Making Optimization Significant: THE ROLE OF STATISTICAL ANALYSIS



by Paul Benson, Mark Casali and Dessislava A. Pachamanova

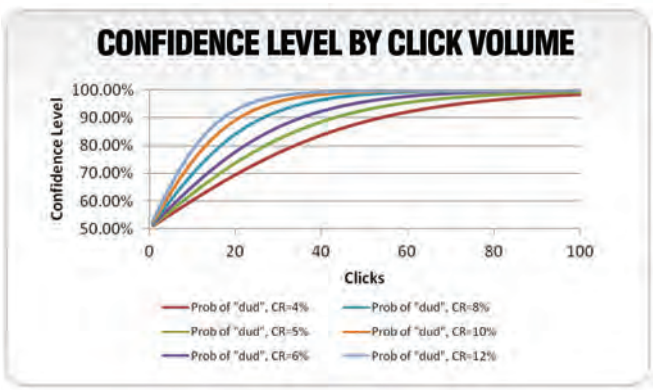
Search marketers are constantly faced with the decision of when to act – when to pause a non-converting keyword, when to end an A/B test, or even when to address a fluctuation in account-level metrics. Most of us address these decisions based on our past experiences, our expertise, or even a sophisticated hunch. Sometimes we get the decision right; sometimes not. This subjective approach to decision-making can jeopardize campaign performance.

In this article, we provide guidance on this predicament. By introducing a suite of statistical models, we aim to objectify the decision-making process and give you tools to approach the question of when to act, thereby reducing the margin of error associated with key choices search engine marketers face every day.

THE AGONY OF CHOOSING SHUT-OFF POINTS

A "shut-off point" refers to the point at which a marketer pauses or suspends a keyword, ad, ad group, or campaign due to poor performance. Traditionally, marketers approach these situations by making an educated guess on whether the account attribute in question has run long enough. But when has a keyword reached a level of clicks that is statistically representative of its ultimate level of performance?

Let's consider a campaign with a cost-per-lead (CPL) goal of $25. A keyword within this campaign has not converted. How many clicks (or what level of spend) do we have to reach, without converting, to get to a point where we can confidently pause that keyword due to poor performance?



Based on a Bayesian model, the graph demonstrates the relationship between shut-off point, conversation rate, cost-per-lead goal, and desired confidence level.

Let's assume our average cost-per-click (CPC) is $1. To reach a CPL of $25 at an average CPC of $1, this keyword would be required to convert at 4% (Required CR = CPC/CPL). The graph traces the shut-off point for a keyword with no conversions that is required to convert at 4%. If we are looking for 90% confidence, that shut-off threshold is 54 clicks. In other words, if after 54 clicks the keyword has not converted, the keyword has a 90% chance of being a dud.

This may go against conventional wisdom, which suggests that if you've reached your CPL goal in spend ($25) and not yet converted, you should turn that keyword off or take steps to optimize it. However, based on the data above, you would only reach a confidence level of 74% after accumulating $25 in spend. This translates to a higher likelihood that you will accidentally suspend a performer, and therefore sacrifice additional conversions and better overall performance.

Naturally, the higher your required level of confidence that a keyword is a dud, the more clicks you need. Thus, if a specific advertiser is not comfortable with a 90% level of confidence and wants a 95% level of confidence, the required number of non-converting clicks will increase to about 73. Logically, the higher the required

conversion rate, the quicker you will reach your shut-off point for a keyword with no conversions. By associating a level of confidence with the number of clicks, a marketer has more information for deciding on a shut-off point. You can prevent the campaign from wasting advertising spend in an area that is not likely to convert and make sure the campaign has run long enough so that there is little chance of missing out on revenue from future potential conversions.

ASSESSING THE OBSERVED SIGNIFICANCE OF A/B TEST RESULTS

As A/B testing becomes more commonplace for search engine marketers, it's increasingly more important to understand whether your test results are statistically significant and not just due to random performance fluctuations. Google has enhanced its functionality, namely through ACE (AdWords Campaign Experiments) and Website Optimizer, to allow advertisers to test in a more controlled environment and identify when results have reached statistical significance.

While these tools are tremendously helpful, there are a few limitations of which you should be aware. First, you need to have either Google Conversion Pixel or Google Analytics tracking in place. Second, Google uses statistical tests that do not work well for small sample sizes. And even if you have the right tracking in place and large sample sizes, you will not have any insight into how close you are to reaching statistical significance or how much more money/time is required.

To overcome these restrictions, you can assess statistical significance on your own using one of several tools. For large sample sizes, chi-square tests, z-tests, and Poisson tests are perfectly suitable. For smaller sample sizes, it's important to leverage more applicable tools, such as Fisher's Exact Test or Poisson Exact Test. All of these tools can be built relatively easily in Excel, and before the test even launches, you will be able to project more accurately the time and money required to run the test. You will also have insight into how much longer a test will need to run given the data you've collected, helping you set expectations internally or with clients.

Whichever path you choose to follow, it's imperative you avoid relying on intuition alone to judge the significance of results. To further illustrate this, consider the following two hypothetical test results with corresponding clicks, conversions, and conversion rates for each landing page.

Example No. 1

Landing Page	Clicks	Conversions	Conversion Rate
LP 1 (Control)	204	22	10.8%
LP 2 (Test)	204	16	7.8%

In this example, the test results indicate a 27% difference in the conversion rates for the control and the test landing pages. This result seems meaningful, and yet a chi-square test run on the results gives a p-value of 30%. Statistical tests usually require a p-value of 5% or less to deem the results statistically significant. These results are not statistically significant and the test needs to continue to run.

Example No. 2

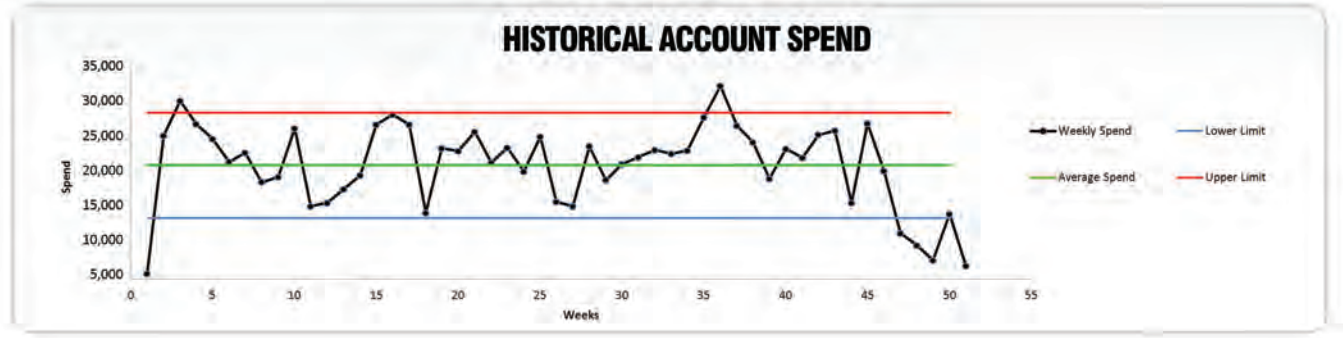
Landing Page	Clicks	Conversions	Conversion Rate
LP 1 (Control)	816	88	10.78%
LP 2 (Test)	816	64	7.84%

In contrast, these results illustrate when a 27% difference in conversion rates can be statistically significant. The p-value in this case is 4%, a relatively small number. The difference here is a larger number of total clicks, which allows us to state a conclusion with greater confidence.

The critical takeaway from these two examples is that it's extremely difficult to determine statistical significance on your own. Leveraging statistical tools will help you make better decisions and save your business time and money. Furthermore, you will know how many more clicks you will need before you reach statistical significance.

EVALUATING ACCOUNT FLUCTUATIONS

Addressing performance fluctuations in an account requires a different approach. Marketers are often faced with the challenge of deciphering spikes and dips in variables like spend, clickthrough rate, conversions, etc. At what level do these fluctuations in an account represent a significant issue as opposed to a normal variation? Control charts can help answer this question.



The above graph illustrates the weekly ad spend for a client across an entire year. It helps determine when fluctuations are unusual by including data outside either the upper control limit or the lower control limit. You can also use control charts to identify trends in data (increases or decreases in a given metric over time).

Control charts are built by identifying the mean for a data set and then establishing upper and lower limits by calculating a multiplier that corresponds to a given confidence level. For example, to achieve a 90% confidence level, you will need to use a multiplier of 1.65 standard deviations. Since control charts are designed to help you focus on troubling data, one standard deviation is typically used, which corresponds to a 68% confidence interval. In other words, this puts the focus on data points that fall outside roughly 70% of your total data set. When evaluating a variable like spend, be sure to make a note of non-representative data. For example, you may have paused campaigns on a holiday, which dramatically reduced a week's total spend. Other changes could be a date on which you made widespread bid changes or added numerous keywords to your campaign.

As a first step, mark any non-representative periods of time on the graph, and then look into the reason. If the cause is identifiable and to a specific action on your part (such as pausing the campaign), you could remove the data point. Be careful, however, not to remove data points just for convenience's sake. If a data point is non-representative, it may also hold a clue about something that went very wrong (or very well) with your campaign. Such data points could be kept in the data set, so that they can incorporate relevant information should a similar event arise in the future. Once the causes of variation in the data have been studied and the set of data to use finalized, control charts will enable you to establish upper and lower ranges to apply to the automated rules or alerts available in the AdWords interface. For example, if you determine that a weekly spend above $5,000 results in an outlier, you can set an alert in AdWords to send you an email any time your costs exceed this amount. Similar alerts and rules can be set for other metrics.

CONCLUSION

Moving forward, we are confident that statistical tools will profoundly impact everyday optimization strategies. Instead of approaching decisions about advertising campaigns based on experience or hunches, statistical tools let us apply objective guidance to narrow the margin of error in our choices. Ultimately, these tools will help us increase revenue, gain efficiencies, properly set expectations regarding anticipated performance, and become more empowered decision-makers.

As the co-founder of Synapse SEM (www.synapsesem.com), Paul oversees the strategic direction and growth initiatives for Synapse clients. Paul has hosted webinars, written articles and conducted trainings focused on advanced paid search strategies and tactics to maximize efficiency.

Mark Casali is a registered CPA in Connecticut. As the co-founder of Synapse SEM, Mark has strived to revolutionize the online marketing industry by applying statistics and other advanced data analysis techniques. Mark holds a BA and an MS in Accounting from Babson College.

Dessislava A. Pachamanova is an Associate Professor at Babson College in Wellesley, MA. Her areas of expertise include statistics, robust optimization, simulation, and financial engineering. She holds an AB in Mathematics from Princeton University and a PhD from the Sloan School of Management at MIT.