Validate_A_B_Test_Yammer

Case Study problem to validate A/B testing results for new Yammer publisher design. Completed as part of the requirements for Springboard Data Science Career Track. Author: Heather Passmore

Mode Analytics Case Study
Case Study: Validating A/B Test Results

Problem Summary: Results from one month of A/B testing of the Yammer publisher design indicate increased message posting in the treatment (new publisher design) group. Determine what the difference in posting rates indicates about control and treatment user groups.

Hypotheses that may explain treatment vs. control differences:

- Treatment and control groups were not evenly distributed (67% of users were in the control group).
    - What proportion of each group was posting?
    - Of users who posted what were the average number of posts per user?
- The types of users in the two groups may not be similar. What if one group consists of all new users who are required to make a first post upon creating their new account?
    - Test this by comparing the 'created_at' dates from the Table 1: Users data table between the control and treatment groups.
- Despite the larger average of messages sent in the test group the standard deviation (variance around the mean) meaning that the group differences are not significant.
    - Look at box and whisker plots of the posts per group.
    - Look at other simple statistics to compare the groups.
    - Interpret the t-test statistic appropriately.
- Are the posts included in the counts per group from similar types of events?
    - Posts in 'event_type' related to signing up and getting started should not be included in counts if the 'event_types' are not evenly distributed between the two groups.
- Other characteristics between treatment and control groups should be checked:
    - Are 'send_message' events from the two groups evenly distributed among device types?
    - Were assignments to this Experiment randomly distributed for 'occured_at', 'location', 'device', etc. characteristics?
- Some other metric is better for comparing valuable user engagement of Yammer.
    - User logins
    - User 'likes'
    - Number of interactions between users.

Validating the results:
Check the statistical tests.

- In the original case study question the 'rate_difference' is calculated as (group average - control group average) which is correct according to Mode's background information.
- However, 'rate_lift' ("The percent difference in posting rates between treatment groups") is supposed to be ((group average / control group average) - 1) and it is calculated incorrectly as ((group average / control group average) / control group average).
- See Updated Statistics table below for 'new_rate_lift' calculated as suggested by the Yammer case study materials.
- Calculations for the Student's t-statistic use the variance instead of the standard deviation as stated in the case study description.
- See Updated Statistics table below, for recalculated 'new_t_stat'. This further inflated the t-statistic and may not be the best way to compare the groups.
- The bottom line for the statistics on this metric: be sure to use appropriate tests and make correct calculations to determine A/B effect.

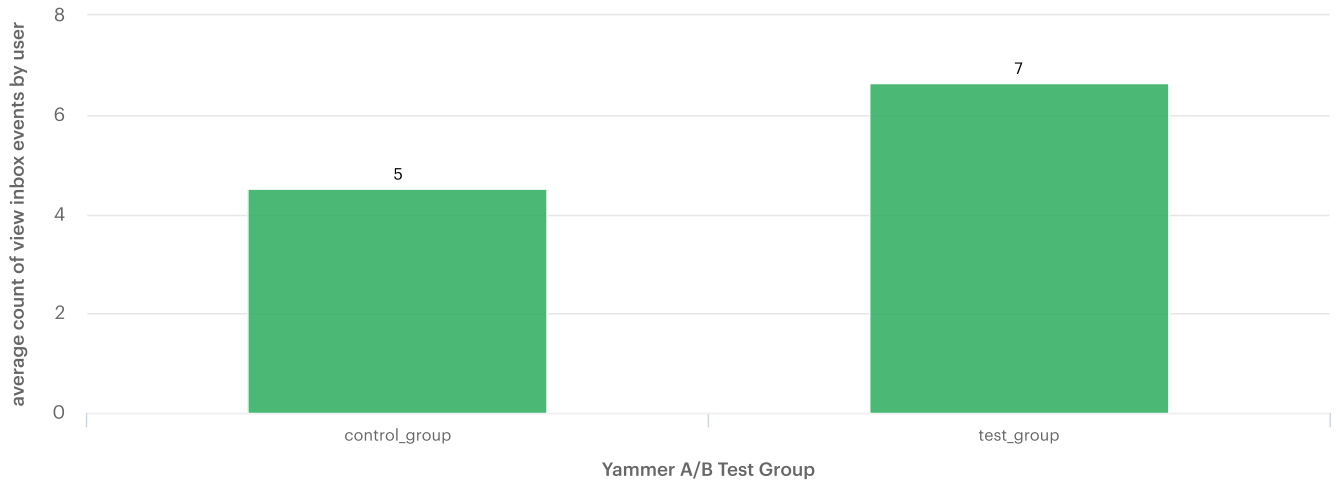
Updated Statistics

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	new_rate_lift
1	publisher_upda…	control_group	1746	2595	0.6728	4660	2.669	0	-1
2	publisher_upda…	test_group	849	2595	0.3272	3460	4.075…	1.4064	0.4064

Check other metrics:

- Are other Yammer user metrics similarly skewed for test group users?
- On average test group users have higher 'view_inbox counts (See 'Comparisons of average 'view_inbox' events per user' chart and table, below). Using the original student's t-statistic calculation method the differences are significant.
- Message likes per user are also higher on average for test group members compared to the control group. These differences are also significant using the original t-test method. (See chart and table below).
- Conclusion: many Yammer user metrics indicate increased Yammer usage by test group members.

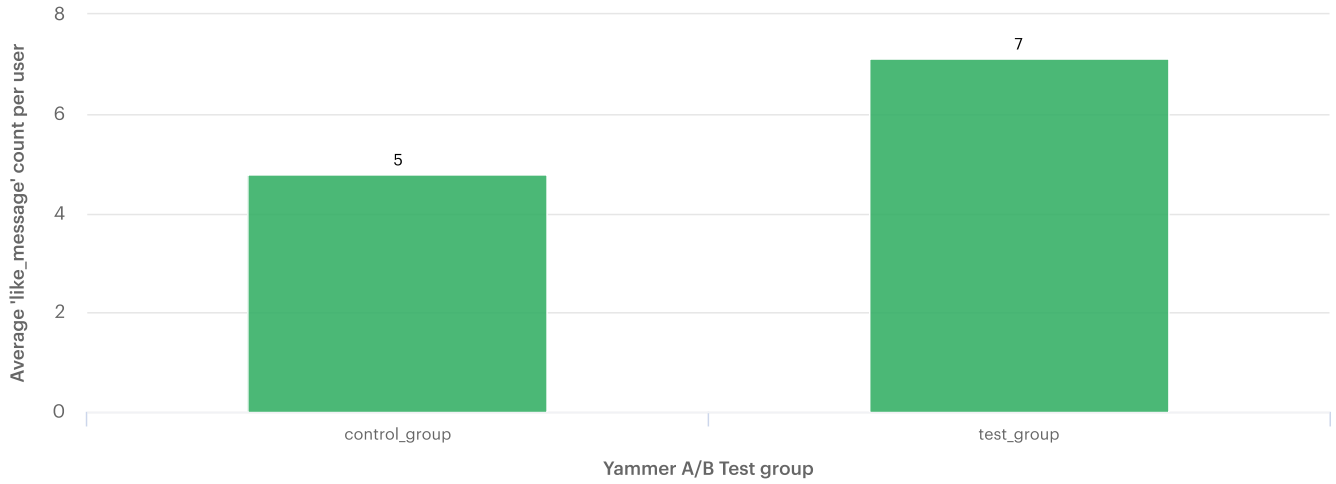
Comparison of average 'view_inbox' events per user



Like_message Count

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	std
1	publisher_upda…	control_group	1746	2595	0.6728	8329	4.770…	0	0	6.(
2	publisher_upda…	test_group	849	2595	0.3272	6036	7.109…	2.3392	0.490…	8.(

Comparison of average message likes per user



View_inbox count

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	std
1	publisher_upda…	control_group	1746	2595	0.6728	7866	4.505…	0	0	5.8
2	publisher_upda…	test_group	849	2595	0.3272	5633	6.634…	2.1297	0.472…	7.4

Check that the data are correct:

- In all control group vs. test group comparisons so far there are more test group members than control group members.
- What if there are different characteristics within the groups that are affecting their behavior?
- Users could be using certain device types in higher proportions in one group.
- Users could be newer in one group and therefore using Yammer at increased rates initially, and not a fair comparison if the other group is dominated by older users.
- To control for the amount of time A/B test participants have used Yammer I filtered Publisher Update experiment participants to only includes those users with 'activated_at' dates prior to May 1, 2014 . This results in smaller but more evenly distributed users counts in control and test groups (n=691 [control_group]; n=646[test_group]
- Below, the chart 'Users joined by May 2014, send message counts' includes the average number of sent messages by users of control and test groups. Here, test_group has a higher average but the differences are smaller. The test statistic is smaller and the p-value is larger but still significant.

Users joined by May 2015, send message counts

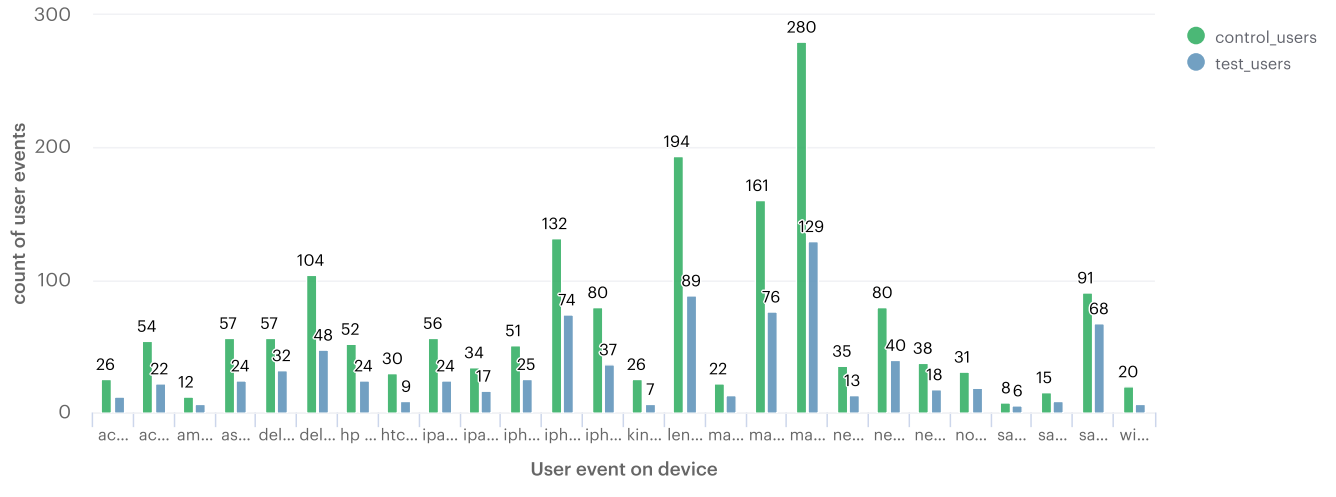
Old Users Send message

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	std
1	publisher_upda…	control_group	691	1337	0.5168	2083	3.014…	0	0	3.8
2	publisher_upda…	test_group	646	1337	0.4832	2749	4.255…	1.2409	0.291…	4.8

Check other factors that may may groups different:

- Control and test groups should be comprised of users with similar characteristics. For example, what if one group is using one type of device more often.
- Below chart 'Comparison of control vs test group user events on devices. Includes the count of user events for the control and test groups, and what type of device they were using. For most devices there are more actions by control users.

Comparison of control vs test group user events on devices



Final Recommendation:

- The comparison between control and test groups for the Yammer Publisher update need to be reviewed and revised before making decisions about the new Publisher format.
- The two experimental groups should be more homogenous. If one group is comprised predominantly of new users the usage comparison may not be fair.
- Device type and other variables should be checked so that characteristics are more homogenous among groups.
- Check several metrics to compare the two groups, not just average of send_message events.
- Appropriate statistical methods should be applied. What is the best measure of error, what kind of t-test, how big should the samples size be to estimate the population.

End of Yammer Case Study report.