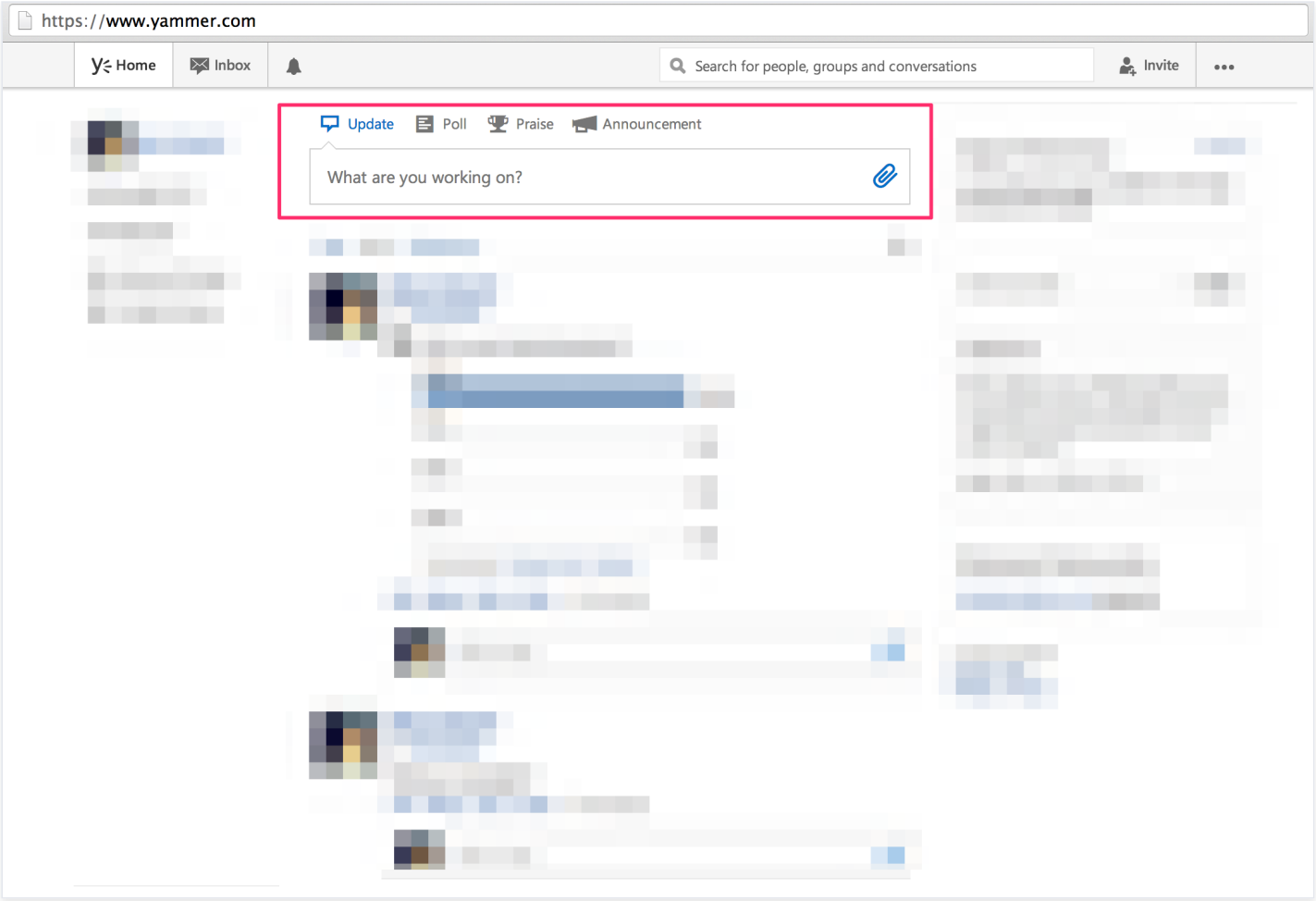
MODE

Validating A/B Test Results

*Before starting, be sure to read **the overview** to learn a bit about Yammer as a company.*

Yammer not only develops new features, but is continuously looking for ways to improving existing ones. Like many software companies, Yammer frequently tests these features before releasing them to all of thier customers. These **A/B tests** help analysts and product managers better understand a feature's effect on user behavior and the overall user experience.

This case focuses on an improvement to Yammer's core “publisher”—the module at the top of a Yammer feed where users type their messages. To test this feature, the product team ran an A/B test from June 1 through June 30. During this period, some users who logged into Yammer were shown the old version of the publisher (the “control group”), while other other users were shown the new version (the “treatment group”).



The problem

On July 1, you check the results of the A/B test. You notice that message posting is 50% higher in the treatment group—a huge increase in posting. The table below summarizes the results:

View Mode Analysis

The chart shows the average number of messages posted per user by treatment group. The table below provides additional test result details:

- **users:** The total number of users shown that version of the publisher.
- **total_treated_users**: The number of users who were treated in either group.
- **treatment_percent**: The number of users in that group as a percentage of the total number of treated users.
- **total:** The total number of messages posted by that treatment group.

- **average:** The average number of messages per user in that treatment group (total/users).
- **rate_difference:** The difference in posting rates between treatment groups (group average - control group average).
- **rate_lift:** The percent difference in posting rates between treatment groups ((group average / control group average) - 1).
- **stdev:** The standard deviation of messages posted per user for users in the treatment group. For example, if there were three people in the control group and they posted 1, 4, and 8 messages, this value would be the standard deviation of 1, 4, and 8 (which is 2.9).
- **t_stat:** A test statistic for calculating if average of the treatment group is statistically different from the average of the control group. It is calculated using the averages and standard deviations of the treatment and control groups.
- **p_value:** Used to determine the test's statistical significance.

The test above, which compares average posting rates between groups, uses a simple Student's t-test for determining statistical signficance. For testing on averages, t-tests are common, though other, more advanced statistical techniques are sometimes used. Furthermore, the test above uses a two-tailed test because the treatment group could perform either better or worse than the control group. (Some argue that one-tailed tests are better, however.) You can read more about the differences between one- and two-tailed t-tests here.

Once you're comfortable with A/B testing, your job is to determine whether this feature is the real deal or too good to be true. The product team is looking to you for advice about this test, and you should try to provide as much information about what happened as you can.

Getting oriented

Before doing anything with the data, develop some hypotheses about why the result might look the way it does, as well as methods for testing those hypotheses. As a point of reference, such dramatic changes in user behavior—like the 50% increase in posting—are extremely uncommon.

If you want to check your list of possible causes against ours, read the first part of the answer key.

The data

For this problem, you will need to use four tables. The tables names and column definitions are listed below—click a table name to view information about that table. *Note: This data is fake and was generated for the purpose of this case study. It is similar in structure to Yammer's actual data, but for privacy and security reasons, it is not real.*

Table 1: Users

This table includes one row per user, with descriptive information about that user's account.

This table name in Mode is tutorial.yammer_users

user_id:	A unique ID per user. Can be joined to user_id in either of the other tables.
created_at:	The time the user was created (first signed up)
state:	The state of the user (active or pending)
activated_at:	The time the user was activated, if they are active
company_id:	The ID of the user's company
language:	The chosen language of the user

Table 2: Events

This table includes one row per event, where an event is an action that a user has taken on Yammer. These events include login events, messaging events, search events, events logged as users progress through a signup funnel, events around received emails.

This table name in Mode is tutorial.yammer_events

user_id:	The ID of the user logging the event. Can be joined to user_id in either of the other tables.
occurred_at:	The time the event occurred.
event_type:	The general event type. There are two values in this dataset: "signup_flow", which refers to anything occuring during the process of a user's authentication, and "engagement", which refers to general product usage after the user has signed up for the first time.
event_name:	The specific action the user took. Possible values include: **create_user:** User is added to Yammer's database during signup process **enter_email:** User begins the signup process by entering her email address **enter_info:** User enters her name and personal information during signup process

	complete_signup: User completes the entire signup/authentication process **home_page:** User loads the home page **like_message:** User likes another user's message **login:** User logs into Yammer **search_autocomplete:** User selects a search result from the autocomplete list **search_run:** User runs a search query and is taken to the search results page **search_click_result_X:** User clicks search result X on the results page, where X is a number from 1 through 10. **send_message:** User posts a message **view_inbox:** User views messages in her inbox
location:	The country from which the event was logged (collected through IP address).
device:	The type of device used to log the event.

Table 3: Experiments

This table shows which groups users are sorted into for experiments. There should be one row per user, per experiment (a user should not be in both the test and control groups in a given experiment).

This table name in Mode is tutorial.yammer_experiments

user_id:	The ID of the user logging the event. Can be joined to user_id in either of the other tables.
occurred_at:	The time the user was treated in that particular group.
experiment:	The name of the experiment. This indicates what actually changed in the product during the experiment.
experiment_group:	The group into which the user was sorted. "test_group" is the new version of the feature; "control_group" is the old version.
location:	The country in which the user was located when sorted into a group (collected through IP address).
device:	The type of device used to log the event.

Table 4: Normal Distribution

This table is purely a lookup table, similar to what you might find in the back of a statistics textbook. It is equivalent to using the leftmost column in this table, though it omits negative Z-Scores.

score:	Z-score. Note that this table only contains values >= 0, so you will need to join the absolute value of the Z-score against it.
value:	The area on a normal distribution below the Z-Score.

Validating the results

Work through your list of hypotheses to determine whether the test results are valid. We suggest following the steps (and answer the questions) below:

- Check to make sure that this test was run correctly. Is the query that calculates lift and p-value correct? It may be helpful to start with the code that produces the above query, which you can find by clicking the link in the footer of the chart and navigating to the "query" tab.
- Check other metrics to make sure that this outsized result is not isolated to this one metric. What other metrics are important? Do they show similar improvements? This will require writing additional SQL queries to test other metrics.
- Check that the data is correct. Are there problems with the way the test results were recorded or the way users were treated into test and control groups? If something is incorrect, determine the steps necessary to correct the problem.
- Make a final recommendation based on your conclusions. Should the new publisher be rolled out to everyone? Should it be re-tested? If so, what should be different? Should it be abandoned entirely?

Answers

If you want to check your work against the answer key, **click here**.

Next Lesson
Validating A/B Test Results: Answers