An introduction to A/B Testing:

In this section we would briefly go through A/B testing, which is one of the most common and important approaches in designing an experiment to improve web-content (or product). A/B testing is normally used in the context of a randomized experiment with two variants of a website, A and B. One of them is the control and the other is the treatment group. A/B testing is a form of statistical hypothesis testing with two variants (two-sample hypothesis testing) which is used in the field of statistics.

Some of the reasons that lead us to conduct an A/B test are the followings:

1) Deciding whether or not we should launch a new feature in our web-page,

2) Quantifying the impact of a feature,

3) Better understanding how do the users respond to addition of a feature and optimizing it.

A/B testing is possible to run off-line as well, it also works when you test a new operating system. In this lesson we would focus on the on-line A/B testing. However, you should be able to figure out some off-line scenarios once you finish the lesson.

We define the following steps to set up an A/B testing:

1) Define a goal and create supporting hypothesis. Also define the null hypothesis that you want to negate.

2) Identify treatment and the control group. Control group is the one where people see the web-content without the new feature and the treatment groups is the one where people see it with the new feature. Also make sure that you always change only one thing between the treatment and the control group.

3) Identify the key metrics to measure, for an website some typical key metrics can be number of searches made, total time spent by an user, click through rate etc.

4) Identify which data do we need to collect to successfully run the test, e.g. User id, cookie id, page loads, any engagement at the web page etc.

5) Make sure to enable logging everywhere , e.g. what happens if we press back button or log-out from the page etc.

6) Decide how much difference do we need to detect in order to consider it is worth investing in the experiment. e.g we want to track an increase of 5% in total time spent by users between the treatment and the control group.

7)What fraction of total visitors do we want to be in our treatment group ?

8) Using the power test decide how much data do we need to collect. Please check the related function in python [here](https://github.com/shubhabrataroy/Thinkful/blob/master/Curriculum/power_test.py).

9) Decide duration of the test, e.g. if you experience around 1000 unique users per day and if from 8 you identify around 50,000 people to be needed to successfully conduct the experiment then you need to run this experiment for at-least 50,000/1000 = 50 days.

10) Run a dummy A/A test where the control and the treatment groups are the same. This step is suggested to identify if there is any systematic bias in the experiment.

An added advantage in A/B testing is that if we observe that the treatment group is outperforming the control group well in advance, we can increase the fraction of visitors in that group to enhance the revenue or other related metric while conducting the experiment.

A/B testing is a versatile tool and can help you make huge improvements to your site. However the limitations of this test are summed up in the name. A/B testing is best used to measure at-most the impact of two to four variables (features) on interactions with the page, that too after tweaking a standard A/B test (multi-variate testing). Tests with more features are difficult to tweak and takes longer to successfully finish the experiment. A/B testing also does not reveal any information about interaction between variables on a single page.

Try it:

White down the steps two perform an A/B testing of a web-page where the only difference between the pages the color of a button. Also assume you want to increase one suitable metric of your web-page by 10%. How much data do you need for that. How long do you need to run your experiment. Can you identify all the 10 steps that we discussed in this section ? Also, try to identify the effect of significance levels (strict vs not so rigid) while you calculate the required data for the experiment.