**Lesson 1 Overview of A/B testing**

**Why do we do A/B tests?**

Testing takes the guesswork out of website optimization and enables data-informed decisions that shift business conversations from “we think” to “we know.” By measuring the impact that changes have on your metrics, you can ensure that every change produces positive results.

**Can we test everything?**

A/B testing is not good for testing new experiences. It may result in change aversion (where users don’t like changes to the norm), or a novelty effect (where users see something new and test out everything).

The two things with new experiences are a) having a baseline and b) how much time needs to be allowed for the users to adapt to the new experience, so you can say what is going to be the plateaued experience and make a robust decision.

Finally, A/B testing cannot tell you if you are missing something.

**What should we do when A/B testing is not useful?**

* Analyze the user activity logs
* Conduct retrospective analysis
* Conduct user experience research
* Focus groups and surveys
* Human evaluation

**What is the workflow of A/B test?**

1. Choose and characterize metrics to evaluate your experiments, i.e. what do you care about, how do you want to measure the effect
2. Choose significance level (alpha), statistical power (1-beta) and practical significance level you really want to launch the change if the test is statistically significant
3. Calculate required sample size
4. Take sample for control/treatment groups and run the test
5. Analyze the results and draw valid conclusions

**Binominal distribution**

For a binomial distribution with probability p, the mean is given by p and the standard deviation is where N is the number of trials. A binomial distribution can be used when

1. The outcomes are of 2 types
2. Each event is independent of the other
3. Each event has an identical distribution (i.e. p is the same for all)

**Hypothesis Testing**

The null hypothesis states that the difference between the control and experiment is due to chance. If pcont and ptest are the control and test probabilities, then according to the null hypothesis H0: pexp−pcont = 0

The alternate hypothesis is that H1: pexp−pcont≠0

**Comparing Two Samples**

For comparing two samples, we calculate the pooled standard error. For e.g., suppose Xcont and Ncont are the control number of users that click, and the total number of users in the control group. Let Xexp and Nexp be the values for the experiment. The pooled probability is given by

p̂ pool =

SEpool =

d̂ =p̂exp - p̂cont

H0: d=0 where d̂ ∼N(0, SEpool)

If d̂ >1.96∗SEpool  or d̂ <−1.96∗SEpool then we can reject the null hypothesis and state that our difference represents a statistically significant difference.

**Practical Significance**

Practical significance is the level of change that you would expect to see from a business standpoint for the change to be valuable. What is considered practically significant can vary by field. In medicine, one would expect a 5,10 or 15% improvement for the result to be considered practically significant. At Google, for example, a 1-2% improvement in click through probability is practically significant.

The statistical significance bar is often lower than the practical significance bar, so that if the outcome is practically significance, it is also statistically significant.

**Size vs. Power Trade-off**

One of the decisions is to determine the number of data points needed to get a statistically significant result. This is called statistical power. Power has an inverse trade-off with size. The smaller the change you want to detect or the increased confidence you want to have in the result, means you have to run a larger experiment.

As you increase the number of samples, the confidence interval moves closer to the mean

α=P(reject null | null true)

β=P(fail to reject null | null false)

1−β is referred to as the sensitivity of the experiment, or statistical power. People often choose high sensitivity, typically around 80%.

For a small sample, α is low and β is high. For a large sample α remains the same but β goes down (i.e. sensitivity increases). A good online calculator for determine the number of samples is [here](http://www.evanmiller.org/ab-testing/sample-size.html). As you change one of the parameters, your sample size will change as well. For example:

* If you increase the baseline click through probability (under 0.5) then this increases the standard error, and therefore, you need a higher number of samples
* If you increase the practical significance level, you require a fewer number of samples since larger changes are easier to detect
* If you increase the confidence level, you want to be more certain that you are rejecting the null. At the same sensitivity, this would require increasing the number of samples
* If you want to increase the sensitivity, you need to collect more samples

**Lesson 2 Metrics**