

Why So Many Rollouts Disappoint

*By Jim Wheaton
Principal, Wheaton Group*

Original Version of an article that appeared in the February 4, 2002 issue of "DM News"

Have you ever wondered why so many of your rollouts are disappointing? Part of the answer lies with statistical sampling theory, and the other part does not. This article will focus on statistical sampling theory, and leave to a future article a discussion of the non-statistical reasons.

Universe vs. Test Panel Performance

When, testing, it is important to understand the difference between the performance of an entire rollout universe and that of an "Nth'd" test panel. The key thing to remember is that test panel performance is generally similar to, but almost never identical to, that of the rollout universe. An example will provide clarification:

Assume that a promotion to a universe of 200,000 yields a 0.80% response. This, of course, is the true responsiveness of all the customers to the promotion.

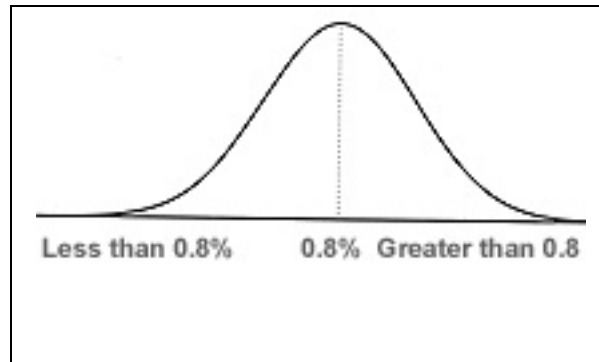
Now, assume that a random sample ("Nth") of 10,000 is contacted, and responds at 0.72%. This is the performance of the test panel and not the true performance of the 200,000 universe. And, what we as direct marketers ultimately are interested in is not this test panel responsiveness, but rather that of the full universe.

Finally, assume that four more Nth'd test panels of 10,000 are contacted, with response rates of 0.82%, 0.86%, 0.79% and 0.78%, respectively. These diverse results illustrate why, from the performance of test panels, we can never know for sure exactly what is the true performance of the full universe.

Although the universe performance is never exactly knowable from testing, we can have a degree of confidence that it will lie within a certain range of the panel results. And, it is statistics that quantifies this confidence and range. In essence, if the universe performance is the "bull's eye," then the test panel results are the shotgun pellets. A few will be way off, and one or two might hit the mark. Most, however, will be somewhere in between.

The Normal Distribution, Or "Bell Curve"

Now, let's assume that we draw many hundreds of random samples of 10,000 from our universe of 200,000, track their corresponding response rates, and then graph their frequency. Statistical sampling theory tells us that the graph will look something like the following (see next page for illustration of normal distribution with 0.8% midpoint):



This is what is known as a “normal distribution” or “bell curve.” The exact quantification of the distances between the various test panel response rates and the one true universe response rate would require a discussion of statistical sampling theory and formulas that are beyond the scope of this article. For our purposes, however, all we need to know is that many of the test panel response rates will be not too far from the true universe rate of 0.8%, and some will be moderately far off. In other words, most of the test panel response rates will cluster with reasonable proximity around the true universe rate.

A handful of the observations, however, will be relatively far away. These are what statisticians refer to as “outliers.” The larger the distance from the true universe response rate, the fewer the number of outliers. In other words, among the pool of outliers, there will be a select few that deviate particularly far from the norm.

This is the same effect that is seen in athletics. To make the starting football lineup at any large high school where the sport is popular, a player is likely to be somewhat of an outlier in athletic ability; that is, a reasonable distance to the right of the midpoint of the normal distribution for male athleticism. To make the starting lineup at a big-time football university, a player has to be even more of an outlier; that is, even farther to the right of the midpoint. And, to make the NFL, a player has to be an extreme outlier; that is, way to the right of the midpoint. The difference is that, in direct marketing testing, we try to avoid the outliers.

In a normal distribution, the shape of the right side of the curve will – by definition – mirror that of the left. Leaving the football arena and returning to our hundreds of test panels drawn from a universe with one true response rate of 0.8%, there will be as many “high response” outliers to the right of 0.8% as there will be “low response” ones to the left. Again, it’s just like football where, for every starter, there’s a klutz who keeps tripping over his own two feet.

True Winner Lists

Assume that a hypothetical direct marketer promotes 50 new lists a year. For each of these lists, let’s focus first on the one true universe response rate rather than on the variable test rates.

For most direct marketers, there are fewer winner lists than there are loser lists; that is, those whose one true universe response rate is above some acceptable cutoff. This makes intuitive sense because direct marketing is a mature industry, and postal rates have been increasing faster than inflation. Mailboxes are glutted, as thousands of catalogers, continuities, magazines, credit card issuers, and store-traffic-hungry retailers compete for the attention of consumers.

With fewer winners than losers, it is generally the case that the true aggregate universe response rate across all rental lists is below that which is required to generate an acceptable acquisition cost per new customer. In other words, if a typical direct marketer were to roll out all of his or her test rental lists, the overall response rate would be below the acceptable acquisition cost. In fact, it is likely that it would be well below this hurdle.

Let's assume that our hypothetical direct marketer's true overall response rate across all 50 test lists is 0.8%, and that 1.0% is required to generate an acceptable acquisition cost per new customer. Some of the lists will have universe response rates that are close to the average of 0.8%, others will be not too far off, and a handful will be a good distance from 0.8%. However, by definition, they will hover around 0.8% rather than the breakeven of 1.0%, which means that relatively few will be greater than the 1.0% required for rollout.

False Negatives

Now, let's switch our focus, from the one true universe response rate for each to the true winner lists, to the variable response rates that are inherent in test panels. Recall from statistical sampling theory that, for each of the lists, the test panel response rate will almost never be identical to the true universe rate. One-half of the time, the test panel will be greater than the true universe rate, or overstated. Likewise, one-half of the time, it will be less than the universe rate, or understated.

Therefore, some of the small number of true winner lists will appear from the test panel response rates to be below the 1.0% cutoff, or losers. These are "statistical mirages" that are referred to as "false negatives." The problem is that, if they are not rolled out or re-tested, they will never be identified.

False Positives

Likewise, some of the large number of true loser lists will appear to be above the cutoff, or winners. These comprise another form of statistical mirage known as "false positives." Only upon rollout will their response rates reflect the ugliness that is their reality. And, these will be the lists that will prove to be disappointing. Mix them in with the relatively small number of lists that are both true winners and whose test panel response rates are above 1.0%, and it is apparent why so many rollouts are disappointing!

Minimizing the False Negatives and False Positives

An important way to minimize the number of false negatives and false positives is to increase the size of the test panels. The higher the test quantity, the “narrower” the distribution of test panel results around the true universe response rate. Going back to our illustration, this means that the “tails” of the distribution will be shorter.

With a narrower distribution, fewer of the true winner lists will display test results below the 1.0% cutoff to create a false negative. Likewise, fewer of the true loser lists will have test results that exceed the cutoff, thereby creating a false positive. Taken together, the effect will be far fewer rollouts that are disappointing.

Of course, this leads to the question of how to determine optimal test panel sizes. This topic was covered in two recent articles, “How Big Should My Test Be?” (*DM News*, Oct. 1, 2001) and “Identifying Millions in Lost Revenue” (*DM News*, Jan. 7, 2002).

Jim Wheaton is a Principal at Wheaton Group, and can be reached at 919-969-8859 or jim.wheaton@wheatongroup.com. The firm specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data warehouses and marts. Jim is also a Co-Founder of Data University www.datauniversity.org.