Validating A/B Testing Results with SQL.

**The Problem**

Yammer has decided to test the improvement of a feature where users can type in their messages to publish what they’re working on via their website. In order to test the effectiveness of this feature to users, the product team ran an A/B test from June 1 to June 30, and during this period, users who logged into Yammer were randomly put into the two groups, the control group being the old publisher, and the treatment group being the improved one.

The problem lies in the average messages sent per group – the treatment group appears to be posting, on average, 50% more than the control. A two-tailed test upon two means is performed and the test statistic yields a significant result. These results either mean one of two things: that the improved feature is absolutely amazing, or the results are too good to be true, and should be investigated further.

**The Hypothesis & Solution:**

The statement of the problem does not necessarily say how the two populations are being sampled. Since the test is performed over the course of a month, we expect that (roughly) the two groups would be increasing at the same rate. The following query, however, yields results that tell a different story:

SELECT DATE\_TRUNC('month',u.activated\_at) AS month\_activated,

COUNT(CASE WHEN e.experiment\_group = 'control\_group' THEN u.user\_id ELSE NULL END) AS control\_users,

COUNT(CASE WHEN e.experiment\_group = 'test\_group' THEN u.user\_id ELSE NULL END) AS test\_users

FROM tutorial.yammer\_experiments e

JOIN tutorial.yammer\_users u

ON u.user\_id = e.user\_id

GROUP BY 1

ORDER BY 1

This query does in fact show that the two groups were sampled roughly equally up until the very end of the month, where it looks like something weird happened with how the two populations were being sampled. Further, this graph also tells us another crucial mistake, that new users were only being considered to be in the control group. While it is important to test the new users on the new features, it is very reasonable to expect existing and new users to behave differently. The suggestion here is to treat the two groups separately, so that perhaps two A/B tests are necessary