Frequentist vs Bayesian Testing:

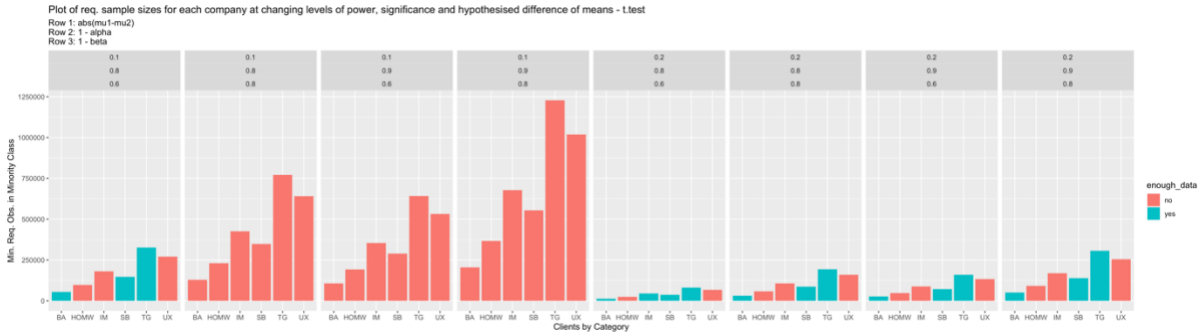
We want to know if and by how much we outperform the control group. This can be defined in terms of spend, conversion, retention or a number of any other metrics we are used to in our dashboarding tools. The question here is how to we bring statistical rigour to our set up.

A multitude of questions are brought up when deciding how to set up our testing:

1) **What are we testing?** ARPU / ARPPU, conversion, median spend, retention, total spend, and any other characteristic can be tested. Based on what we are testing the tests we use will change, to test for conversion we wouldn't use a t-test like we would for ARPU/ARPPU. The granularity of data also influences the tests we use given it changes the amount of data we get and the properties that data holds. A few simple questions have to be answered and were answered:
    a. **What are we testing for?** Takeaway: After long discussions about the data granularity we picked to go with meta_date/user_id/test_group on a daily basis and on a x days since first login basis and conversion.
    b. **How long are we comfortable with running the test for?** Takeaway: Decided this is about 30 days max, but sooner better.
    c. **Which tests do we use?** This can be split into Bayesian vs Frequentist:
        i. **Frequentist:** Given we have high sample sizes we can use parametric tests given assumption of independent and identically distributed data. This assumptions means CLM (Central Limit Theorem) applies and given high sample size (dependent on underlying distribution, finding the right minimum value requires for an assumption of underlying distribution and distribution having finite first 4 moments). So a t-test (even z-test) would be sufficient for ARPU/ARPPU tests and something like a proportion test would be sufficient for conversion tests. Takeaway: No need to go into non-parametric testing.
        ii. **Bayesian:** Many different ways of doing this and it's very dependant on what we want to do. A testing protocol will depend on how the data comes in. The way we do it is well described in the VWO paper linked below. Takeaway: Now looking back our assumption of conversion being a Bernoulli process is likely wrong, but not sure how big of a problem this is. What I mean: Assume spend is driven by conversion followed by spend, this is correct either your spend or not, and then you spend x. If the underlying process is just spend without conversion then P(x = 0) = 0 and there would be no non-spends (given point probability of continuous process is 0). But for us:
            1. 1) If we take meta_date, user_id, test_group level data: There is more than 1 possible conversion on each day so spends are being summed for users with multiple spends. 2) If we take day x in game, user_id, test_group level data: There are many many more than just 1 possible conversion.
            2. **Leads to:** spend process is different between these two testing protocols, which means log-normal might not be the right rpicess for both testing protocols.

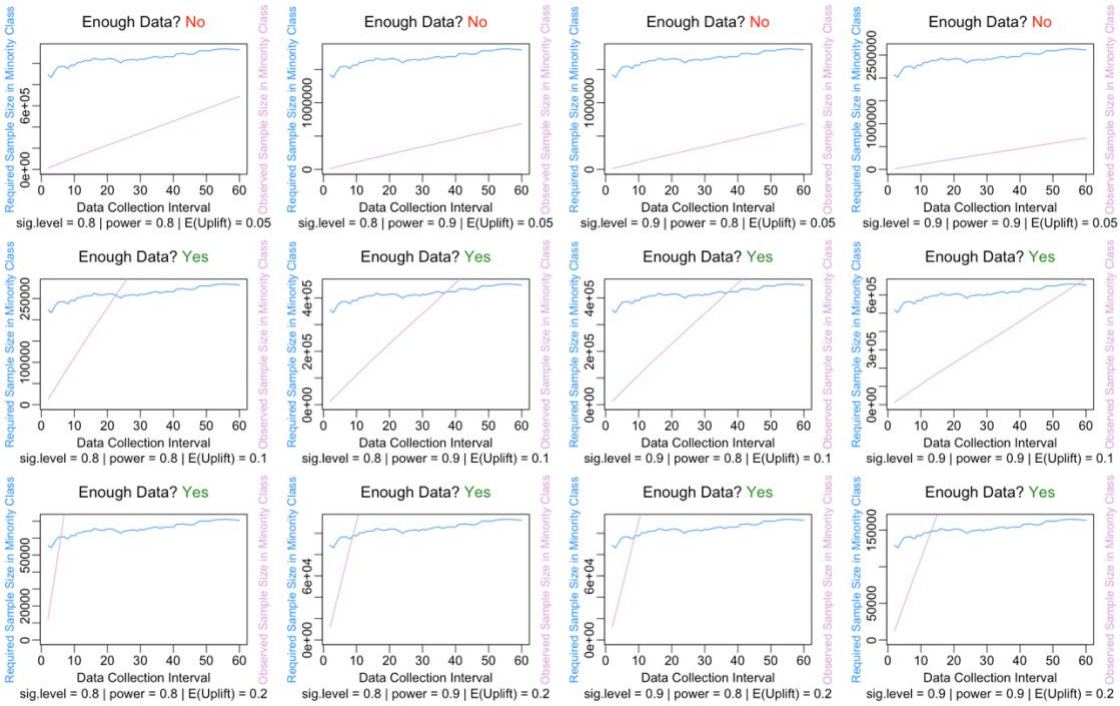
3. **Let's assume underlying distribution is log normal with mu, sigma:** Then the only thing changing between protocol 1, 2 and the most simple protocol possible (view level data), is the frequency of buying possibilities. So why are we assuming different priors for the two testing protocols? Because frequency gets loaded onto the log-normal process…
4. **Fix:** Model conversion as a discrete process with some parameters, something like Poisson distribution or binomial distribution. Allowing us to use the same spend priors irrespective of how we collect the data, only changing frequency of spends priors…

2) **Why are we doing Bayesian instead of Frequentist?** <u>Takeaway: Essentially because of how our data looks frequentists testing would take too much time.</u>
Long Answer: Given an A/A test variance and the desired confidence level, test power and minimum observable difference we can infer the minimum required minority sample size for a two sample, welched, t-test. Which as described above is an okay testing protocol given i.i.d. assumption and high sample size.



a. It shows if we get enough data to satisfy certain conditions for the minimum sample size of the minority class given:
    i. Test expected uplift, i.e. what is the expected difference in ARPUs between the two groups, the higher the expected observed difference the less data we need to get a conclusion for such a one sided hypothesis. Values looked at: 10%, 20%
    ii. Test desired significance level, so 1-alpha. The lower this is the less users needed. Values looked at: 80%, 90%
    iii. Test desired power, so 1-beta. The lower this is the less users needed. values looked at: 60%, 80%.
b. What do the colours mean: The values in red show **not enough** data in 31 days to run test. The bars in green show **enough** data in 31 days to run test.
c. Notice for HOMW and UX even the most conservative of tests (with the very very strong assumption of 20% uplift) show not enough users in 31 days.

Similarly looking at just HOMW data we see that only uplifts > 10% can be detected within 31 days which is rather poor performance:



3) **Now how do we do our Bayesian Testing?** Kaja knows best and VWO paper is the logical basis of this. But one of my caveats on underlying distribution selection: "One worry might be that selecting a model after considering the data is HARKing (hypothesising after the results are known; Kerr 1998). In that famous article, Kerr discusses why HARKing may be inadvisable. In particular, HARKing can transform Type I errors (false alarms) into confirmed hypothesis ergo fact. I.e. the non-linear trend in the population data might be a random fluke, and the better fit by the non-linear distribution might be a random fluke. Bayesian analysis is always conditional on the assumed model space. Often the assumed model space is merely a convenient default. The default is convenient because it is familiar to both the analyst and the audience of the analysis, but the default need not be a theoretical commitment. There are also different goals for data analysis: Describing the one set of data in hand, and generalising to the population from which the data were sampled. Various methods for penalising overfitting of noise are aimed at finding a statistical compromise between describing the data in hand and generalising to other data."
4) **What granularity of data is best?** This was made as a business decision however it is vital for assuming i.i.d. of data which is assumed in both frequentist and Bayesian testing I've discussed.

See Notebooks + Confluence:

- assetario.atlassian.net/wiki/spaces/AN/pages/1166475265/AB+testing

Data Simulations

See Notebook + Confluence:

- sims_sequential_simple.ipynb –https://github.com/assetario/ab-testing/blob/pnov/self_run_sims/sims_sequential_simple.ipynb

- https://assetario.atlassian.net/wiki/spaces/AN/pages/1243021313/AB+Testing+-+Generated+Data+Simulation+Studies+-+Results

Example Application

See Notebook:

- self_run_sims.ipynb - https://github.com/assetario/ab-testing/blob/pnov/self_run_sims/self_run_sims.ipynb

Reference:

https://vwo.com/downloads/VWO_SmartStats_technical_whitepaper.pdf