Statistical Significance in Evaluating Advertising Response Rates

Background

Despite the rather off putting title of this brief, we hope to provide a basic understanding of the concept of statistical reliability and hope to do so in a way that non-math/stats geeks will understand.

The ability to generate response data gives marketers a great tool to maximize their advertising activity. But we have seen marketers who don't recognize when the data is unreliable and worse; begin to make important decisions based on what is clearly unreliable data. Surprisingly we've had discussions with advertising professionals who simply didn't understand the entire concept. We get it: numbers on a spreadsheet appear concrete and since these numbers are often expressed to four decimal points they give the appearance of exactness. But this exactness can be anything but.

We promise not to mention statistical terms like standard deviation[1] etc. There is a tiny bit of math ahead but we only included it for those that want it. If you find the math off-putting just ask us for our spreadsheet. We'll happily send it to you and then you will only need to input the data and read the results.

The Problem in a Nutshell

The value of results data is found primarily in its predictive value. In technical terms, reliability is the overall consistency of a measure. Hypothetically we know creative treatment A has performed better than creative treatment B because the data shows a better set of metrics for creative treatment A vs. B. What makes this data valuable—and actionable—is our belief that it is predictive. But what if it isn't predictive? And why might it not be predictive?[2]

Rather than think about the numbers in abstract, let's think about how people might respond to an ad. Let's start with the recognition that response rates for online advertising are generally measured in

---

[1] Oops. We promise not to do it again.
[2] There may be a variety of reasons why it might not be predictive including environmental factors unrelated to the effectiveness of the different treatments. For our purposes here we'll restrict our thinking to the relative performance of the ads without any consideration for uncontrollable factors.

fractions of percentage points. One study has click thru rates[3] averaging 0.06%.[4] If we use this 0.06% CTR then six people out of every 1,000 will click. But this doesn't mean that of every 1,000 viewers exactly six will click. We may get three responses in the first 1,000 followed by nine in the next 1,000. This sporadic nature of the clicks is problematic. As the sample size grows the "lumpiness" of the results tends to average out.

So When Does the Data Become Reliable?

The short answer to what makes data more reliable is a larger sample size. But that begs the question: how big does the sample size need to be? The answer: it depends on the results themselves; sadly there is no simple answer to this question. The closer the comparison between two or more options the larger the sample size needed.[5] Let's look at some exciting real world examples.

Peppermint Patties, Tootsie Pops and a Poke in the Eye

If we ask people if they would prefer a Peppermint Pattie or a poke in the eye with a stick, we'd find out pretty quickly that the overwhelming majority prefer the Peppermint Pattie; we wouldn't need

to ask several thousand people. We'd be confident in our data after the first 10 responses and by 20 or so we'd be extremely confident. But what if we asked them if they preferred a Peppermint Pattie to a Tootsie Pop? Let's assume that the final answer to this question is a preference for Peppermint Patties by a margin of 51% to 49%. In the first 10 responses you could get eight saying Tootsie Pops and only two opting for the Peppermint Pattie. If you used this small sample size as a predictor you'd act on incorrect data (even though the spreadsheet looked authoritative). Ask another 10 people and you might have the result at 12 for the Tootsie Pop and eight for the Peppermint Pattie. Ask 1,000 and you'd likely get very close to final outcome of 51% to 49%. You wouldn't need to ask anywhere close to 1,000 people about the Peppermint Pattie vs. the poke in the eye preference. We'll look at the actual statistical significance numbers in a moment but for now we hope we've conveyed the concept that the closer the comparison the greater the sample size needed to confidently determine a comparative winner.

---

[3] For the moment we're dealing with click thru rates but the same analysis applies to downstream metrics as well.
[4] Note that we would consider this rate a failure for most of our clients. This is where good targeting and placements make a huge difference. Targeting and context matter. Source: http://www.smartinsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/
[5] There are some other technical points here as well not the least of which is the required confidence level. But we've promised to keep this relatively simple.

So How Do We Know When the Data is Reliable?

The only way to know is to do some math. To deliver on our promise to make this brief usable for non-math/stats geeks we've avoided any serious math to this point. But some will want to know the formula we use. If you don't want to get into this detail we can always share our spreadsheet with you and then you only need to enter the data to see the results.

The formula:

$$(((Response\ Rate*(1-Response\ Rate))/Sample\ Size)^{0.5})*1.96$$

We admit this is a bit clunky to read so what we have here in English is:

The Response Rate (in percent) multiplied by 1 minus the response rate.
Divide that by the sample size.
Find the square root of that number.
Multiply by 1.96.

This will give you the margin of error expressed as a percent. This margin of error will get smaller as the sample size grows.

Real Life Examples Come to Life

In our Peppermint Pattie vs. the sharp stick scenario let's assume that out of the first 10 respondents one person prefers the sharp stick.[6] If you examine the graph below you'll see the results presented with the upper and lower bounds. Even after just 10 responses we can confidently determine that the Peppermint Pattie will be preferred over the sharp stick.
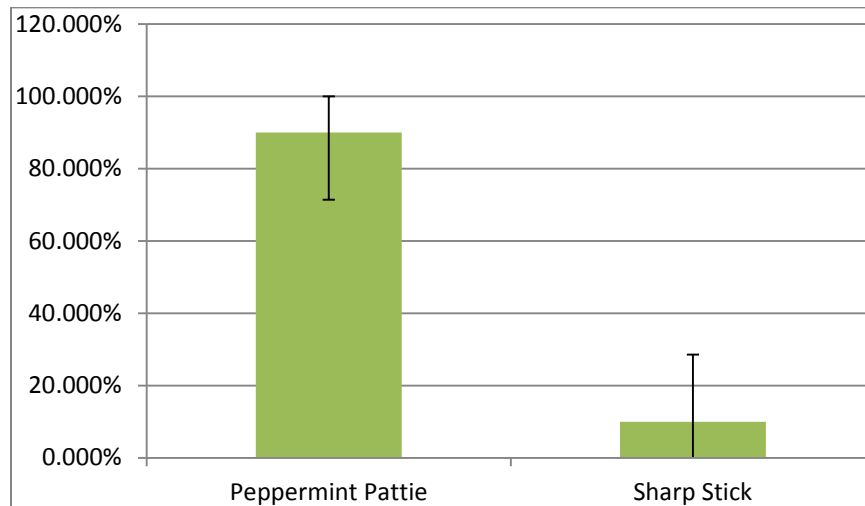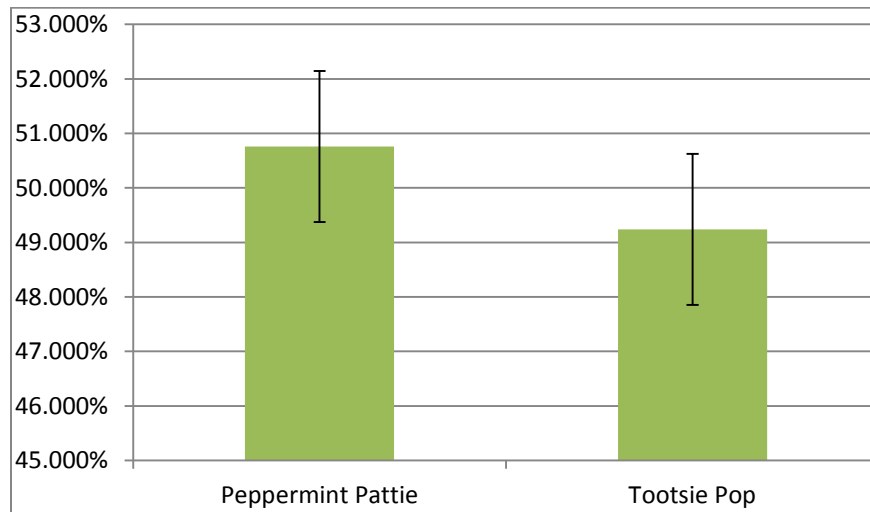

Sample size of 10 Margin of error plus or minus approximately 2. But even with a sample size of just 10 we can confidently predict that however large the sample size the Peppermint Pattie will be preferred over the poke in the eye with sharp stick.

---

[6] You can draw your own conclusions here as to why that preference. Perhaps the respondent didn't hear the question. Perhaps he's a very contrary individual. Or perhaps he has a potentially fatal peppermint allergy.

Let's see what happens when the comparative results are closer. We've asked 5,000 people to tell us their preference between the Peppermint Pattie and the Tootsie Pop. Here are the gathered results.

| Choice	Data	Percent	Margin of Error
Peppermint Pattie	2,538	50.76%	+/- 1.39%
Tootsie Pop	2,462	49.24%	+/- 1.39%

Sample size of 5,000 Margin of error plus or minus approximately 1.39%. Even with a sample size of 5,000 we can't say with confidence that the Peppermint Pattie is our winner. Notice the error bars. If the Tootsie Pop were to move upward towards the top of that range it would be nearly equal to the Peppermint Pattie. And, if the Peppermint Pattie were to move just bit toward its lower range then Tootsie Pop could be our winner here despite being behind with the first 5,000 respondents being tallied.

Based on these numbers discerning consumers prefer the Peppermint Pattie to the Tootsie Pop by about 1% point. But what is our margin of error here? Using our formula above, we see that the margin of error is 1.39%. This margin of error is less than the difference between these two options so we don't have data that can be confidently relied on. If these two choices were creative treatments we wouldn't recommend any optimization; we'd continue to serve them equally.

Let's change the scenario a bit. Let's assume we're a candy retailer and we're advertising our two favorite candies and here are our impressions and click data with the calculate margin of error.

Creative	Impressions	Clicks	CTR	Margin of Error
Peppermint Pattie	1,482,563	2,538	0.17%	+/- 0.007%
Tootsie Pop	1,496,347	2,462	0.16%	+/- 0.006%
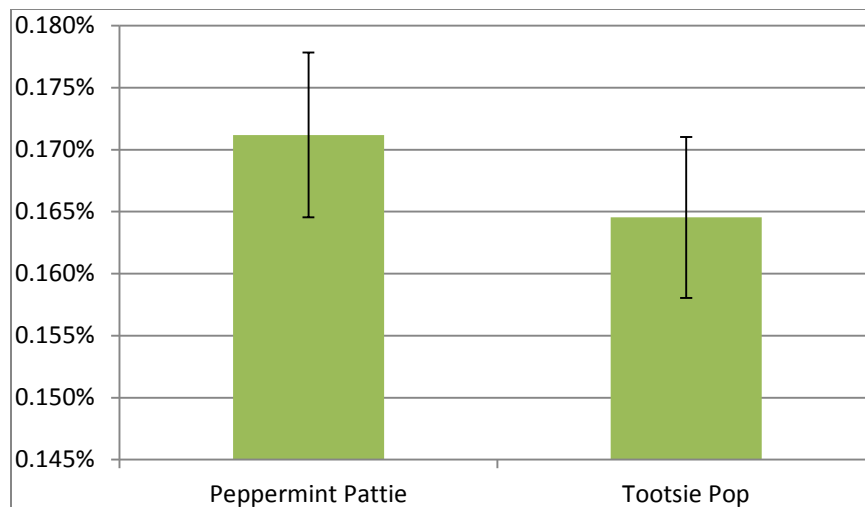
The results here are again extremely close. The margin of error is less than 0.01% but that is sufficient to reverse these results. This potential extra 0.01% response rate is significant. Multiply all of these numbers by 10 (meaning the impressions and clicks) and the CTRs remain identical but the margin of error becomes 0.002%, which means we would have sufficiently reliable date to have confidence that the Peppermint Pattie ad is our best performer between these two.

Why Does this Matter?

The difference between 0.17% and 0.16% is just a basis point or one percent of one percent. How important is that in the grand scheme of things? We believe it is quite important. That is approximately a 6% difference between these two outcomes.[7] If you knew you could increase your overall results by 6% simply by deploying the more responsive creative unit you'd happily do so. This would mean a 6% reduction in your cost per metrics. But what if the initial data suggested the lesser of the two ads was performing better because the sample size was too small? Ouch. You'd make a decision to deploy inferior creative because you didn't have a sufficiently large sample size.

Other Metrics

To this point we've really only addressed click thru rates. But your marketing and advertising activity involves post click activities and likely these metrics are of greater importance to you than the CTR. The concepts we've explained here apply to these other metrics and we advise that it is critical to keep the concept of statistical reliability in mind for these other metrics. It is here where many marketers are quick to make decisions without understanding that they might be a long way from having usable data. To return to our candy retailer, let's assume that of the visitors to the retailer's site that came from the two ads (Peppermint Pattie and Tootsie Pop) some of them purchase one of these delicious treats. Before we do any calculations or even look at the data, we know one thing: The visitors that purchase

---

[7] Without getting too mathy, 0.16% is 5.88% less than 0.17% but 0.17% is 6.25% more than 0.16%. The difference here is the denominator as the numerator—.0001—stays the same in both cases.

will be a subset of the total visitors. Further, it is extremely likely that the percentage of purchasers will be a small subset of these visitors. Which places us in the position of having a small response rate which, as we've seen, means it takes a larger sample size to get to reliable data. We won't belabor this point with more results or calculations but be aware that if you're making decisions based on post click behavior you generally need a LOT of data (by which we really mean sample size) to generate statistically reliable data.

This can be important if you're testing landing pages. Ideally you would run a multivariate test of various landing page elements rather than a simple A/B test. Unless you have a highly trafficked landing page it is very difficult to generate a sufficient sample size. This is particularly true for B to B advertisers who are almost always dealing with small numbers. In the event you find it difficult to generate sufficient traffic then we suggest starting with a simple A/B testing program and as you determine your winner test a second element in a subsequent A/B test.

The Orbis Marketing Margin of Error Spreadsheet

Did we mention our margin of error spreadsheet? We certainly meant to. We may have hinted that you can get your hands on it simply by asking us. For your free copy give us a call or send us an email. You can find our contact details on our web site www.orbismarketing.com.

About Orbis Marketing

Founded in 2001, Orbis Marketing has offices in Chicago and Los Angeles. Since our founding we've purchased media on behalf of over 50 different clients from over 100 different publishers. Our media planning and buying services are driven by our results oriented approach to advertising. The tracking of advertising results coupled with our proprietary research allows us to know where to advertise, and critically, what rate to pay.