charts in "We're all doing fine… on average" were taken from www.weather.com.
2) Charts and story in "You want more drama? Use less data!" are taken from Teaching Statistics: A Bag of Tricks, by Gelman and Nolan.
3) The example and chart from "Proving that storks carry babies" are taken from Statistics for Experimenters, by Box, Hunter, Hunter.
4) The chart and story from "Assuming the unknown" are taken from A Mathematician Reads the Newspaper, by John Allen Paulos.

Gordon H. Bell is president of LucidView, a consulting firm helping industry leaders increase marketing ROI and learn best practices in the art and science of marketing testing and analytics. He can be reached at www.lucidview.com or 888-lucidview.