Evolving from the Use of P-Values to Multi-Armed Bandits Algorithms for Marketing Analytics

David Fogarty

National University, United States

Abstract

Marketing Analytics is an important application of analytics in modern firms. Being able to evaluate and improve the impact of marketing investments drives sales, acquisition. retention and customer loyalty. The use of hypothesis testing to evaluate the effectiveness of marketing campaigns is an important application of marketing analytics. However, marketing analysts either follow the social sciences gold standard of rejecting the null hypothesis if the probability is less than .05 or are unsure of which p-values to use to evaluate statistical significance. This research takes a critical look at the typical marketing investments and evaluates how the .05 level of significance is inappropriate for these types of investments. Finally, the research suggests a repeat testing approach via multi-armed bandit algorithms which are more suited to the types of investment marketing organizations are making and what is expected from their stakeholders.

Keywords: P-Values, Marketing, Investments, Machine Learning, Artificial Intelligence, Quantitative, Data Mining.

Introduction

Marketing analytics and its related tools are commonly used to unveil customer preferences and behavior patterns in order to provide insights for product development and service launches. Marketing analytics is defined by Wordstream as "the practice of measuring, managing and analyzing marketing performance to maximize its effectiveness and optimize return on investment (ROI). Understanding marketing analytics allows marketers to be more effective and efficient at their jobs and maximize the ROI of marketing investments. Other names for Marketing analytics include growth analytics, customer value management and customer relationship management. No matter what the name all have in common something to do with evaluating and influencing customer and client behavior all within acceptable ROI targets. This growth of marketing analytics is the trend away from mass advertising to certain broad segments of the population which while still effective for many of the big brands has proven to be an inefficient use of capital today especially with some of

the targeted channels available for a firm to get its message across. The growth of the IoT or Internet of Things is also fueling the growth of marketing analytics with data being captured from multiple devices. Proctor and Gamble for instance in very interested in the data from the power buttons when people order laundry detergents directly from their washing machines. While at the current time they don't envision having over 1000 different buttons required to cover every household good purchased there is bound to be further technical innovations around this and the data gathering, and prediction possibilities are proving to be very exciting. Sensors in the refrigerator can be used by grocery chains not only to replenish household food stocks but also can be used for marketing new products based on consumer habits. Grocery and food product executives imagine a day when the profile of a person can be used to automatically stock their refrigerator and cabinets with existing food products and even go one step further to proactively allow them to try new items. Think about this occurring all within a set budget for the consumer. This would be a real value add especially when most consumers are complaining about the rising cost of food. If they had a machine that they thought was "watching their back" this would represent a big plus. Different themes on this could be played out with selections based on healthy eating and also based on weight loss or types of diets. Perhaps the justification of the recent acquisition of Whole Foods by Amazon is based on related ideas and/or concepts. The use of Marketing Analytics by firms is even more important today given the concept of the "Modern Marketer" described by the American Marketing Association which boils down to 5 key traits which include being digital, customer centric, strategic versus tactical, always optimizing and have a focus on revenue-based outcomes.

Some of the key areas for the application of Marketing Analytics are listed below:

- Marketing Design of Experiments / Test and Learn / Champion Challengers / A|B Testing
- Pricing Analysis - Price Elasticity, Price Sensitivity, and Willingness to Pay analyses
- Product Analysis – Product Decision Trees – Product Cohort Analysis
- Distribution Analyses (i.e., store location decision model)
- Digital Marketing Attribution and Channel Optimization
- Indirect Attribution, Marketing Mix Modeling and RCQ (Reach, Cost and Quality)
- Customer, Broker/Producer and Client Segmentation
- B2B B2B2C and B2C Retention and Engagement Analytics and Modeling
- B2B B2B2C and B2C Acquisition Analytics and Modeling
- Satisfaction/Engagement/NPS/Moments that Matter/Moments of Truth
- Customer Lifetime Value/Prospect Lifetime Value/Client Lifetime Value
- B2B Lead Generation for new prospects and cross-selling

- Propensity Modeling
- Consumer and Commercial Choice Modeling
- Sales Forecasting
- Marketing Optimization Modeling (Channel, Response, Pricing, Conversion, Engagement, Next Best Offer or Next Best Action)
- Campaign Reporting and Testing
- Brand Equity / Brand Valuation / Brand Health Evaluation
- Aggregate/Individual-Static/Dynamic Advertising Response Models
- Marketing Funnel Analysis
- Social Media Analytics
- Conjoint Analysis for valuing attributes (feature, function and benefits) in products in services

The list above represents only a current snapshot of Marketing Analytics activities and it should be noted that activities will need to be added to the inventory as the function evolves especially in new areas such as the digital space when the collection of data and online interactions make the use of analytics critical. As we advance into the digital space the list of analytics activities under Marketing Analytics will continue to grow and therefore it is expected that this area is ripe for future research. However, for the purposes of this research a deep dive will be taken on one of these which is Marketing Design of Experiments / Test and Learn / Champion Challengers / A|B Testing. Firms which regularly use testing are better able to quantify the results of their investments and are more likely to attain greater levels of success. Fogarty (2013) describes testing as an often-overlooked method which enables firms to try new innovations while reducing the risk of making significant investments without realizing adequate returns. Focusing on experimentation in large Fortune 500 companies this researched explored the hypothesis that firms can sustain unprecedented growth by combining normal growth initiatives with a disciplined test and learn approach. The theory around testing is that companies who adopt the test and learn approach will be more equipped to take on additional risk and thereby increase their returns which can be measured sustained in cumulative annualized growth rates. The evidence based on the results of the study garnered support for the hypothesis that firms can use testing as a one of the potential tools to mitigate risk, accelerate growth and stay ahead of the competition over a significant time period. Please note that while this study will focus on Marketing Design of Experiments / Test and Learn / Champion Challengers / A|B Testing each of the marketing analytics activities in the above inventory is also relevant since each one will

require some form of testing to determine efficacy. All modern marketers must demonstrate the value of their marketing investments and hypothesis testing is a required element of this.

Marketing Design of Experiments / Test and Learn / Champion Challengers / A|B Testing

Marketing design of Experiments or DOE is all about setting up tests for marketing programs and marketing campaigns which can measure the impact of these activities before a full launch. These experiments can cover all areas of marketing including branding and advertising, direct marketing campaigns, customer communications, customer service etc. Setting up some of these experiments can be quite difficult especially on activities like national advertising and branding campaigns where not exposing participants to the treatment is virtually impossible. With these it's possible to set up quasi-experimental designs and/or develop before and after methods of evaluating the test when it's implemented. Most marketers are not skilled in designing and maintaining control groups and this can be a problem. Even more problematic is when the business environment is not conducive to creating control groups. One of these environments is in the healthcare field where creating a no treatment control group is unethical as this can induce illness or even death. There are also other instances like in a B2B where clients will require that all of their employees or customers receive the same treatment. While this may be good from a coverage perspective the firm will never really know how effective they are at the program and may find it difficult to convince the client to keep the program going when capital is scarce and competition for other investments is fierce. In my experience the same hold true in retail where analysts try to do a test in certain retail stores, but the retailer wants to launch in every store immediately with the hope of building sales volume.

Retailers are notorious for doing this despite the fact that in other areas of their businesses they are highly sophisticated in terms of analytics. Tesco was a pioneer in the analytical testing for retail area boasting discount coupon participation rates higher than 25% which at the time was virtually unheard of in the industry. The author personally worked on a similar and related case in the UK where UK subsidiary of Wal-Mart known as ASDA, we conducted an analysis of the customer private-label card activity and discovered that a great many customers were buying products at Boots which is equivalent to CVS or Walgreens in the United States. What was found via data analysis was that these ASDA customers were primarily

women and were buying some of their cosmetics at Boots instead of ASDA. So, what we did which was a first in Wal-Mart history was to offer incentive coupons for discounts on cosmetics to ASDA shoppers who were frequently shopping at Boots for what we surmised to be cosmetics. The willingness of Wal-Mart to try this in their remote operations was because of the retail business culture in the UK at the time being led by Tesco who as discussed earlier used to boast of unprecedented 25% plus coupon fulfillment rates due to targeting and analytics. The incentive coupons offered a discount at ASDA if the customer purchased something from the cosmetics department. Of course, a randomized control group was created in order to control for the treatment effect. The customers in the control did not receive the promotion in the mail. By setting up the test group which did receive the coupon and the control group which didn't we were able to ascertain whether or not the incentive worked in terms of being able to stimulate incremental sales. Some sales in the control will occur anyway. Therefore, the difference in sales between the test and the control is known as the incremental sales or the response "lift". If the sales lift generates enough incremental sales and profits to pay for the cost of the mailing and generates a decent ROI then the promotion is deemed to be a success. Of course, sometimes it takes quite awhile for a firm to determine the ultimate profitability of the campaign as it can have a knock-on affect to a customer's lifetime value. Other big areas for test and control are B2B and B2C marketing campaigns executed through direct mail, email or other channels. Direct mail is usually the easiest and is measured typically through direct mail control cells. Email can also be measured in the same way although the same guarantee of delivery and comfort that the recipient will at least consider reading it is not the same as first class print mail. However, there are many variables to deal with even above and beyond whether or not the mail gets delivered. For instance, in a recent mailing at a Fortune 500 company the US Postal Service sorting machines scuffed up a whole batch of color mailings with their postal sorting machines which are more geared toward envelope enclosed mailings versus large post cards. The alternatives to mitigate this problem included enclosing the postcard in an envelope, and several types of glossy and non-glossy acquiescent coatings. Of course, the scuffed mailers are still generating a decent response rate and ROI so with any alternative the firm must be careful not to bring down the response and increase the cost which if these are both moving in the wrong direction could impact the ROI dramatically. The test would have to be a sample run and mailing on a sample of people from virtually every state in the US since the researchers would have to have the mailers

sorted in the various regional mailing facilities. Of course, this is a difficult sample for a business to run as where would they be able to send mailers to in order to test whether certain sorting machines are causing the problem or not. Moreover, how to create small production runs of different creatives is another problem with this test since many of the print shops wanted a certain quantity in order to achieve the desired cost requirements via scale of efficiencies. As one can see a test like this can be a real challenge for smaller firms and many today are literally at the mercy of the post office especially since it's a government agency and not fully subjected to the economic laws of commercialism.

Email also has interesting properties for testing. Unlike direct mail one can have interim statistics with email by tracking KPIs such as click-through-rates, open-rates, unsubscribe rate, delivery rate, spam complaint rate, site traffic, delivery rate, time on site. Of course, these are just interim measures trying to get to the ultimate which is conversion rate or retention. The key challenge for any analyst is the forecast the conversion rates and retention rates from these interim KPIs. This is especially true when it takes quite awhile to be able to read the ultimate KPIs of a test like for example ones where the researcher wants to measure customer lifetime value or in healthcare where retention is not measured until after the annual enrollment event. This is an example where not only testing is critical where a probability sample is drawn but also modeling is needed in order to forecast what will happen in the future. Consumer credit companies do this very often when they are booking new accounts since they don't know what the true behavior or profitability of new applicants will be until after the fact.

A new marketing channel which is currently receiving a lot of attention in terms of testing is the digital channel. This channel includes paid search, programmatic ads and social media. One of the advantages of using the various opportunities available to get your message across in the digital channels is the ability to adapt the marketing message on a real-time basis. Another advantage of this channel is the ease and low cost of testing since one does not have the traditional expense of postage. However, among the challenges of testing in this new channel is the fact that while most people get their mail on a daily basis which bodes well for direct mail testing the same doesn't hold true for searching the web which can vary widely between different segments of the population. This in effect can create an unknown and difficult to measure response bias.

Evaluation of Marketing Tests

Now that an overview of testing in marketing analytics and some very specific examples have been provided, we will discuss how these tests are evaluated. Typically, these tests rely on the statistical hypothesis testing concept to evaluate whether or not the test group or groups are significant over the control population. These tests typically include Analysis of Variance, T-tests, Chi-Square Tests and Kolmorogov-Smirnoff Tests. Most of these statistical tests including the previous examples mentioned rely on a P value in order to reject the null hypothesis if the calculated probability is in the rejection region. The P value is a measure of nonagreement of the fit of a model aka the "null hypothesis" to a particular sample dataset. This process was suggested by Fisher (1936) and later refined by Neyman (1950). The mathematical notation for this process was well stated by Vander Pas, S., L. (2010):

"**Definition 1.1** (sample space) The sample space X is the set of all outcomes of an event that may potentially be observed. The set of all possible samples of length n is denoted X n."

"**Definition 1.2** (test statistic) A test statistic is a function T : X → R. (see Figures 1 and 2) We also need some notation: P(A|H0) will denote the probability of the event A, under the assumption that the null hypothesis H0 is true. Using this notation, we can define the p-value".

"**Definition 1.3** (p-value) Let T be some test statistic. After observing data x0, then p = P(T(X) ≥ T(x0)|H0). Figure 3: For a standard normal distribution with T(X) = |X|, the p-value after observing x0 = 1.96 is equal to the shaded area".

The statistic T is usually chosen such that large values of T cast doubt on the null hypothesis H0 thereby resulting in a rejection. This is known as setting the critical level of alpha. Informally, the p-value is the probability of the observed result or a more extreme result, assuming the null hypothesis is true. This is illustrated for the standard normal distribution in Figure 1. The rationale behind this test is Fisher's well-known disjunction: if a small p-value is found, then either a rare event has occurred or else the null hypothesis is false. Overall, this process enables to gather evidence for or against our alternate hypothesis which is the inverse of the null hypothesis (Vander Pas, 2010).

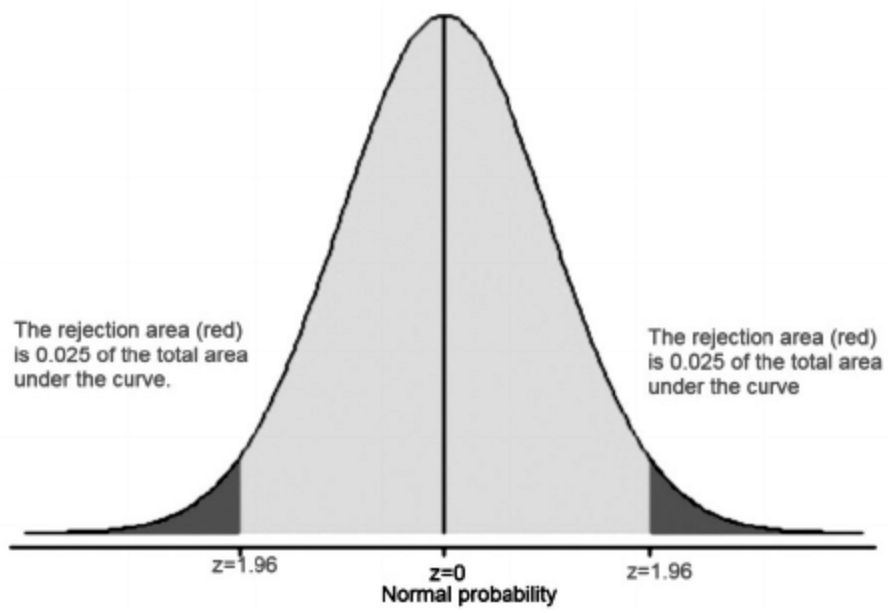
Figure 1: z-score for proportions

$$z = \frac{\left(\hat{p} - p_0\right)}{\sqrt{\frac{p_0\left(1 - p_0\right)}{n}}}$$

Figure 2: t-score for means

$$t = \frac{\left(\bar{x} - \mu_0\right)}{\left(S_x / \sqrt{n}\right)}$$

Figure 3: The standard normal distribution with alpha = .05 (two-tailed)



In the case of marketing this process lends to the evidence that our marketing treatment worked. Otherwise, we fail to reject the null hypothesis and conscript to further testing. The decision on setting the critical alpha level in marketing is usually driven by executives' understanding of statistics of which they were exposed to in college with an undergraduate course and/or a graduate level course depending on whether or not they received an MBA with a quantitative discipline.

Challenges to Traditional Marketing Test Evaluation

However, despite their popularity in business and in marketing there are some pretty heavy challenges with the p-value concept being used today. Burnham and Anderson (2014) point out that defending the use of P-values in the 21st century is flawed due to the many advanced in statistics and data collection which have taken place. One of the big challenges of the current P value system was defined by Gelman (2013) who pointed out the P values across different tests cannot be easily compared due to the fact that the difference between a highly significant P value and a nonsignificant P value is itself not statistically significant. Moreover, there are also challenges with using the .05 level of significance as the scientific gold standard (McShane, Gal, Gelman, Robert and Tackett, 2019). Many of the current studies challenging the use of p-values are worried about statistical significance and committing Type-I errors. However, in marketing tests the opposite is true in that marketers are typically are more

concerned with Type-II errors where one could be missing critical opportunities at relatively low business risk if a false negative were to occur.

Suggested Evolution of Marketing Test Evaluation

One of the recent challenges the author ran across using the .05 alpha level is after a series of expensive testing the null hypothesis on a critical healthcare test failed to be rejected at the .05 level of significance using an independent t-test of proportions. The problem is the level of significance was actually < .07. Theoretically, the P value is a continuous measure of statistical evidence. However, historical hypothesis testing methodology tends to discretize the measure would require us to preset the alpha level and fail to reject the null hypothesis if the p-value was greater than .05. Our conclusion to the business given all the effort put into creating and executing the test along with the hypothesis being part of the core value proposition of the business was that the danger of committing a Type-II error was of far greater risk than a Type-1 error which the .05 level of significance was designed to protect.

Some of the purist statisticians would never allow such a decision as it fails to follow the rules of statistical hypothesis testing. However, in statistical quality control we make decisions on trends in a run chart all the time without the line crossing into the upper or lower control limits. Crossing into the upper and lower control limits in control charting terms is the statistical equivalent of rejecting the null hypothesis in a hypothesis test. The use of runs in quality control goes back to the 1940s and 1950's where several authors suggested a concept of a sequence of points falling within/outside the control limits which could be used to identify trends. This was later systematized by Western Electric (1956) which outlined in cookbook fashion how to set up and evaluate a control chart. This publication reinforced the worldwide use of runs in the field of Statistical Process Control (SPC).

The main objective for suggesting this set of rules was to increase the sensitivity of the Shewhart charts and improve their potential to detect non-random patterns. Some of the first statistical validations of runs rules were provided by Page (1955) and Mosteller (1941). More recently, Koutras, Bersimis and Marakelakis (2007) evaluated the statistical basis for runs rules in Shewhart Control Charts and found that despite criticism, the Shewhart charts have solid theoretical foundations and are among the easiest charts to implement.

The real question becomes whether given the absence of data we can treat marketing campaigns as a continuous process where we can measure and improve the effect over time. Multi-Armed bandit algorithms (Slivkins, 2019; Vermorel & Mohri, 2005; Mannor and Tsitsiklis, 2003; Auer, Cesa-Bianchi & Fischer, 2002; Auer, Cesa-Bianchi, Freund, Y & Schapire, R., E., 2002) do just this but do not explicitly discuss it in the same context as a control chart.

Multi-Armed Bandit Models

With multi-armed bandit models there is a Bayesian apriori probability which allocates subjects to the marketing treatment in advance and then further subjects are allocated to the same treatments in different proportions as the posterior probabilities are calculated (Kaufmann, Cappe Garivier, 2016; Villar, Bowden, Wason, 2015). Think of it as if one were walking into a casino with a bag of coins or tokens and there were six one armed bandits on the table. The casino manager informs you that each of the slot machines are biased. You want to spend all your quarters but maximize the return. So, you start by placing a single coin in each slot machine until you get a win. The winning slot machine will get two coins for every one in the losing slot machine until another wins. Then the second winner will get more coins and so on. The principle is that you heavy-up in the winning machines but do not bet the farm on them in case it's just a random run. The advantages of using multi-armed bandit models is that instead of waiting until one of the treatments (in the above case the on-armed bandit) to be classified as better from a traditional hypothesis testing methodology you can capitalize on some of the gains earlier in the cycle via optimization and being greedy about those treatments receiving earlier wins. While the casino analogy is important to the understanding of multi-armed bandits given their namesake it just as important to point out that there are many marketing applications such as the use of dynamic creative content delivered through a digital channel in order to convey a message to a customer or product.

More advanced multi-armed bandits enable researchers to model the individual characteristics if the people who respond to the treatment and optimize the treatment features around these characteristics creating a continuous improvement. They are also being referred to as contextual one-armed bandits (Surmenok, 2017) and often utilize artificial intelligence and

machine learning to actually model the type of treatment that individuals would prefer using all of the characteristics available on those specific individuals. These characteristics could include demographics, psychographic, financial, automotive, homeownership, shopping, web browsing etc. Another important function of the bandits is being able to truly codify the creative characteristics that really matter. This codification can serve to reduce the reliance on pure creative resources to design marketing content and place a greater reliance on quantitative analysis. Moreover, it serves to enhance or augment the creative resources since their work would now be less of a guessing game and more informed via direct response information on consumer preferences rather than those obtained via primary research.

The problem with insights derived via primary research is that what people say they are going to do is often at odds with what they will actually do. Take consumer credit research for example. In consumer credit research if you ask people about how much interest they would pay they for the most part will reply as little as possible. However, individual borrowing behavior demonstrates that when they need to borrow, they will actually pay more than they are willing to say in an interview. In healthcare primary research will demonstrate that everyone wants to engage in healthy behavior and pay no deductible for healthcare. However, in practice many people do not do what they say and will choose a healthcare plan that fits their individual budget. Therefore, actual in-market testing with control groups which can actually track human behavior are most appropriate. Fogarty (2013) talks about this in his study of the impact of testing and growth at Fortune 500 firms in that testing should be a fourth type of research to accompany qualitative and quantitative primary research along with exploratory data mining. Overall, in his study it was discovered that firms which engage in enhanced testing activities overall grew much faster than those that did not. Of course, correlation does not always imply causation but firms which are more successful very often use testing as one of their tools to learn about how to get better. Its important with testing to not only succeed but also to have planned failures. When the author worked at GE they used Six Sigma to evaluate and measure everything including the number of tests. The motto from Peter Drucker that was followed closely using Lean Six Sigma was “What gets measured gets managed”.

Just measuring the number of tests is not useful since the impact based on the number of tests is based on the size and complexity of the organization. Other measures include the number

of tests which turn into a positive outcome. However, all testing should have a fair number of failures in addition to successes. If you just measure and award the positive outcomes, then managers will be afraid of reporting failures and as such will not conduct the proper learning necessary to grow the business or protect it from risks. Therefore, the proper measures could be the number of tests along with the ratio of success to loss in the tests. Managers may want to keep their loss ratios slightly higher to make sure they are taking enough risks in their tests. So far, no theories or research on this loss ratio have been made and it's a suggested topic of further research.

Many companies like Capital One report the number of tests they conduct on an annual basis and use this to demonstrate how well they are doing in terms of using data analysis to drive the business. The marketplace reacts to this by boosting their stock and also getting the competition to follow in kind. When the author worked in analytics at GE Capital in the early Aughts this was exactly the case as the standard was set by Capital one on the number of tests they were conducting which was being communicated by bank officials at industry conferences and in business publications. The leaders of GE Capital would ask them to count the number of tests run in our businesses and then set goals to achieve Capital One level standards. At first, they followed suit but of course realized they were way behind as Capital One was running some 10,000 tests per year. Then it was realized from some of the new people they hired from Capital One that the number of tests which Capital One reported included some of the subtests and levels of the tests in the campaigns which the business is running. When this criteria was applied to the measurement it became easy for GE Capital to run an equivalent number of tests over the same time period. Since GE was a Six Sigma driven company and the author was a certified Master Black Belt in Six Sigma at the time there was a very heightened interest in creating measurements related to testing. Moreover, there was the distinct appetite to go above and beyond what Capital One had achieved so that they GE Capital could become the new leader and not a follower trying to catch-up with the competition. Therefore, it was decided to not only measure the number of tests being run but some of the suggested metrics above related to testing success outcome ratios. These enhanced metrics represent the quantification of the learnings being generated as a direct result of testing.

It should be noted that sometimes its not good to follow the competition in metrics. An example if this is Wells Fargo. GE Capital Global Consumer Finance got the idea in the early Aughts to follow the leader in cross-selling Wells Fargo who measured their success in the number of products per customer. GE Capital Global Consumer Finance was interested in replicating the success of Wells Fargo and set a similar goal for its own consumer finance business. Unfortunately, we were to learn only later that this metric was responsible for the sales team at Wells Fargo to conduct unethical activities and this subsequently led to a negative brand image and eventual apology from the senior leadership to customers. The good news is that at GE we achieved our products per customer goal via providing our lists to affinity insurance companies like Cigna and AIG. Therefore, there was no miss-selling as these outsource companies followed strict rules in terms of regulations for outbound telemarketing sales and following compliance rules for insurance selling disclosures. This outsourcing of insurance and sundry products along with the use of advanced analytics to drive cross-selling by identifying the right lists and booking customers who will eventually have a high number of products or those we currently had a low share of wallet allowed us to achieve their goal ethically.

The benefit of this thought process of treating marketing campaigns as a continuous testing process is due to the fact that in the author's previous 20 years of experience very few occasions have occurred where marketing campaigns were successful at the outset. It usually takes about 6-12 months of continuous testing including the development of propensity models before one can show a positive result with a campaign. This is especially true when one is trying to change a new behavior for the first time without any established practices. For example, it may be fairly well-known how to convince a person to respond to a pre-approved credit offer. However, if a change in the industry results in a new observed economic behavior in consumers such as rate surfing then executing and testing new programs to mitigate rate surfing could take months or even years to perfect. This perfection will include testing creative, offers, media, exclusion criteria, modeling etc. All of these would have to be tested, adapted and modified. Moreover, as long as the product is competitive in the marketplace and/or has sufficient brand health and awareness then the direct marketing campaigns can be adapted to make a profitable direct marketing business. The trick is to be able to survive at a loss while the learning is taking place and convince senior management that there will be a winning business as soon as the

testing cycle is mature. The author generally finds that if you can show actuaries or finance managers that there is a testing plan leading to anticipated successful outcomes then they will be more willing to continue funding the program. However, one must meet these outcome goals within the agreed deadlines of the investor will quickly lost patience and/or begin to be worried that their investment will not be recouped. Sometimes a test which is expected to work will not work, and it may not work in the traditional way where the effect is not statistically significant but may instead work in the opposite way where an effect is statistically significant but performing in the opposite way as the hypothesis intended. Another way to put it would be a creative test which uses imagery and phrases to attract people to buy a certain product. The message is to be served in a digital advertisement. The imagery and phrases are to be developed based upon a segmentation analysis which divides consumers based on their preferences for the product. The segmentation could have age and lifestyle variables as components, and these could be represented in the creative through imagery which looks like the individuals from an age standpoint or shows people engaging in the relevant lifestyle. Fitness could be one example of a lifestyle. Outdoor activities like kayaking or skiing could represent another lifestyle. The hypothesis would be that people would be attracted to the lifestyles which fit their own profiles and thus would be more responsive to the product or service intended to be sold via the message. Sometimes these messages will need to be tweaked in certain ways in order to be attractive to customers or prospects, so a few insignificant tests are to be expected.

One example of changes which can have a significant impact on many direct marketing campaigns is subject line testing to get people to actually open emails. Subject line phrases are very sensitive to changes and a few words can make or break an entire campaign response rate. We all get many emails each day usually too many to open without some type of filtering. Subject lines are a great way to apply this filtering. If the phrase doesn't attract us the mail doesn't get opened. The reason is that subject lines are a big factor in enticing people to open their emails which is needed for them to respond or react to one's email offer. Researchers don't want stakeholders getting discouraged in the whole direct marketing program from a few bad subject line tests. Researchers can get to the proper subject line after only a few tests or testing multiple subject in parallel. This challenge of learning as you go is to be expected when testing so what we are proposing is a new method of evaluating tests using elements from control

charting where we can measure the effectiveness of test along a continuum rather than at a fixed point in time. Then stakeholders can become much more comfortable that the test they are sponsoring is performing consistently over time. They can also see how their investment is progressing and if after repeated testing it looks like the hypothesis is not being supported by the data, they can choose to pull the plug on the test. The probability of the stakeholders doing this earlier than when a clear decision can be made is lower due to the greater transparency of the testing method used in this procedure.

Conclusion

In this study an alternative to the traditional hypothesis testing approach to marketing test evaluation was suggested. In today's ever-changing environment with increasingly expanding digital marketing and the ability to change targeted communications in real-time there is a need to rethink how marketing departments are measuring and tuning the sensitivity of their tests to ensure that no opportunities are missed. The techniques evaluated in this research are a step in this direction.

This study represents an attempt to break the cycle of decades or practice on using traditional hypothesis testing to evaluate the efficacy or marketing activities ranging everywhere from advertising and branding campaigns to retention and acquisition campaigns. It flies in the face of everything we learned in school about statistics but nonetheless makes perfect sense from a practical perspective. Traditional hypothesis testing is a model and as the famous statistician George Box said, "all models are wrong, but some are useful".

So, in final summary the case was presented that the classic hypothesis test in which marketers reject the null hypothesis at the .05 level of significance is not very useful given the fact we as marketers are more worried about committing a Type-II error and even still we are more worried about multiple tests and the performance over time. Therefore, a recommendation was made to drop p-values altogether and instead pursue all marketing positive marketing outcomes with greater levels of investment (in an optimal manner using a Bandit Algorithm) until we can conclude definitely that there is a superior treatment. Research shows that this is the most effective way to allocate resources to ongoing marketing campaigns where testing is continually being administered.

This will allow us to arrive at the most optimal solution versus depending on an inaccurate single p-value.

References

Auer, P., Cesa-Bianchi, N., Fischer, P. (2002) Finite Time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2/3), pp. 235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1), pp. 48–77

Burnham K., P., Anderson, D. R. (2014). P values are only an index to evidence: 20th- vs. 21st-century statistical science. *Ecology*, 95(3), pp. 627–630.

Fisher, R., A. (1936). *The Design of Experiments*. Oliver and Boyd. London, UK

Fogarty, D., J. (2013) Global 500 Test and Learn: The Relationship between Business Experimentation and Explosive Growth in Large Global Firms. *International Journal of Business and Management Tomorrow*, vol. 3, no. 5, pp. 16-25.

Fogarty, D., J., Harrison, P., Jing, L., Yip, S. (2014) Beyond Sales and Awareness: Using Marketing Analytics for Improved Health Engagement and Outcomes. *Applied Marketing Analytics*, vol. 1, no. 1, pp. 13-20.

Driscoll, D., L., Appiah-Yeboah, A., Salib, P., Rupert, D., J. (2007) Merging Qualitative and Quantitative Data in Mixed Methods Research: How To and Why Not? *Ecological and Environmental Anthropology*, vol. 3, no. 1, pp. 19-28.

Gelman, A (2013) Misunderstanding the p-Value. *Statistical Modeling, Causal Inference, and Social Science*, Retrieved from: https://statmodeling.stat.columbia.edu/2013/03/12/misunderstanding-the-p-value/

Hawton, K., Kingsbury, S., Steinhardt, K., James, A. & Fagg, J. (1999). Op cite.

Johnson, R., B., Onwuegbuzie, A.,J., Turner, L.,A. (2011) Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, vol. 1, no. 2, pp. 112-133.

Kaufmann, E., Cappe, O., Garivier, A. (2016). On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1-42.

Koutras, M., V., Bersimis, S., Maravelakis, P., E. 2007) Statistical Process Control Using Shewhart Control Charts with Supplementary Runs Rules. *Methodology and Computing in Applied Probability*, vol. 9, no. 2, pp. 207-214.

Mannor, S., Tsitsiklis, J.N. (2003). The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. In: *Sixteenth Annual Conference on Computational Learning Theory.*

McShane, B., Gal, D., Gelman, A., Robert, C., Tackett, J. (2019) Abandon Statistical Significance. *The American Statistician*, Vol. 17, no 4, pp. 23-34.

Mosteller, F. (1941). Note on application of runs to quality control charts. *Annals of Mathematical Statistics*, vol. 12, no, 1, pp. 228–231.

Neyman, J. (1950) *First Course in Probability and Statistics*. Henry Holt, New York.

Owens, D., Horrocks, J. & House, A. (2002). Fatal and non-fatal repetition of self-harm. *British Journal of Psychiatry, 181,* 193-199.
Page, E., S. (1955). Control charts with warning lines, *Biometrics*, vol. 42, no.2, pp. 243–257.

Slivkins, A. (2019), "Introduction to Multi-Armed Bandits", *Foundations and Trends® in Machine Learning*: Vol. 12: No. 1-2, pp 1-286. http://dx.doi.org/10.1561/2200000068

Surmenok, P. (2017). Contexual Bandits and Reinforcement Learning. *Towards Data Science*. Retrieved from: https://towardsdatascience.com/contextual-bandits-and-reinforcement-learning-6bdfeaece72a

Tourangeau, R., Rips. L., J., Rasinski, K. (2000). *The Psychology of Survey Response*
Cambridge: Cambridge University Press, Cambridge UK.

Vander Pas, S., L. (2010). *Much ado about the p-value: Fisherian hypothesis testing versus an alternative test, with an application to highly-cited clinical research*. Unpublished Thesis. Mathematisch Instituut, Universiteit Leiden.

Vermorel J., Mohri M. (2005). *Multi-armed Bandit Algorithms and Empirical Evaluation*. In: Gama J., Camacho R., Brazdil P.B., Jorge A.M., Torgo L. (eds) Machine Learning: ECML 2005. ECML 2005. Lecture Notes in Computer Science, vol 3720. Springer, Berlin, Heidelberg.

Villar, S., S., Bowdon, J., Wason, J. (2015). Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical Science*, vol. 30, no. 2, pp. 199-215.

Western Electric Corporation, (1956). Western Electric, Statistical Quality Control Handbook, Indianapolis, Ind.