Online Retail Store Subscription Pop-up A/B Testing Report

By: Anqi Wu and Britney Wang

Introduction

As loyal online shopping customers, we've noticed that pop-up windows requesting sign-ups for discounts are a common sight immediately upon entering an online store website. This observation prompted us to question whether the immediate presentation is the most effective timing to encourage customer sign-ups, or if there might be a more optimal moment that could enhance the likelihood of converting visitors into members. Therefore, we plan to conduct an A/B testing on the timing of the subscription pop-up to investigate whether adjusting the timing impacts the customer sign-up rate. The question that we want to solve is how to increase the number of customers signing up for an account on the website.

Test Design

Our hypothesis would be altering the timing of when the sign-up prompt is displayed, we can increase the conversion rate for email sign-ups. In the control condition, users will be presented with the sign-up prompt immediately as they enter the website, which will be their first interaction. For the treatment condition, the prompt will only appear after the user has engaged with the website in some way, such as by navigating to a new page or clicking on a product. To ensure a fair and unbiased comparison, we will randomly assign users to either the control or treatment group using a random number generator, with each group having a 50/50 chance of receiving a particular user. Adding on to that, this randomization can increase the statistical power of the test, so we are more likely to detect a true effect if one exists.

Data Collection

We gathered our data through a survey hosted on Qualtrics, designed with questions based on potential factors we hypothesized might influence someone's decision to sign up. Our questionnaire is divided into two main sections: user behavior and experimental inquiries. The user behavior section captures information such as the participant's gender, their online shopping frequency, whether they prefer to create an account or proceed as a guest, their likelihood of signing up during a purchase, and their concerns about privacy related to account creation. These questions are presented before any experimental queries, with the goal of gaining insight into our user base. This information is useful for us to understand the varied behaviors of users across different experimental conditions.

The experimental questions are presented to users after they have been informed about the condition they are in, which may influence their responses. Initially, we simply ask whether they would sign up through the popup window in the presented scenario. If they answered no, a follow-up question will inquire about their reasons for not signing up. Subsequently, all participants will be asked to evaluate the timing of the popup to gain insights into their satisfaction with the scenario they experienced. Additionally, they are to rate how the popup's timing might influence their perception of the website's credibility and their propensity to engage further with the website. These questions aim to gather data on the user experience and the potential impact of the experiment on their interaction with the website.

The survey begins with a questionnaire to gather background information from all participants. Following this, a randomizer evenly distributes users between the control and treatment groups, presenting a question about their willingness to sign up. If a participant answers 'no', a subsequent question probes for their reasons. Finally, all participants are presented with the remaining experimental questions. With this carefully structured survey, our aim goes beyond simply measuring the immediate impact of our strategies. We're also looking to unravel the less obvious factors that shape how people behave when they're shopping online.

Data Analyzing

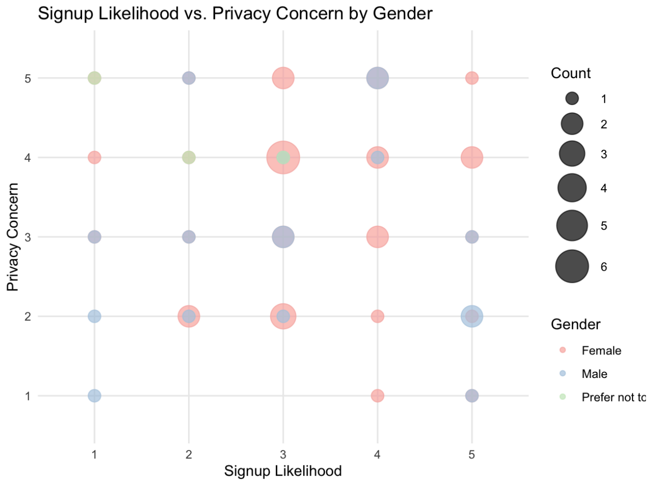
Data Cleaning:

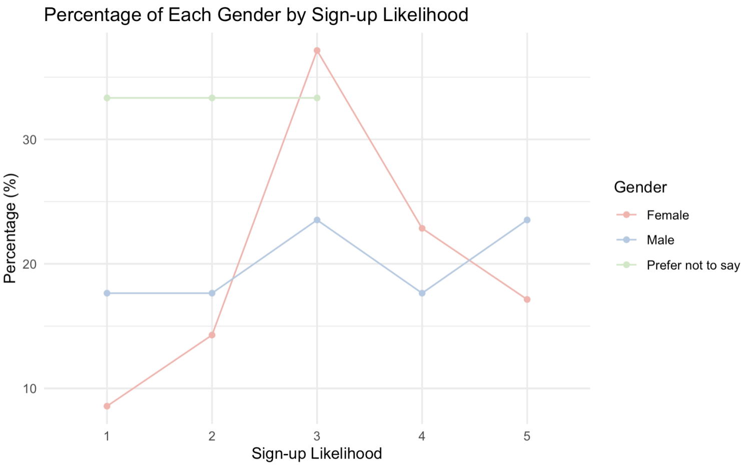
We received a total of 55 responses, with 27 participants assigned to the control group and 28 to the treatment group. The dataset, exported from Qualtrics, included several additional columns such as start and end dates, IP addresses, progress, duration, and recorded dates, among others. Additionally, Qualtrics automatically included the question count and headers within the first two rows. As a result, the raw dataset comprised 34 columns and 57 rows.

During the data preprocessing phase, we began by removing uninformative columns and renaming the remaining ones for clarity and coherence. Due to the survey's structure, responses to the key question—whether participants would sign up given a scenario—were split across separate columns. Furthermore, these columns contained blank responses, depending on whether the participant was in the control or treatment group. To streamline the data frame for analysis, we merged these columns and introduced a new ‘Group’ column to indicate the participant's group assignment. Following data cleaning, our refined dataset contained 13 columns and 55 rows.

User Behavior Analysis:

To gain deeper insights into our respondents, we have employed dot graphs to illustrate aspects of their backgrounds. The following figure 1 shows the relationship between signup likelihood and privacy concerns, with gender distinctions highlighted through color coding. Participants rated their privacy concern on a scale from 1 to 5, where 1 indicates minimal concern and 5 signifies extreme concern regarding privacy in the context of account registration. Similarly, for signup likelihood, the scale runs from 1, representing an unlikely intent to sign up, to 5, denoting a definite propensity to register during purchases.



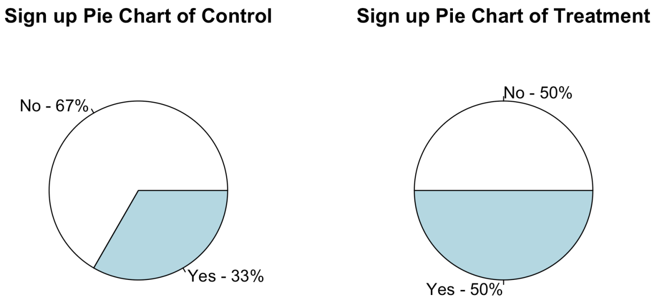


The data visualization presents an interesting distribution of responses concerning signup likelihood and privacy concern, segregated by gender. In the bubble chart, the size of each dot correlates with the number of individuals from each gender group; larger bubbles indicate a higher number of responses. The distribution of bubbles across the chart does not demonstrate a clear correlation between signup likelihood and privacy concerns, indicating the complexity and multiplicity of factors that influence users' decisions to sign up. The pink-toned bubbles, representing female respondents, appear more frequently and with larger sizes, which suggests a higher participation rate among females in the study. The overlapping of pink and blue bubbles, those purple dots, indicates areas where responses from female and male participants converge.

The line chart further elucidates the nuances in the data, with each gender plotted to show the percentage of respondents by signup likelihood. It reveals that females reported a peak in willingness to sign up at a moderate likelihood level (a score of 3), which may indicate a particular inclination towards creating accounts on retail websites post-purchase. In contrast, male responses show less variation across the likelihood scores.

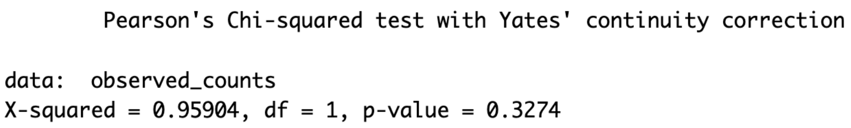
Accessing the Treatment Impact:

In the next phase of our analysis, we examined the signup proportions within the control and treatment groups. Our goal was to see which group exhibited a higher percentage of participants inclined to sign up under their specific conditions. In the following graph, we put two pie charts side by side to effectively contrast the signup rates between the control and treatment groups, which have similar numbers of observations. It appears that the treatment group has a higher proportion of affirmative responses to signing up via the popup window compared to the control group. While the count of positive sign-up responses did not surpass those saying ‘No’, there is a noticeable increase in the treatment group's signup rate when contrasted with the control group, which aligns with our hypothesis that applying the treatment will boost the sign up rate.



Sign up Pie Chart of Control

Sign up Pie Chart of Treatment



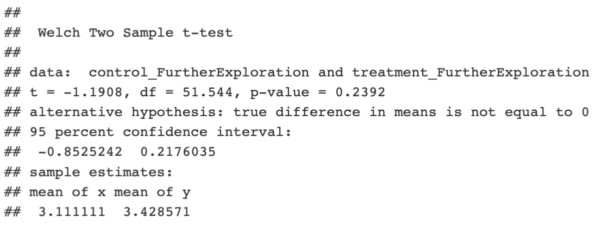
```
            Pearson's Chi-squared test with Yates' continuity correction

data:  observed_counts
X-squared = 0.95904, df = 1, p-value = 0.3274
```

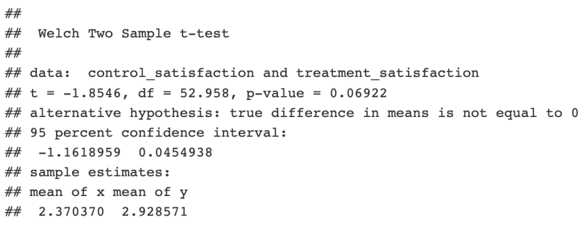
To evaluate the statistical significance of the observed differences in signup rates between the control and treatment groups, a Chi-squared test of independence was performed. The results of this test yielded a p-value of 0.3274, indicating that this difference is not statistically significant at the conventional 5% significance level. Consequently, while there is an observed increase, the evidence is not strong enough for us to conclusively state that the treatment led to a significant improvement in signup rates based on the data at hand.

Two sided t-test:

Subsequently, we analyzed the post-experiment outcomes to examine the impact of the treatment on user satisfaction, their perception of the website's credibility, and their willingness to explore the website further. To assess whether the treatment led to statistically significant differences between the control and treatment groups on these factors, we employed a two-sided t-test. Each of these three aspects was evaluated on a scale from 1 to 5, where a score of 1 indicates that the timing of the popup negatively affected user satisfaction, diminished the website's perceived credibility, and deterred further exploration by the users. Conversely, a score of 5 suggests the opposite effect.



```
##
##  Welch Two Sample t-test
##
## data:  control_FurtherExploration and treatment_FurtherExploration
## t = -1.1908, df = 51.544, p-value = 0.2392
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8525242  0.2176035
## sample estimates:
## mean of x mean of y
##  3.111111  3.428571
```

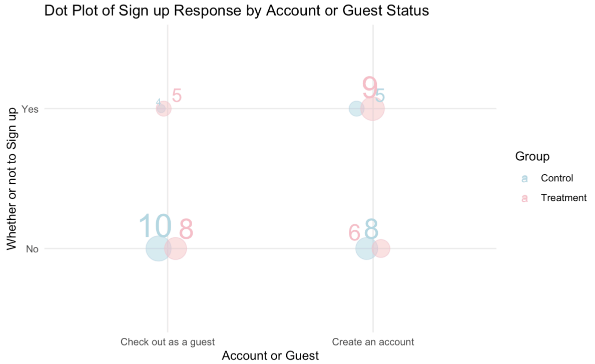


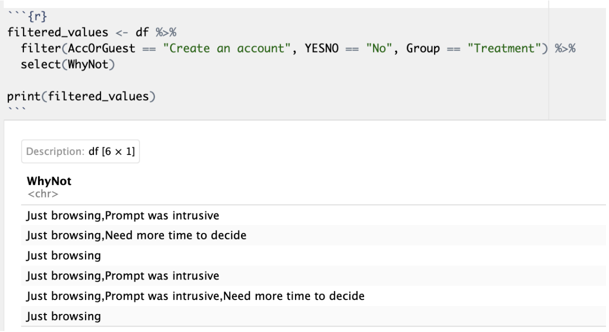
```
##
##  Welch Two Sample t-test
##
## data:  control_satisfaction and treatment_satisfaction
## t = -1.8546, df = 52.958, p-value = 0.06922
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1618959  0.0454938
## sample estimates:
## mean of x mean of y
##  2.370370  2.928571
##  Welch Two Sample t-test
##
## data:  control_credibility and treatment_credibility
## t = -2.2048, df = 50.212, p-value = 0.03208
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.97061928 -0.04525374
## sample estimates:
## mean of x mean of y
##  2.777778  3.285714
```

The results of the two-sided t-test revealed that only the credibility aspect had a p-value below 0.05, signifying a statistically significant effect of the treatment on users' perceptions of website credibility. By comparing the mean scores of the control group (x) with those of the treatment group (y), it becomes evident that users exposed to the treatment perceive the website as more credible. While the p-values for the other aspects did not indicate statistical significance, it is noteworthy that all three factors—user satisfaction, credibility perception, and willingness to explore—exhibited higher mean scores in the treatment group. This observation suggests that with a larger sample size, we might be better positioned to detect significant differences across these dimensions.

Difference-in-Differences (Diff-in-Diff) :

Utilizing the Diff-in-Diff approach, we examined the treatment's effect on user signup decisions. This method compares the pre- and post-treatment shifts in outcomes between a treatment group and a control group to isolate the treatment effect. In our dot plot, the initial 'sign up or guest checkout' choice is plotted against the post-treatment signup decision, a format echoing a truth table. However, unlike a standard truth table, our analysis is enriched by the Diff-in-Diff technique, which discerns behavioral changes attributable to the treatment.




~~~~
```{r}
filtered_values <- df %>%
  filter(AccOrGuest == "Create an account", YESNO == "No", Group == "Treatment") %>%
  select(WhyNot)

print(filtered_values)
```
~~~~

Description: df [6 × 1]

WhyNot <chr>
Just browsing,Prompt was intrusive
Just browsing,Need more time to decide
Just browsing
Just browsing,Prompt was intrusive
Just browsing,Prompt was intrusive,Need more time to decide
Just browsing

In the left graph, we can derive the Difference-in-Differences formula to gauge the treatment's impact. Initially, 15 individuals in the treatment group were inclined to create an account, compared to 13 in the control group. After the treatment was applied, the number of affirmative responses in the treatment group was 14, while the control group had 9. Employing the Diff-in-Diff calculation, (C-D)-(A-B), which represents the treatment effect, we find that (14-9)-(15-13) equates to 3. This figure signifies the net influence of the treatment on those who were treated.

Then, we can shift our attention to particularly insightful cases: the number of users who were initially reluctant to sign up but were swayed by the treatment, as well as those who initially planned to create an account but retracted their decision after the treatment. Notably, of the 13 individuals initially not considering signing up, the treatment managed to persuade 5 to reconsider and ultimately opt for signing up via the popup window presented after certain interactions. This outcome is marginally more effective compared to the control group, which only influenced 4 out of 14 individuals to change their stance. Then, we focused on individuals who initially were inclined to create an account but decided against it post-treatment and we delved deeper to explore the underlying reasons for their refusal. The filtered results, depicted in the graph to the right, indicated that these people all have 'Just Browsing' as their reasons in common. This suggests that while the treatment has potential in converting hesitant users, it may also inadvertently deter users who are in the exploratory phase of their engagement with the platform.

Results

The goal of our A/B testing experiment is to figure out the impact of timing on the effectiveness of subscription pop-up prompts in an online retail setting. We tested the hypothesis that adjusting the timing of when the sign-up prompt is displayed—immediately upon website entry versus after some engagement with the website—can increase the conversion rate for email sign-ups. To this end, we randomized users into a control group, which received the prompt immediately, and a treatment group, which encountered the prompt after engaging with the site.

In summary, the analysis of our data involved several steps, including data cleaning, user behavior analysis, and assessing the treatment impact through statistical tests and the Difference-in-Differences (Diff-in-Diff) approach. Our cleaned dataset comprised 55 responses, split nearly evenly between the control (27) and treatment (28) groups, and has 13 columns and 55 rows.

As a result, our observations indicated a trend where the treatment group demonstrated a greater tendency to sign up in comparison to the control group, which received an immediate prompt. This trend might imply that a delayed pop-up prompt could potentially be a more strategic approach in persuading visitors to subscribe. However, it's important to note that the difference in signup rates, while noticeable, did not reach statistical significance according to the Chi-squared test results (p-value of 0.3274). Consequently, the assumption that timing is a crucial factor in digital marketing strategies, although supported by the trend in our data, cannot be confirmed as statistically impactful based on this particular experiment.

The results of our two-sided t-test further underscored the treatment's effectiveness, particularly regarding users' perceptions of website credibility. Those in the treatment group rated the website as more credible on average than those in the control group. This finding indicates that the delayed presentation of the sign-up prompt not only influences signup rates but also positively affects users' trust in the website.

Our application of the Difference-in-Differences (Diff-in-Diff) analysis method provided additional insights into the treatment's effectiveness. This approach revealed a net positive effect of the treatment on the decision to sign up among users who were initially reluctant. Specifically, the treatment persuaded 5 out of 13 users who were not considering signing up initially, a conversion rate marginally better than that of the control group.

However, the study also brought to light a potential drawback of the treatment. While it was effective in converting a portion of users, it appeared to deter those in the exploratory phase, or "Just Browsing," from signing up. This observation implies that while the treatment shows promise in increasing sign-ups among certain users, it may not be universally appealing across all user types.

Limitations:

Our study encountered several limitations that are important to address for a better understanding of its outcomes and for guiding future research in this area.

First and foremost, the sample size of 55 participants significantly restricts the depth of insights we can derive from the data. With such a limited pool of responses, our ability to detect nuanced behaviors or make broad generalizations about user preferences and behaviors is constrained. This small sample size also limits the statistical power of our analyses, making it challenging to identify smaller but potentially meaningful effects of the treatment on user sign-up decisions. For instance, expanding the sample size could yield more insightful results for our t-tests, particularly in exploring user satisfaction and their responses to further engagement.

Another critical limitation arises from the constraints of our survey methodology, particularly the Overall Evaluation Criterion (OEC). The nature of the survey, which primarily utilized yes/no, multiple-choice, and ranking questions, limited our capacity to capture the complexity of user behaviors and their nuanced reactions to the subscription pop-up prompts. This format inherently restricts the depth of insights we can gather, especially concerning the reasons behind users' decisions to sign up or not.

To enhance the validity and applicability of our findings, a more robust method of data collection is needed. Observing actual user behavior on a website—such as click-through rates, time spent on specific pages, or engagement with particular elements—would provide a richer dataset from which to draw more detailed conclusions. Moreover, designing an experimental environment where users can interact with the website in real-time, rather than relying on hypothetical scenarios and screenshots, would offer a more accurate representation of their natural behaviors and preferences.

Additionally, tailoring the experiment to align with the users' interests could significantly reduce bias and increase the relevance of the findings. Recognizing that user preferences vary widely, an approach that matches the content or products shown with individual interests could lead to more genuine engagement and, consequently, more insightful data on their willingness to sign up. Users are more likely to express interest or engage with content that resonates with their preferences, which could, in turn, affect their response to subscription prompts. Incorporating these considerations into the experimental design would not only make the study more fair and accurate but also provide a more nuanced understanding of the factors influencing user sign-up decisions in online retail environments.