Controlled Experiments for Decision-Making in e-Commerce Search

Anjan Goswami, Wei Han, Zhenrui Wang, Angela Jiang
WalmartLabs
860 W California Ave, Sunnyvale, CA 94085
{agoswami, whan, zwang, ajiang0}@walmartlabs.com

***Abstract*—With the explosion of big data, companies both small and large are increasingly motivated to make data-driven decisions. For web-based companies in particular online controlled experiments or A/B tests have become essential scientific tools for decision-making. Large scale organizations like Google, Amazon, eBay, Facebook, LinkedIn, Yahoo, and Microsoft have built mature systems and support for controlled experiments and have helped to popularize the methodology of A/B testing for guiding product development. In e-Commerce, A/B tests are used extensively to understand how customers respond to new features and to use statistics at scale to make decisions rather than relying on the highest paid person's opinion. However, unlike other web-based businesses, e-Commerce exhibits particular qualities and behaviors where generalized controlled experimentation guidelines are not nuanced enough to provide meaningful insight into decision-making. Specific functions of an e-Commerce website present further complexity in testing, especially in essential and highly visible functions like search.**

Per week, more than 260 million customers visit Walmart's retail units and e-Commerce websites. A single decision can result in not only large financial repercussions but also cascading effects across such an expansive customer base that traverses between digital and physical purchase points. For the retail sector at large, e-Commerce is a new business as most operations are still primarily physical. The transformation to web-focused purchase points is burgeoning but the application of data-driven decision-making remains difficult. Generalized guidelines from technology companies have not been able to fully serve the problems specific to e-Commerce.

In this paper, we discuss our experiences in running controlled experiments at WalmartLabs Search and specific guidelines in bias, test design, measurement, analysis, financial constraint and decision-making for e-Commerce. We share examples and key lessons from our work in each of these focal areas where we believe the reader will benefit from interpreting real results. Our work provides the growing e-Commerce analytics community with guidelines for running experiments within their own retailers and highlights sector specific challenges in the application of controlled experiments.

***Index Terms*—Online experiments, controlled experiments, A/B testing, search, e-commerce**

I. Introduction

Controlled experiments or A/B tests in websites have become essential scientific instruments for making decisions in organizations including Google, Microsoft, Facebook, eBay, Yahoo, Amazon, LinkedIn, and Walmart. The application of controlled experiments illuminates the explicit choice of granting merit to the customer's opinion over that of the highest person paid and the shifting corporate governance of today's companies. The mathematical theory of A/B testing originated largely from the work of R. A. Fisher, J. Neyman, and E. S. Pearson [1] [2] with the current commercial use of A/B testing utilizing a simple variant of this theory. Even so, the practical application of A/B testing has proven to have several complex challenges - many of which have been noted by Kohavi and his Bing research group in their observations and experiences with A/B testing at various large scale websites [3] [4] [5].

While the literature on the generalized nature of A/B testing and its application to websites have given experimenters and researchers greater insight into running controlled experiments, it highlights the fundamental maxim about application: the devil is in the details. In our experience applying controlled experiments to e-Commerce sites including Amazon, eBay, Elance-oDesk, and most recently WalmartLabs, we have found important distinctions and unique complexities in running controlled experiments in e-Commerce and in its most core function: product search. For an e-Commerce business, search is the fundamental method by which customers find and discover products and as a result, it is a key business lever on financial performance. A change in search can translate directly to a change in an e-Commerce company's financial health - for better or for worse.

For the retail sector at large, e-Commerce is still a new business as most operations originated from physical purchase points and most companies in the space continue to service their physical operations alongside their growing digital presence. The transformation to web-focused purchase points is burgeoning as new technologies make it easier for these companies to become omni-channel in their pursuit of the customer. The expansion of big data, the success of large technology companies in generating value from that data, and the growing demand for faster improvements has led many retailers into the digital foray.

With big data, e-Commerce sites face the imposing challenge of evaluation at scale. Evaluation in e-Commerce is two-fold: (i) we want to evaluate the success of a given feature in its functionality; and (ii) we want to evaluate the purchase intent that results from the feature. At small amounts of data, e-Commerce sites can employ manual methods to seek evaluation for these two criteria. However, with big data, manual methods no longer suffice to be able to provide meaningful and accurate feedback regarding the e-Commerce site and other scientific tools such as controlled experiments must be

employed. Big data further imposes additional difficulties in evaluation where rare probability events begin to surface more readily and alter the perceived notion of the current state of the e-Commerce business. As a result, the application of data-driven decision-making remains difficult. The use of controlled experiments as derived from generalized guidelines have not been able to fully serve the problems specific to e-Commerce.

In this paper, we discuss our experiences in running controlled experiments at WalmartLabs Search and specific guidelines in bias, test design, measurement, analysis, financial constraint and decision-making for e-Commerce. We share examples and key lessons from our work in each of these focal areas where we believe the reader will benefit from interpreting real results. Product search on e-Commerce sites is an interactive information retrieval system. When discussing search features, we refer to any change in the human-computer interface, ranking, recall, navigation, and assistance which impact the user's search strategies or ability to service an information need. In the context of this paper, search relevance features are specifically changes in ranking and recall. These search features operate on an explicit transaction relationship that has a large potential impact on retail financial performance. This dynamic warrants a rigorous decision-making process that incorporates well understood implications from A/B tests. The need for advanced A/B test analytics is absolutely crucial, perhaps even more so than for other types of companies. Our work provides the following contributions:

- We explain aspects of A/B testing for e-Commerce sites that have important distinctions
- We provide the growing e-Commerce analytics community guidelines for running experiments within their own retailers through a collection of our experiences and learnings from conducting A/B tests on various e-Commerce sites
- We highlight sector specific challenges for the application of controlled experiments for the research community

The structure of this paper is as follows. We start first with a discussion on bias as it relates specifically to e-Commerce and to search. In the following sections, we elaborate on establishing metrics, defining hypothesis tests, understanding the population, the impact of financial constraints, and analyzing the results, respectively. Finally, we conclude with a series of open questions for the research and analytics community that offer new challenges.

II. Know Your Bias

The definition of bias varies in different contexts. In machine learning, bias is defined as an error from erroneous assumptions in the learning process. In statistics, bias is an error in which a measurement is consistently different from its expected value. In A/B testing, we utilize the results from controlled experiments to make a launch decision on a feature. The culmination of successes and failures of features performing on the whole site ultimately lead to the future financial success or failure of the e-Commerce company. In this context, anything which results in a misestimation of this decision is a bias. In e-Commerce search, we identify several forms of bias:

- Visit level bias
- Query level bias
- Item level bias

Avoiding or reducing any of these biases is not a trivial task. In statistical experiment design, a randomized or full factorial experiment [6] can be used to reduce bias. However, in application, neither fully randomized experiments nor full factorial experiments are possible. Generally, multiple A/B tests are run simultaneously on the site and traffic is primarily split by visit due to constraints in engineering implementation. In e-Commerce search at WalmartLabs, we see over 45 million visits per month to our core Walmart.com site. An A/B test for a search feature must be run for a sufficiently long period of time and receive enough traffic to ensure the power of the test [7]. Further, the practical nature of the business favors rapid product innovation. Speed to market and a rapid feedback loop are crucial elements in the success of an e-Commerce business. These limitations force us to reduce bias through domain knowledge, randomization algorithms, analysis, and test design.

A. Visit Level Bias

Visit level bias is a type of error that originates from assigning visits to controlled experiments. Most experimentation platforms split traffic by some type of session or user identifier. Bias can result from using different types of identifiers in building metrics and defining visits. Most e-Commerce sites will employ a user identifier and a session identifier for retaining information about both repeating and new customers. Established users are often far along experience curve; they retain a familiarity with the functionality of the site and are likely well-oriented with various search strategies to find products. Subtle changes in ranking for example may prove ineffective to these users. New users, on the other hand, will lack these capabilities and may exhibit exaggerated responses to subtle changes. Choosing an identification method that fails to incorporate this phenomenon can result in bias.

At WalmartLabs, we attempt to eliminate this type of bias by building our metrics using both a session identifier and a user identifier. Even careful selection of visit identification is not always enough to reduce bias. Take for example this particular case from our own work: we launched A/B tests for several search features at the same time. Surprisingly, all tests showed reduction in a crucial business metric: Revenue Per Visit (RPV). All tests where subsequently turned off to prevent further financial loss to the site. After several rounds of analysis, we realized that another ongoing test was causing these simultaneous drops. During the time of testing, Walmart was transitioning to a new site design and a controlled test had been deployed to slowly ramp the transition. If one user was assigned to the new site, then the system remembered the user and consistently redirected the user to the new site. Therefore, established users were more likely to be assigned to the new site and new users to the old site. The differences between

users was noted remarkably in their financial contribution: established users generally provided higher RPV than new users. The resulting bias led us to the wrong conclusion: that our features were negatively impacting RPV.

Randomization algorithms and assignment methods are crucial to reducing bias. Randomization algorithms can follow various techniques including (i) pseudorandom with caching and (ii) hash and partition as long as they satisfy the four major properties of proper randomization to eliminate bias as outlined by Kohavi [8]. Assignment methods may also vary and each present unique challenges of their own [8].

B. Query Level Bias

Query level bias is a type of error that originates from differences in queries issued by users. Query level bias is crucial to understanding experimentation with search relevance features. Search relevance features are designed at the query level and do not impact all queries equally. Furthermore, typical experimentation platforms will operate on a visit level and we can generally expect users on e-Commerce sites to make multiple queries for products within a single session.

A notable example from our own work illustrates the importance of eliminating this bias: we developed a search relevance feature using query level data and ran a controlled experiment. The results of the A/B test came back negative, labeling the feature as a failure. We then performed query level analysis, conducting bootstrap samples and computing the mean difference in the metric of interest. The results proved positive. Because there was a difference in queries between our control and variation, using only the original visit level results would have resulted in the decision to not launch a value-adding feature. However, using query level analysis and understanding the query level bias that is likely to be present in e-Commerce search testing, we were able to make the appropriate decision.

Queries in e-Commerce search platforms may vary by traffic, product category, and search strategy. Typically, we will segment queries by traffic to identify head, torso, and tail cut-offs and conduct query level analysis within each segment. In our experience, we found that performance of search relevance features can vary significantly based on traffic segmentation. In another one of our features, we saw marked improvement in conversion rate for only head and torso queries but a notable decline in performance for tail queries. As a result, we launched the feature for only head and torso queries. Behavior may also vary by product category; we have seen relevance features show improvements in distinct categories of products while simultaneously showing regressions in other product categories. Lastly, users may also vary the strategies that they choose to employ to obtain their information need. Some users may try including color, brand, size, or gender attributes to formulate complex query strings while other users may start with a simplistic single token query and then narrow or refine their search through search assistance features, left-hand navigation filtering, or item recommendations.

C. Item Level Bias

Item level bias is a type of error that originates from differences in assortment and can include demand, pricing, and item data quality. While the general goal of assortment planning is to obtain a supply of products that meet customer demand, there can be significant differences in assortment across various dimensions. Some assortment may be more competitively priced, have more demand, and have higher data quality than others and vice versa. These differences may roll up to the query level during feature development and at the visit level during experiment runs.

At Walmart, we see a variety of items from both internal purchases and third-party marketplace vendors. These two item types exhibit distinct differences in demand and pricing and have different requirements resulting in differences in data quality. In certain product categories, these differences can be drastic - so much so that we have been able to develop features that capitalize on item differences by product category.

Performing analysis on common queries between different variants and understanding the performance of different item types on the e-Commerce site in question are important to reducing this bias. Domain knowledge in both assortment performance and search relevance is important; insight and analysis derived from questioning ranking or recall differences by item type can become inputs for consideration in metric or hypothesis testing design.

III. Know Your Metrics

Defining metrics for controlled experiments in e-Commerce search should include both consideration for the e-Commerce business as well as the nature of search. The objective of e-Commerce search is tied to financial objectives for the overall e-Commerce platform and as thus, the following are conventional business metrics which are tracked for all controlled experiments:

- Product View Rate (PVR): the percentage of visits that have clicked items
- Add to Cart Rate (ATC): the percentage of visits that have added items to cart
- Conversion Rate (CR): the percentage of visits that have converted
- Average Order Size (AOS): the total revenue over the number of orders
- Revenue Per Visit (RPV): the total revenue over the number of visits

These metrics identify the conversion funnel for users throughout the e-Commerce site, allowing us to identify major cut-off points in the user flow and the overall relation to financial success. The aggregation method for metrics varies depending on the stage of the controlled experiment. During the actual execution of the A/B test, metric collection and aggregation is by visits. Once the A/B test is complete, further analysis is done at more granular levels of aggregation including query level and item level. Consistent sessionization is necessary to ensure that metrics are translatable between various aggregation methods.

While these standard metrics are generally recorded for all tests, it is important to understand the tradeoff between statistics and to identify which metrics are primary and which are diagnostic for a given experiment. In one search relevance feature that we developed, we utilized historical order information to promote items with higher orders. The feature resulted in a significant reduction in AOS but a higher CR. Items for which the feature promoted tended to be lower priced, encouraging users to convert more frequently. A decision here based on an individualistic interpretation of each metric can prove to be difficult; instead, a holistic understanding is more prudent.

Selection of the appropriate primary metric or overall evaluation criterion (OEC) is fundamental to making appropriate decisions regarding features and should capture the overall business goals [9]. In selecting the OEC, experimenters should conduct thorough analysis of each metric including (i) descriptive statistics: mean, median, variance and quantiles; (ii) non-normality / weak normal / skewness; (iii) signal to noise ratio; (iv) time series behavior of each metric and (v) correlation between metrics. As a final step, we apply the general litmus test: is it possible to do something simple and wrong that will meet the OEC but not the real business goal? [10]. Without these, determination of the OEC can prove to be detrimental in decision-making. Common mistakes include:

- Creating conflicting metric pairs like AOS and CR which make it difficult to make a fundamental decision on launch as detailed in our earlier example.
- Optimizing for a metric that does not meet the business goal which can destroy long-term business value. In a frequent pitfall, e-Commerce businesses seek to optimize for PVR or the click-through rate for items as a carry-over from other web businesses. By selecting PVR as a primary metric, e-Commerce businesses fail to meet the real business goal. A simple application of the litmus test would prevent this mistake. In Figure 1, we show the relationship between query level revenue and query level PVR which is weak. Optimizing for PVR would fail to meet the goal of generating revenue.
- Creating elaborate sets of metrics in hopes to move a single metric which increases family-wise type I error.

Beyond OEC determination, analysis of metrics should be done to determine the appropriate sample size for the desired power. A typical test will have a confidence level of 95% and power of 90% [8]. Removal of outliers and identification of bots should also be done using descriptive statistics to identify thresholds for unlikely behavior [11]. Tests for normality should be used to determine the appropriate hypothesis test when designing experiments. For example, RPV is a highly skewed metric with more than 95% zeros; a standard t-test on a small sample size would be misleading. Time series analysis should be used to identify any seasonality or notable changes in behavior over time. In e-Commerce, weekends show different behavior than weekdays and certain product categories, like clothing, can be exposed to extreme

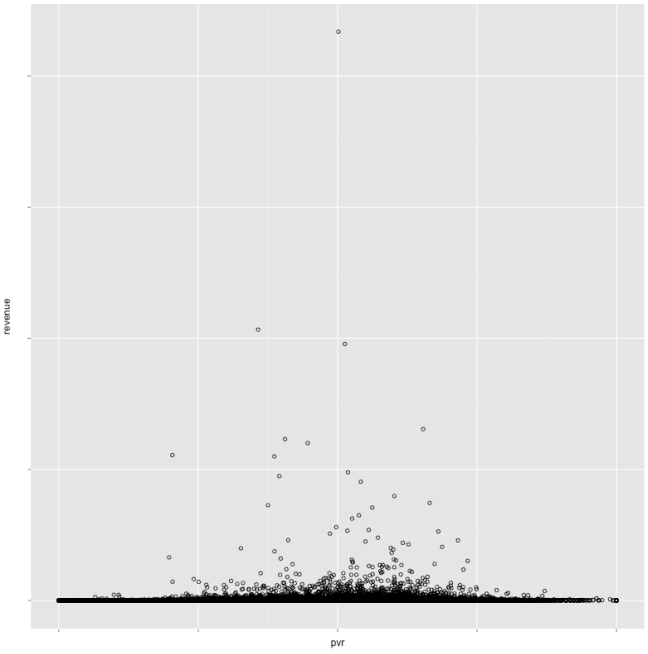

Fig. 1: Scatterplot of query revenue vs. query PVR

seasonality. Depending on the identified behavior, controlled experiments should be run over an appropriately long period of time and cover time frames for which differences can be appropriately captured. Depending on behavior, transformations or decomposition of metrics may also be done. For metrics that exhibit a long tail distribution like AOS, the logarithm of AOS may prove to be more robust. Complex metrics like RPV may be better understood as the product of CR and AOS. Features can then be designed to target improvement in one metric while keeping the other metric consistent.

IV. Know Your Hypothesis Tests

A/B testing relies on hypothesis test to examine the causal relationship between a metric of interest and the potential impacting factor that splits data into control and variation groups. Hypothesis tests draw conclusion on the validity of two contradicting hypotheses (i.e. null hypothesis H_0 and alternative hypothesis H_a) with consideration for randomness in data. Prior to launching an A/B test, it is crucial to quantify the effect on the metric of interest and to determine if your hypothesis test should be parametric or non-parametric. Most e-Commerce businesses choose to use third party out-of-the-box A/B testing solutions that will typically default to either a one-sample or two-sample t-test. Experimenters operating within e-Commerce sites should verify that the necessary assumptions are met when working with such third party solutions: (i) independent observations and (ii) normality of the distribution for sample means for the metrics in consideration within each sample or (iii) approximately normal distribution where the central limit theorem applies.

The misuse of t-tests on time series data is a common pitfall and a violation of these assumptions. A typical scenario

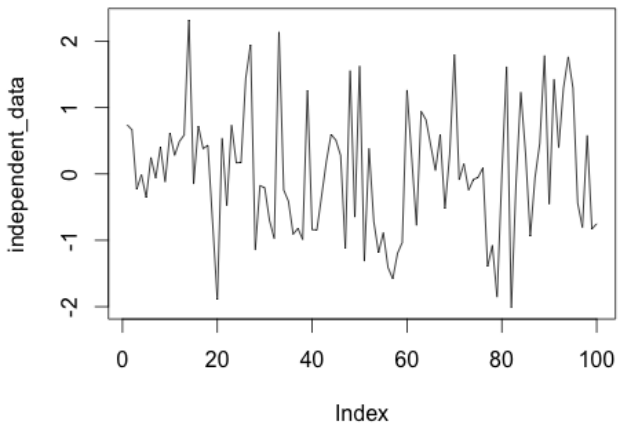

Fig. 2: Raw data of independent observations

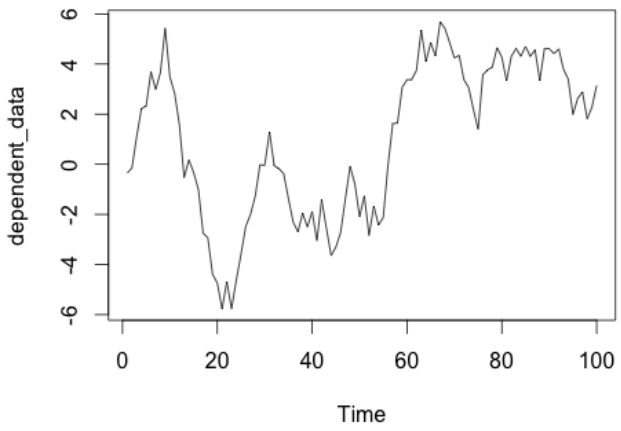

Fig. 3: Raw data of correlated observations

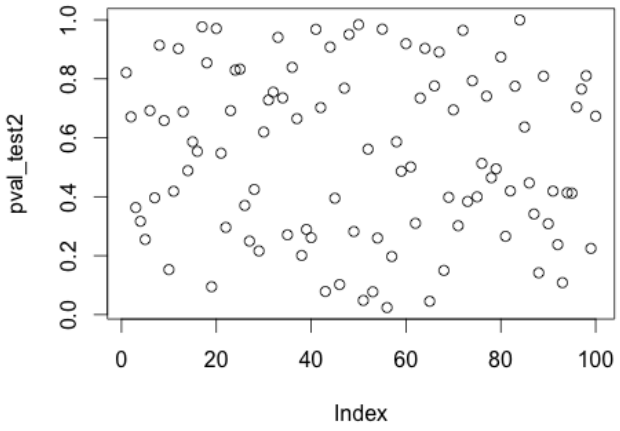

Fig. 4: p-value of t-test on independent data

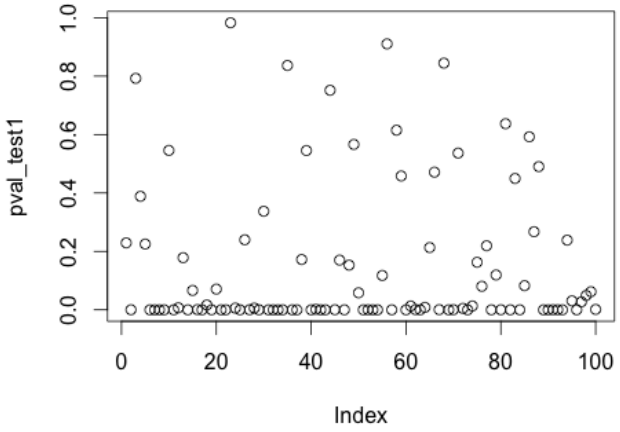

Fig. 5: p-value of t-test on dependent data

follows: an e-Commerce experimenter wishes to perform a day-to-day comparison of daily data to decide if the population mean in the variation group is significantly higher than that of the control group. A two sample paired t-test is done by calculating the difference in daily metrics between control and variation and building a one sample t-test on it. This produces misleading and inaccurate results because of autocorrelation. To demonstrate, a simulation experiment is conducted to generate two sample set A and B with 100 replications.

Sample A is collected from an AR(1) process with ϕ_1=0.9. Sample B contains 100 observations from a normal distribution N(0,1). Both sets have a population mean of 0. The raw data, along with autocorrelation plots from one replication are contained in Figure 2 to Figure 3. P-values of the t-test (with $\mu = 0$ as null hypothesis) on these 100 replications of Set A and B are collected and shown in Figure 4 and Figure 5. There is a much higher proportion of low ($\leq$0.05) p-values for Set A as compared with Set B, indicating a significantly higher type I error. In contrast, the percentage of low p-values in Set B is close to the expected significance level of 0.05.

In situations where the central limit theorem cannot be applied (e.g., test statistic is not the sample mean or sample size is small), non-parametric tests can be used. It has been our observation that more often than not, e-Commerce data will be non-normal and contain outliers. The use of non-parametric hypothesis tests is not as frequent in practical application on e-Commerce sites as parametric tests, but for certain situations, consideration for such tests may be warranted as prescribed in the following examples. For testing location parameters, experimenters can use rank based tests such as Wilcoxon rank-

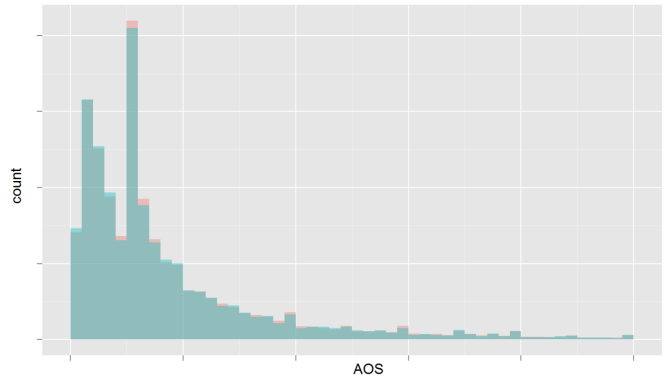

Fig. 6: Original AOS data

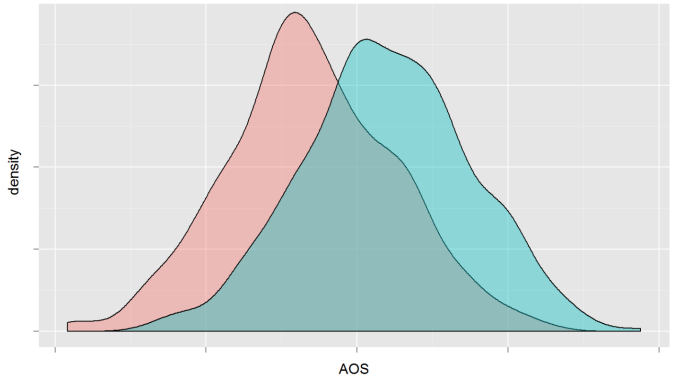

Fig. 7: Bootstrapped AOS data

sum test and Wilcoxon signed-rank test and bootstrapping approaches. As always, assumptions for both the rank-sum test and the signed rank test should be verified. Bootstrapping is a more flexible approach as it allows testing on various distribution parameters. The simplest procedure involves resampling with replacement on observations to create a large number of bootstrap samples, calculating test statistics on each bootstrap sample, and building a interval confidence on the calculated statistics values. We encounter this frequently in our work: in Figure 6, where the red curve indicates control and the blue curve indicates variation test groups, we show the skewness of the original AOS variable from which hypothesis testing provided results with no statistical significance. In Figure 7 we show the bootstrapped AOS variable from which we were able to obtain statistically significant results under hypothesis testing.

The result of a hypothesis test is defined by its p-value or the probability of the test statistic being at least as extreme as the observed one under the hypothesis that H_0 is true [12], [13]. E-Commerce experiments generally obey a 0.05 p-value threshold or type I error of less than 0.05 as rule of thumb. Type II error should be calculated accordingly using type-I error, factors of sample size, distribution of test statistic, and acceptable level of difference between the hypothesized value and true value. Both type I and type II errors should be identified prior to making a decision on the result of a test as both errors result in a decision that reduces the business value of the e-Commerce site. In the case of a false positive, the e-Commerce business will launch the feature which may prove to have little or negative impact. In the case of a false negative, the e-Commerce business will not launch the feature which means that the engineering investment become costs without return. Depending on the state of business and risk tolerance, appropriate error acceptance levels should be set.

V. Know Your Population

Understanding the population distribution of e-Commerce metrics is an important part of making inferences from controlled experiments. Knowing the empirical distribution of metrics: whether it has a long tail, is unimodal or multi-modal, and is normal or non-normal provides valuable information about user behavior on the e-Commerce site beyond just the location parameter which is primary focus of A/B testing. With knowledge about the population, experimenters can make better determinations on outlier detection, metric definition and behavior, OEC determination, and hypothesis testing.

One of the primary methodologies in analyzing the population is to run an A/A test [11]. A/A tests are tests where there is no difference between control and variation and the target metric distribution is expected to be identical in the two groups. A/A tests should be used by experimenters to identify potential biases in testing and to assist in experiment design, especially in sampling. Data segmentation and visual inspection should be combined with knowledge of potential types of bias to identify the existence of any such bias.

In comparing the results of the A/A test, tests for homogeneity of the distribution rather than tests on location parameter should be used. A common pitfall in A/A testing is to use t-tests to validate control and variation test group similarity. However, in conducting an A/A test, we expect identical distributions between control and variation and a goodness-of-fit test would be more appropriate than a t-test. For example, the distribution for RPV is heavily skewed right with a mixture of a large zero component and a long-tail non-zero component. In analyzing the distributions between test and variation for this metric, we should conduct (i) a Chi-squared test on proportion to test for the same mixture between zero and non-zero components and; (ii) a Kolmogorov-Smirnov test on the non-zero component to test for the same distribution. Figure 8 shows the distribution of transformed non-zero components of RPV from control and variation groups in an A/A test. Both visual inspection and hypothesis test suggest that the distributions are the same.

A particularity of e-Commerce businesses is the severity of the effects seasonality to financial performance. In most retail businesses, holiday periods are perhaps the most notable example of seasonality and generally consist of the last two months before calendar year close which are referred to holistically as the holiday season. According to survey results from the National Retail Federation, sales during the holiday season have consistently composed approximately 20% of annual sales for the retail industry. For e-Commerce specific

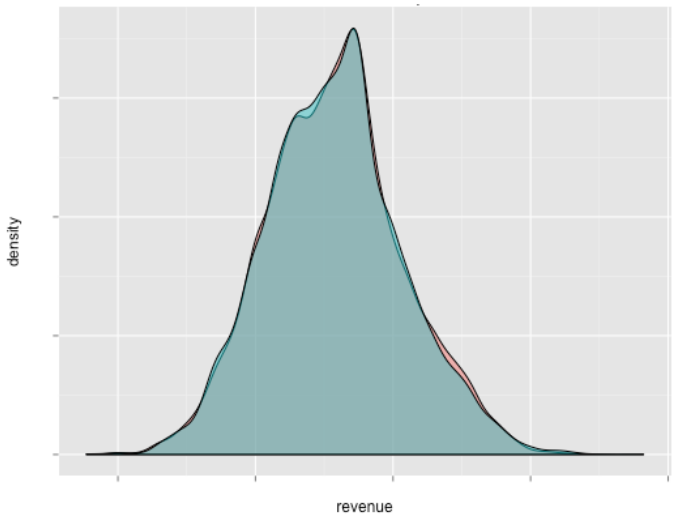

Fig. 8: RPV distribution in control and variation

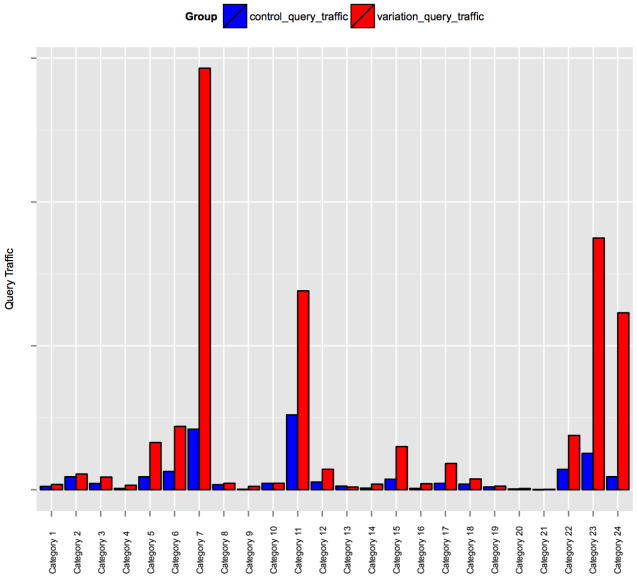

Fig. 9: Comparison of traffic by product category

operations of retailers, the impact and financial importance of this time is even more concentrated.

User behavior during the holiday season is markedly different when compared to the non-holiday, but is generally very similar across years. Figure 9 shows the traffic distribution by product category during holiday season compared to the non-holiday season where the differences are substantial. Figure 10 shows the traffic distribution before the calendar year-end for the past three years. The time series clearly shows trends in user behavior based on time of day and between weekdays and weekends that is consistent across years.

Because of the business importance associated with this time frame, A/B tests and production launches are typically halted or carried out with extreme caution. As a result, e-

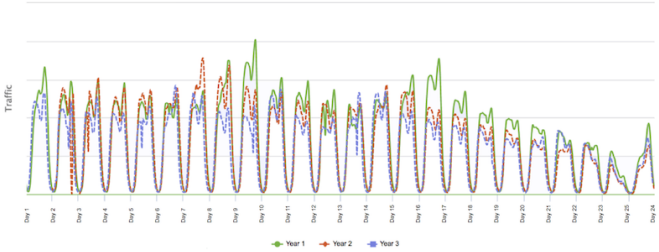

Fig. 10: Traffic trends before year-end

Commerce businesses face difficulty in being able to use A/B test results during non-holiday seasons to infer effects during the holiday season and to develop features that capitalize on gains during the holiday season. E-Commerce experimenters can understand the behavior of their users during the holiday seasons by conducting A/A tests which will not provide problems with production support and performance. Based on the findings from the A/A test, search features can be optimized for holiday season performance. However, it is important to note that the primary constraint is time: e-Commerce sites need to be able to make a decision in a very short period of time using controlled tests in order to capitalize on holiday traffic. A/B tests conducted during the holiday season can be run for a considerably shorter time frame as the volume of traffic reaches such levels that obtaining sufficient power is often not a problem. Experimenters may also wish to use variation reduction techniques [14] to reduce the sample size needed in order to reduce the time to a decision.

VI. Know Your Budget

A unique property of running controlled experiments on e-Commerce websites is its explicit relationship with money. Because the e-Commerce business is nearly entirely transactional, there is a limit to running A/B tests that reduce revenue. If an A/B test is generating revenue in its initial few days, there is frequently little imposition as to the length of its run. However, if an A/B test is losing revenue instead, a budget is set and the test will be stopped once it has reached its defined limit.

This property is particularly distinct in search where a single feature may impact a large amount of revenue. Stopping, analyzing, and the re-developing a feature or re-starting a test is a time intensive and costly process. The financial restrictions that define the nature of e-Commerce often force us to consider some form of prioritization prior to running A/B tests and to consider the possibility for certain features to forgo A/B tests to launch. In order to prioritize features for controlled experiments, we define a series of analyses and quality checks through which our feature development work flows. For search relevance, we measure the potential impact by comparing the change in ranking of a representative sample of queries and scoping priority based on the percentage of search queries, query traffic, conversion, and revenue touched. A team of search evaluation experts or a crowd of workers are

assembled to provide relevance ratings for query-item pairs and the Normalized Discounted Cumulative Gain (NDCG) [15] [16] is computed for the live site and for the feature in question. If there is a combined large potential impact and positive NDCG gain from the feature, we prepare the feature for an A/B test.

VII. Know Your Results

Analysis of an A/B test is crucial to reducing bias, avoiding a decision derived from false positives, and to discovering new opportunities for further development. At WalmartLabs, we perform a series of controlled experiment analyses including:

- Time series plots for all major metrics
- Visit level, query level, and common query level aggregations, bootstrapped results, and hypothesis tests
- Various segmentations including query traffic, product category, and user device
- Distribution of various other dimensions such as item price

Post-test analysis can be an extensive process with substantial engineering and data science effort. Rarely do e-Commerce organizations have the luxury to have an on-hand staff of statisticians to perform in-depth analysis for each and every controlled test. In order to scale the data-driven decision-making process at WalmartLabs, we implemented a pipeline and framework to generate automated analytics and easy data retrieval at the completion of each search A/B test. Having an automated analytics pipeline for controlled experiments is key to integrating applied science at scale in areas like e-Commerce search where feature development is highly algorithmic while decision-making is largely driven by business heuristics. Beyond scale, automated analytics also enables other benefits including reduction of error through standardization, comparability of features, monitoring of live tests, and a deeper understanding of feature performance. Through these analytics, we have identified several new insights which have been used to develop new versions of search relevance features.

In this process, data visualization becomes an important delivery method for insights. As the analysis becomes more complex and contains more slices of data, visualization becomes more effective in understanding results. Figure 11 shows a visualization of performance of three metrics by product category and the respective traffic distribution. From this visualization, we are able to identify which product categories performed well or poorly for metrics that concern our decision-making. Further, we can easily cross check the traffic distribution for each product category to temper our conclusions. The complexity of analyzing results by multiple dimensions is simplified through data visualization. We employ this philosophy throughout our automated analytics. In order to truly effect controlled experiments at scale in e-Commerce, both experimenters and business stakeholders need to be able to interpret results for effective decision making.

Fig. 11: Metric performance by product category matrix

VIII. Conclusion

In e-Commerce, controlled experiments, driven by the advent of big data and the need for retail organizations to become data-driven, have become fundamental scientific tools for decision-making. The applied science behind these controlled experiments have existed in a generalized form, popularized by large scale web technology companies and by research practitioners like Kohavi and his Bing research group. While these contributions have proven useful to other experimenters at large, there are sector specific difficulties and challenges that lie in the applied science of controlled experiments in e-Commerce and product search. In our paper, we highlighted the main distinctions of A/B testing for e-Commerce sites, provided the growing retail analytics community with guidelines for the application of controlled experiments, and shared sector specific challenges from our work. These contributions were addressed in the areas we believe to be fundamental in running successful controlled experiments: bias, metrics, hypothesis tests, population, budget, and results. In doing so, we hope to provide a more thorough understanding of e-Commerce behavior, the function of product search within retailers, and the details behind the applied science of controlled experiments throughout the domain of e-Commerce search. Although our work is specific to e-Commerce search, we believe that several of the topics we discussed can have

application in other functional areas including transaction-oriented businesses, content discovery products, and two-sided markets among others.

In closing, we present the research community with a series of open-ended questions that we are also currently actively exploring. These questions are both open areas for further exploration where there is little or no literature and represent important problems or potential value-adding contributions to e-Commerce businesses.

- How do we prioritize projects for A/B testing for general retailers who may not have as much traffic as big box retailers?
- How do we conduct controlled experiments during the holiday time frame when the traffic and sales profiles are considerably higher and time to action is shorter?
- How do we estimate the minimal loss budget for controlled tests when the tests can have significant revenue impact? This requires considering controlled experiments in a theoretic game setup and to design a process that guarantees a minimal loss.
- Is it possible to speed up feature launches by surpassing controlled testing using recent advances in counterfactual based simulation techniques? This is already an open area of research in the context of web search [17] and we think that such techniques may have value in the context of e-Commerce search particularly when the retailers may need to make fast product decisions before the holiday time frame.

ACKNOWLEDGMENT

The authors would like to thank Professor ChengXiang Zhai of the University of Illinois at Urbana-Champaign for his guidance and the members of WalmartLabs Search Analytics and Walmart eCommerce Testing and Optimization teams.

REFERENCES

[1] E. L. Lehmann, "The fisher, neyman-pearson theories of testing hypotheses: One theory or two?" *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1242–1249, 1993.

[2] R. A. Fisher, "Design of experiments," *British Medical Journal*, vol. 1, no. 3923, p. 554, 1936.

[3] R. Kohavi and R. Longbotham, "Online experiments: Lessons learned," *Computer*, vol. 40, no. 9, pp. 103–105, Sep. 2007.

[4] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: Survey and practical guide," *Data Min. Knowl. Discov.*, vol. 18, no. 1, pp. 140–181, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1007/s10618-008-0114-1

[5] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu, "Trustworthy online controlled experiments: Five puzzling outcomes explained," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 786–794. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339653

[6] S. R. A. Fisher, S. Genetiker, R. A. Fisher, S. Genetician, R. A. Fisher, and S. Généticien, *The design of experiments*. Oliver and Boyd Edinburgh, 1960, vol. 12, no. 6.

[7] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, 3rd ed., ser. Springer Texts in Statistics. New York: Springer, 2005.

[8] R. Kohavi, R. M. Henne, and D. Sommerfield, "Practical guide to controlled experiments on the web: Listen to your customers not to the hippo," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07. New York, NY, USA: ACM, 2007, pp. 959–967. [Online]. Available: http://doi.acm.org/10.1145/1281192.1281295

[9] R. Kohavi and R. Longbotham, "Online experiments: Lessons learned," *Computer*, vol. 40, no. 9, pp. 103–105, 2007.

[10] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham, "Seven pitfalls to avoid when running controlled experiments on the web," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 1105–1114. [Online]. Available: http://doi.acm.org/10.1145/1557019.1557139

[11] R. Kohavi, R. Longbotham, and T. Walker, "Online experiments: Practical lessons," *Computer*, vol. 43, no. 9, pp. 82–85, 2010.

[12] S. Goodman, "A dirty dozen: Twelve p-value misconceptions," *Seminars in Hematology*, vol. 45, no. 3, pp. 135 – 140, 2008, interpretation of Quantitative Research.

[13] A. Gelman, "Commentary: P values and statistical practice," *Epidemiology*, vol. 24, no. 1, pp. 69–72, 2013.

[14] A. Deng, Y. Xu, R. Kohavi, and T. Walker, "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ser. WSDM '13. New York, NY, USA: ACM, 2013, pp. 123–132. [Online]. Available: http://doi.acm.org/10.1145/2433396.2433413

[15] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002. [Online]. Available: http://doi.acm.org/10.1145/582415.582418

[16] Y. Wang, L. Wang, Y. Li, D. He, and T. Liu, "A theoretical analysis of NDCG type ranking measures," in *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, 2013, pp. 25–54.

[17] L. Li, S. Chen, J. Kleban, and A. Gupta, "Counterfactual estimation and optimization of click metrics in search engines: A case study," in *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2015, pp. 929–934.