A Better Statistical Method for A/B Testing in Marketing Campaigns

Scott Burk

Marketers are always looking for an advantage, a way to win customers, improve market share, profitability and demonstrate value to the firm. Many forms of market research are about trying new things and seeing what works and what doesn't. This may be called campaign marketing, advertising and promotion research or e-marketing. It is about testing ideas and attempting to drive return on investment. A very common method of testing these ideas is A/B testing. The most common way to compare the results of an A/B experiment is a simple t-test. We present a better way to perform the statistical analysis associated with these experiments. Control charts have been used in manufacturing and service industries for years. By adapting these robust methods marketing professionals can add scientific rigor to their creative process. We provide practical illustrations to the interpretation, construction and comparison of control charts to traditional forms of analysis.

Keywords: Control charts, A/B testing, quantitative marketing research

Introduction

A very common method of testing in advertising, promotion or e-marketing is the A/B split. In A/B testing, you normally introduce two different versions of a design and see which performs the best. By design we mean creative layout, promotion or special offer. For our purposes here we will use the terms design and treatments synonymously. This has been the classic method in direct mail where companies often split their mailing lists and send out different versions of a mailing to different recipients. This experimental data is often collected in 'batches' meaning that test groups and control groups are considered treatments and then data from those treatments are complied as separate groups or independent samples.

The internet has made this form of testing even easier where it is simple to present different page versions or different promotions to different visitors. For our purposes we will be considering e-marketing website design where we are introducing different creative layouts or promotions via the internet. However, the testing methods discussed here are also appropriate for many other areas of marketing research whenever runs can be conducted in a serial fashion.

According to Nielson (2005), compared with other promotion research methods, A/B testing has four huge benefits:

1. It measures the actual behavior of customers under real-world conditions. You can confidently conclude that if version B sells more than version A, then version B is the design you should show all users in the future.
2. It can measure very small performance differences with high statistical significance because you can throw boatloads of traffic at each design.

3. It can resolve trade-offs between conflicting guidelines or qualitative usability findings by determining which one carries the most weight under the circumstances. For example, if an e-commerce site prominently asks users to enter a discount coupon, user testing shows that people will complain bitterly if they don't have a coupon because they don't want to pay more than other customers. At the same time, coupons are a good marketing tool, and usability for coupon holders is obviously diminished if there's no easy way to enter the code.
4. It's cheap: once you've created the two design alternatives (or the one innovation to test against your current design), you simply put both of them on the server and employ a tiny bit of software to randomly serve each new user one version or the other. Also, you typically need to cookie users so that they'll see the same version on subsequent visits instead of suffering fluctuating pages, but that's also easy to implement. There's no need for expensive usability specialists to monitor each user's behavior or analyze complicated interaction design questions. You just wait until you've collected enough statistics, then go with the design that has the best numbers.

The ensuing statistical analysis for this type of A/B experiment is typically a simple t-test to compare means between the two test versions. Averages or means are often compared as pre and post treatment measurements to determine if there is a statistical difference in the variable or variables of interest. The data types involved in these experiments are normally numerical and binary. Examples of numerical data types are revenue and margin. An example of a binary data type is conversion, that is, whether someone responds or does not respond to a promotion.

We propose an alternative method of analysis which can be performed more quickly with a smaller sample size, yields results that are more intuitive and can be easily explained, can easily be varied for different experimental conditions and are more flexible to analyzing different data types. This method is using control charts.

What is a control chart?

Control charts were invented by Walter A. Shewhart while working for Western Electric (the manufacturing arm of Bell Telephone). Dr. Shewhart created the basis for the control chart and the concept of a state of statistical control by carefully designed experiments. Dr. Shewhart concluded that while every process displays variation, some processes display controlled variation that is natural to the process, while others display uncontrolled variation that is not present in the process causal system at all times. For more information on this history of control charts see Adams and Orville (1999).

The original purpose of a control chart was to determine whether a manufacturing or other process was performing or behaving as expected. From a manufacturing standpoint if a process was performing as intended it was ‘in control’ if something happened so that the process had changed the process was considered ‘out of control’. Therefore the original intent of control charts was for things to be in control.

Although the original intent of control charts was to monitor a process we can use the inherent nature of a control chart for experimental testing purposes. We can allow a process to run under

normal conditions and then intervene (perform a test). If the process stays in control the findings were not statistically significant. If the process is determined to be out of control, the process has changed and our test results are significant.

Donald Wheeler has coined some useful terminology in intuiting the nature and use of control charts. In his book (Wheeler 2000) he calls them "process behavior charts" because we are really talking about whether a process is behaving in a consistent fashion or the behaviour, thus the underlying process has changed.

As to construction, a control chart is simply a time series plot with certain limits that have been calculated. One simply compares a data series to these 'control limits' and determines whether a statistically significant event occurred. There are numerous kinds of control charts for different types of data and situations, but the basic rules and interpretation of these charts is the same. Additionally, these charts are useful to determine if external forces such as negative press releases or even a meteorological event such has a hurricane has an affect on customer behavior.

Example 1 – A Simple t-test vs. a Control Chart

Let's examine a simple example contrasting a typical t-test method of analysis to the use of control charts. Suppose we have a small e-commerce business that sells retail products on-line. The average conversion rate runs about of 2%, but varies from day to day and seasonally. We want to see if a test design will outperform the current layout. We split the traffic to the site and half of the visitors are routed to the new test page and half get routed to the traditional layout. We receive about 10K visitors a day to the section of the website where our test is running so we run this test about twenty days until we receive about 100K visitors to the new page and about 100K to the traditional (control) page.

We collect the data and the control group generates a total of 1,996 sales and the test (new design) group generates a total of 2,211. We perform a simple t-test and the difference is highly statistically significant, $p < 0.001$ (the data here was simulated from a binomial distribution with 100K observations in each group with a 10% effect difference. For the control group, $\pi = 0.02$ and for the test group $\pi = 0.022$).

Suppose now we take the same underlying data model and allow the control group to run for the first ten days and collect the same data for the control group. On day eleven we 'turn on' our test site. This differs in that we are no longer splitting the data as visitors arrive, but are collecting data in a serial fashion. We collect the data shown in Table 1.

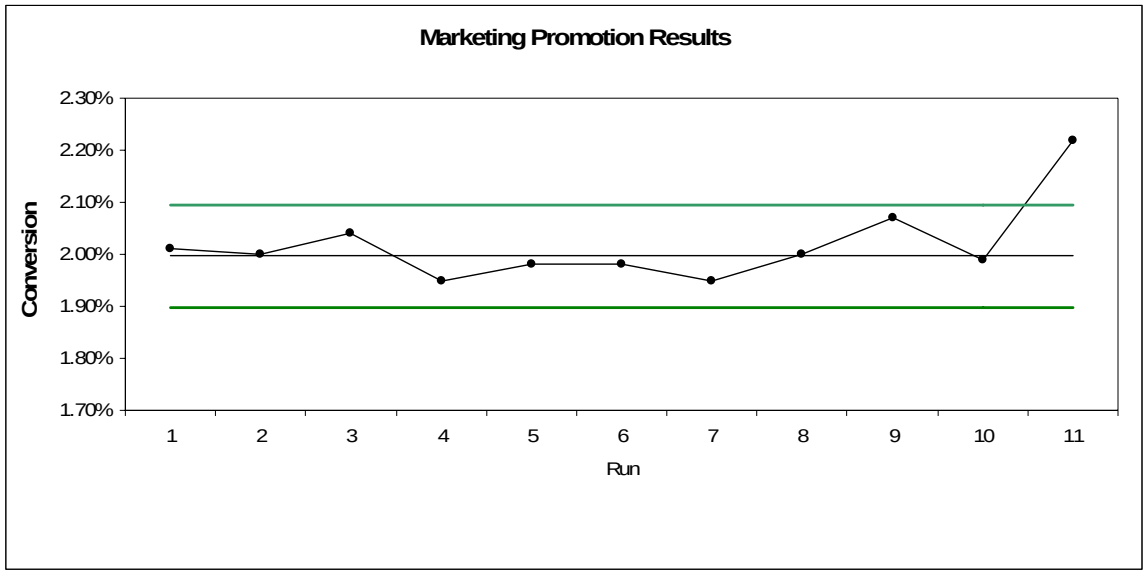
We see that there is a definite increase on day eleven, but is this result significant? A control chart is an excellent way to answer this question. And as we will soon see it is the only effective way to answer this question at this point in data collection.

Table 1. Conversion rate data for daily campaign runs

Run Number	Conversion Rate
Day # 1	2.01 %
Day # 2	2.00 %
Day # 3	2.04 %
Day # 4	1.95 %
Day # 5	1.98 %
Day # 6	1.98 %
Day # 7	1.95 %
Day # 8	2.00 %
Day # 9	2.07 %
Day # 10	1.99 %
Day # 11	2.22 %
N = 11	

Figure 1 illustrates the data in a control chart. The data is plotted in a time series with three additional lines. The straight line is the average and is called the center line. There are two boundary lines, the upper straight line being the upper control limit and the lower straight line being the lower control limit. The idea behind a control chart is to determine whether 'rises' and 'falls' in the data series are due to normal variation or 'special cause' variation. Normal variation is present in any process. By special variation we mean that the process has shifted up or down, that is a statistically meaningful change has occurred.

Figure 1. Individuals control chart for the marketing promotion data



There are several ways to determine if a significant event has occurred, but the simplest is whether a point rises above the upper control limit or falls below the lower control limit. Here

we see that on the introduction of our test design the conversion rate rose above the upper control limit. This means we had a statistically significant increase in conversion on day eleven and our test design is significantly better than the control design. Note: We have performed the same A/B test in eleven days where it would have taken 20 days with a traditional A/B split and subsequent t-test.

There are many software packages that will construct control charts, but they are easy to construct by hand or in a spreadsheet. The control chart in Figure 1 was produced in Microsoft Excel®. We will give a simple overview of the charts construction. For a comprehensive reference on how to construct a similar type chart (this chart is known as an individuals chart) as well as many other types of control charts see Wheeler and Chambers (1992) or Montgomery (2004). The center line is just the average of the data series. The lower control limit is the average less two and one half times the standard error and the upper control limit is the average plus two and one half times the standard error (significance level = 0.05). The calculation of the standard error here is NOT simply the standard deviation divided by the square root of the number of observations. The easiest way to calculate the standard error for this control chart is to perform the following steps:

1) Take the absolute moving deviation ABS ($X_{n+1} - X_n$ for $n=1$ to k where k is equal to the number of runs).
2) Take the average of the absolute deviations.
3) Divide this average by 1.128 (the appropriate constant here).

Table 2 demonstrates how easy it is to perform these calculations in Excel®. The data illustrated is the data which appears in Figure 1.

Table 2. Control chart calculations demonstrated in Excel®

Run #	Conversion Percent	ABS Deviation	Average of Deviations	0.000444
Day # 1	2.01 %	NA	Constant	1.128
Day # 2	2.00 %	1E-04		
Day # 3	2.04 %	0.0004	Standard Error (SE)	0.000394
Day # 4	1.95 %	0.0009		
Day # 5	1.98 %	0.0003	SE * 2.5	0.000985
Day # 6	1.98 %	0		
Day # 7	1.95 %	0.0003		
Day # 8	2.00 %	0.0005	Average	1.997 %
Day # 9	2.07 %	0.0007	Lower Control Limit	1.898 %
Day # 10	1.99 %	0.0008	Upper Control Limit	2.096 %
N = 10				

It should be noted that we are testing for a change in the conversion rate or in other words, whether there is a significant change in central tendency. A control chart can easily be constructed that tests if there is a significant change in the variation between periods. We will

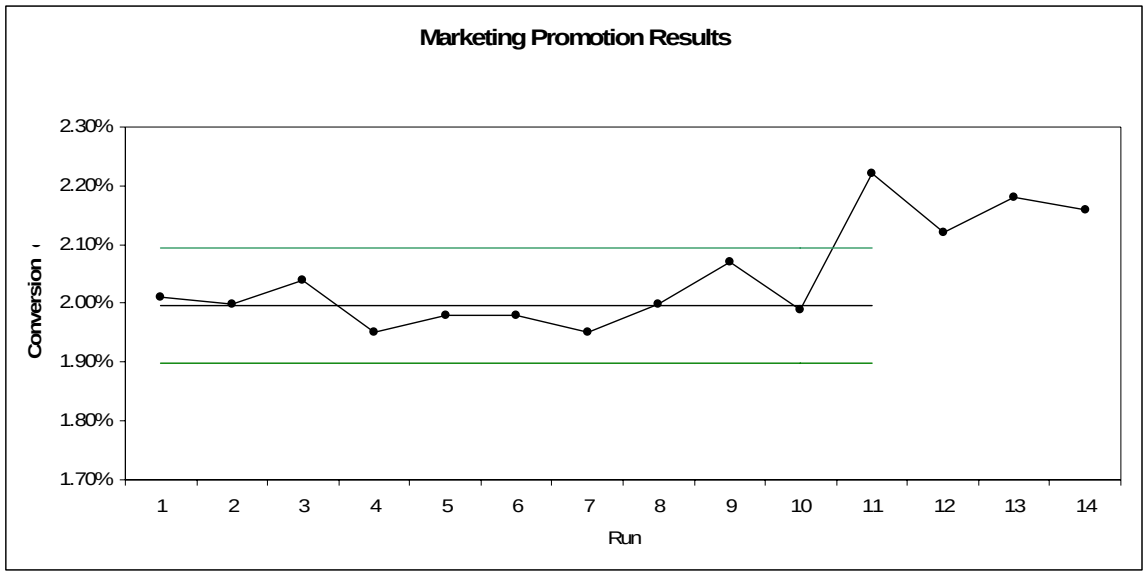
forgo this detail for the purpose of our discussion here, but one can easily find the construction of a chart for this purpose in one of the references listed above (see range or s-chart).

Suppose instead of constructing the control chart above we perform a simple t-test comparing the 100K visitors (ten days) in the control group to the 10K in the test group (one day running). What would the results be? For the test group we have a proportion of 1.997% and the proportion for the test group is 2.22% using a standard t-test with pooled variance. The result here is not statistically significant at a 0.05 significance level with a p-value = 0.146 One of the reasons that the control chart is more sensitive is that it takes the time series component into effect. Traditional A/B testing using t-tests batches the data into 2 groups and compares the means of these groups. Yet a control chart measures the signal at the very point of an intervention and that allows for a much more sensitive test.

Extending the Application – First Results Found Significant

We have illustrated an example where using a control chart leads to a more sensitive statistical test on a marketing campaign and we have significant results. The next step is to continue using the improved design and to validate the results. We continue plotting the data and see if the results hold as one might argue that there was 'something' special about day eleven (see figure 2). However this is not a problem using the control chart method, but would be using the simple t-test.

Figure 2. Individuals control chart with three additional daily runs



This method lends itself well for future testing. After we have about eight to ten daily runs with the new design we recalculate the control limits and then we are ready to test the next idea for a promotion or layout. Therefore treatments can be tested rapidly in succession in a serial fashion. Each improvement serves as a baseline for which new alternatives can be tested against.

To summarize the advantages of the control chart:

1. Requires less data.
2. Can determine significant results sooner, in fewer runs.
3. Uses the time series component of the data and is a more sensitive test.
4. Using the control chart method of testing lends itself to subsequent testing.

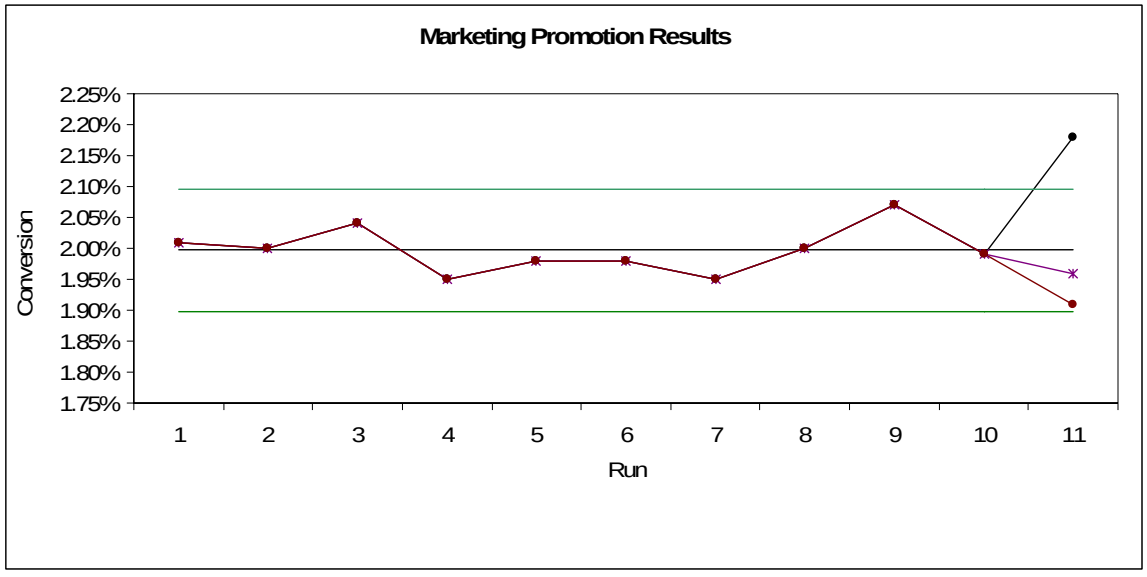
Extending the Application – First Results Found Insignificant

Suppose we run the experiment as above, but we do not find that our idea is an improvement. Do we have to start all over? Fortunately, the answer is 'no'. The nice thing about this form of testing is that we can run experiment after experiment in a serial fashion. We can continue with the same testing group and try another creative change, incentive, etc. We continue this until we find that there was a significant change. The idea is to continually keep a containment group for further testing and refining.

Example 2 – Multiple Factor Testing

One can easily extend this method beyond A/B testing to testing multiple factors or designs at one time. By factors we mean an independent treatment variable whose settings (values) are controlled and varied by the experimenter. So an individual design or treatment is composed of many factors, most of which we will not manipulate. We can also test for possible interactions between factors. Suppose we have a baseline example as before, but now instead of testing a single test design, we test three designs at once. We can construct a control chart using the first ten baseline points and then plot the additional three series (see Figure 3).

Figure 3. Individuals control chart for multiple factors



Suppose we have design A which is a form of subject line personalization, design B which includes either the contents left in the shopping cart or alternatively products based upon a personalization algorithm if nothing is left in the cart and Design C which is the presence of the both A and B i.e. we want to test the interaction of the subject line and product personalization.

Two of the test designs (A and B) were not found to be statistically significant. However for design C the interaction of A&B was found to be statistically significant. This is a powerful illustration of this form of testing. We have found an interaction of two factors to be significant whereas neither factor by itself was significant.

Consider the testing procedure here that is normally employed. A simple t-test is no longer applicable to this scenario. We could construct four groups (the control and three factors) and construct *m-1* pair wise t-tests in this case six or we could perform some form of ANOVA (Analysis of Variance) procedure. However ANOVA is considered a more sophisticated statistical technique and many practitioners find it more difficult to understand. Also, traditional ANOVA deals with numerical data and all our examples in this discussion have been binary (see Almquist & Wyner, 2001). This typically requires additional knowledge and methods like something involving a transformation of the data prior to analysis.

Summary

We have shown that control charts are a very effective way to test A/B experimental results. They have many advantages over the traditional method of performing t-tests. They are flexible, intuitive, sensitive and are more efficient and effective in performing successive tests. They can also be extended to testing multiple factors and have many of the same advantages as the more sophisticated statistical method of ANOVA without the complexity.

Control charts are easy to construct and intuitively friendly. They are powerful, yet simple methods to determine statistical significance. They can be used on proportional data like conversion or continuous data such as gross profit margin. Yet, control charts are extremely useful and will be seen as practical tools in many emerging areas.

References

Adams S & Orville B (1999). *Manufacturing the Future: A History of Western Electric*. Cambridge: Cambridge University Press, ISBN 0521651182.

Almquist E & Wyner G (2001). Boost Your Marketing ROI with Experimental Design, *Harvard Business Review on Marketing*, Harvard Business School Press.

Montgomery D (August, 2004). *Introduction to Statistical Quality Control – 5th Edition*, John Wiley & Sons

Nielson J (August, 2005). Putting A/B Testing in Its Place. (www.useit.com)

Wheeler D & Chambers D (June 1992). *Understanding Statistical Process Control.* SPC Press, Inc. 2nd Edition. ISBN 0945320132

Wheeler D J (2000) *Normality and the Process-Behaviour Chart*. ISBN 0945320566

Scott Burk is a Senior Scientist in Pricing Science at Zilliant.com