

Answers to Four Common List Questions

*By Jim Wheaton
Principal, Wheaton Group*

Original version of an article that appeared in the November 4, 2002 issue of "DM News"

The October 1, 2001 issue of *DM News* included an article titled "How Big Should My Test Be?" The article contained a formula to assist in answering the question that I have been asked countless times as a direct marketing consultant.

Based on the number of emails that I received, the article struck a responsive cord in the direct marketing community. Based on the questions, it is clear that direct marketers have many questions about testing. Four of them will be answered in this article.

Question #1: Take a confidence level of 90%. Does that mean I can be 90% confident that the test panel response rate is what I will see on the rollout?

Answer: Unfortunately, no. Instead, it means there is a 90% chance that the result relates in some specified way to a reference point. Consider a scenario in which we want to compare a test and control treatment. Here, there is a 90% chance that the test treatment will be superior to the control if it is promoted to the entire universe, and a 10% chance that will be inferior to the control.

Often, when statisticians speak of "confidence," they are referencing two related but distinct concepts. The first is Confidence Level and the second Confidence Interval. This works as follows:

Assume we test an outside rental list and observe a 1.0% response. Based on the test and rollout quantities, and using the appropriate statistical formula, we determine there is a 90% chance that the rollout will perform between 0.9% and 1.1%. The "90%" is the Confidence Level, and the "between 0.9% and 1.1%" is the Confidence Interval.

Without the Confidence Interval as a reference point, the Confidence Level is meaningless. We can always be 90% confident of something having to do with a test result. The only question is what that "something" is. For example, with a sufficiently-low test quantity, we can be 90% confident that the rollout will perform between 0.2% and 1.8%. For a direct marketer, such a result is of little use!

Question #2: When running tests of significance between a mailed group and control, what level of confidence do you recommend? Many in the industry say 90%, although I have also heard 95%. Or, do you recommend something different?

Answer: In many ways, 90% is our industry's "gold standard." Generally, however, I use lower confidence levels, but in a carefully-controlled way. Often, I find anything over 75% intriguing enough for additional investigation.

The problem with 90% is that it makes it very difficult to beat the control. By definition, it is saying that there is a 90% chance the test treatment will be superior to the control if it is promoted to the entire universe, and a 10% chance that it will not be. This is a ratio of nine-to-one.

Consider a test treatment for which there is an 80% chance of being superior to the control if the entire universe is promoted. This is a ratio of four-to-one which, in my book, is pretty darn good odds. However, by using the 90% threshold, such a treatment will be rejected. The same is true for 75%, which translates into a ratio of three-to-one; that is, a 75% chance of being better than the control, and a 25% chance of being inferior.

The beauty of direct marketing is that we can be prudent and incremental in our decision-making. Generally, additional testing is appropriate, especially when confidence levels are relatively low.

It can even be argued that a treatment with 66% confidence is worth a retest in certain circumstances. At a ratio of two-to-one, it is twice as likely as not to be superior to the control. Direct marketers often use a new control piece across many, many promotions, comprising millions of total contacts. With these sorts of volumes, a successful test treatment can translate into substantial incremental revenues and profits.

Question #3: The formula you provided in your October 1, 2001 article for determining how big a test should be includes the "Rollout Universe Quantity." A formula that appears in several direct marketing books and articles does not include the Rollout Universe Quantity. Therefore, it provides results that are different from your formula. Please explain what is going on.

Answer: The formula that is most often used by direct marketers to determine how big a test should be does not contain what is called a "Finite Population Correction Factor." Implicit is the assumption that all rollout universes are infinite in quantity. If that were true, direct marketers would only have to identify one such successful universe to generate infinite revenues and profits!

In the real world, all rollout universes are finite. For niche direct marketers, most rollout universes are small, and often in the 5,000 to 200,000 range.

A "reduction to absurdity" will provide intuitive clarity as to why the Rollout Universe Quantity can dramatically affect the appropriate test panel quantity. Assume a test quantity of 10,000 and a corresponding rollout universe of 10,100. Common sense suggests, and statistical theory supports, that we can be much more confident of the test results than if the universe size were – say – ten million. This is because, with a test quantity of 10,000, essentially the entire rollout universe of 10,100 has been sampled.

The Finite Population Correction Factor plays a larger and larger role as the Rollout Universe Quantity gets smaller and smaller. For example, assume an expected (“observed”) test panel response rate of 1.0%, and that we want to be 80% confident the rollout (“actual”) response rate will be at least 0.9%:

- With an infinite rollout universe quantity, the required test quantity is 6,985.
- With a rollout of 200,000, the test quantity is 6,750.
- With a rollout of 20,000, the quantity is 5,177.
- With a rollout of 5,000, the quantity is 2,914.

Question #4: All of the testing formulas that I see have to do with response rates. However, dollars is what is most important to my business. Are there any formulas that deal with dollars? Or, can the response rate formulas be translated into dollars?

Answer:

The following example illustrates the problem:

Consider two test panels, one with a quantity of 6,010 and the other with 5,270. The first panel has a response rate of 1.29% and the second 1.03%. Using a statistical formula that we will not discuss in this article, we can be 90% confident that, upon rollout, the first response rate will be greater than the second.

In scenario #1, the two panels have Average Order Sizes of \$110 and \$90, respectively, which translate to corresponding Dollars Per Piece Mailed of \$1.42 and \$0.93. Here, Average Order Size has accentuated the difference in response rate.

In scenario #2, the respective Average Orders Sizes are \$80 and \$99, which translate to \$1.03 and \$1.02. Here, Average Order Size has essentially neutralized the difference in response rate.

Unfortunately, the calculation of dollar-driven formulas is much more involved than response-driven formulas. To do so requires data manipulation and statistical capabilities. It is also problematic to derive dollar-driven formulas from response-driven formulas. However, an important consolation is that response-driven formulas are much better than nothing!

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.