**How Ubisoft Used A/B Testing To Increase Lead Generation By 12%**

The test was conducted for the Ubisoft’s video game: For Honor

**Flow of an A/B test**

1. Set Null hypothesis
2. Choose your Confidence level
3. Do Power Analysis
4. A/A test
5. Standard deviation
6. Standard error
7. Statistical significance
8. Interpret and present the result

Let’s just assume the hypothesis is that the new design will generate better leads.

Let’s rephrase the hypothesis as: the newly designed page will drive the revenue generation up by at least 5% by getting more customers to buy.

**Assumptions:**

Control:

Let’s calculate the approx. average revenue generated per week.

3000 customers per week and usually 60% of them end up buying.

Cost of one game is $35

Average Revenue generated is ((0.60\*3000)\*35)+(.40\*3000\*0)/ 3000= $21

Hypothesis: The newly designed page will drive the revenue by at least 5%

How would I approach the problem?

1. Set Null hypothesis: H0: there’s no difference between the revenue generated by the control and that of the treatment.

Ha: There is at least 10% increment in the revenue generated by the treatment. Thus there is at least (>=) 10% positive difference.

1. Choose your Confidence level: The confidence level is (1-alpha). The Confidence level is the probability of failing to reject (i.e., retaining) the null hypothesis when it is true. A 95% confidence level is commonly used for evaluating one Treatment versus a Control. That’s the case here so I ought to use 95% confidence level.
2. Do Power Analysis: I ought to understand how long should I run an experiment for? Power can be calculated using the formula: n=(16\*(sigma)^2)/(delta)^2

One has to have an understanding re sigma (standard deviation) and the delta is the minimum detectable effect that the experimenter wishes for. Delta is calculated based on what’s change is expected (either increase mostly or decrease) re the value of the agreed upon metrics. The 16 in the power formula corresponds to the 80% power.

n=16(15)^2/(21\*0.05)^2 and the n = 3265 approx. Point to be noted: this n here is for each treatment and control separately. In practice it’s probably a good idea to stick strictly, again this depends on the business context re how much money and resources will be spent for every additional number of customers sent to variants, with the number you have got out of the power calculation in order to run a test efficiently.

1. A/A test: a. Sanity checking for the metrics value. I will store the percentage of customers sent to both control and treatment in a record and verify the percentage against that of the record. Am I leaking data anywhere?

b. Also I will check for the specifics of a/b testing splits. For example: If I have decided to do equal percentage of customers per type of browsers then I will check whether I have achieved the intended effect using A/A test.

c. Next, I will check for the values of key metrics that are agreed upon. For example: if I am measuring the number of people who purchased the game, I will split the number of customers into two groups and calculate the CTR leading to purchase using how many people clicked the baseline (A). So I will be using same number of people in baseline1(A1) and baseline2(A2), but A1 will get the type of customers assigned to actual baseline and A2 will get the type of customers assigned to treatment. Next thing is to calculate CTR from both versions (by that I mean: I would have assumed an appropriate statistical test to measure the metric) and compare whether the results are equal.

d. I might also like to plot the p-value and make sure that the shape of the plots is not far off from a uniform distribution. This will tell me whether my assumption is correct, data is unbiased, and the system of instrumentation is set accurately for the experiment.

1. Standard deviation
2. Standard error
3. Statistical significance
4. Try to bring OEC into the picture: The concept of thinking: how does changing a button or a small area on a webpage for example, affects the rest of the page and ultimately how the change translates to the benefits of the business/product.

Why consider OEC: For example, a metric might be proven statistically significant using standard formulas for standard deviation and variance, but when considered with regards to OEC it might not be a useful one. We might not ship the change even it has proven significant. The point to be noted in such situations is that the standard formulas used in finding the variance of a metric tend to underestimate values, especially the case in low varying metric: CTR. In such cases, bootstrapping randomly drawn samples to find the variance has proven to be an accurate measurement. Bootstrapping is especially advised when different units of user assignment been used for A/B splitting and calculating the variance.

So when considering OEC try to calculate the variance based on the combination of different metrics involved.

1. Interpret and present the result:

Interpret the result

The standard practice is to calculate the p-value and compare it with the threshold at alpha=0.05. If p-value less than or equal to alpha then we reject the null hypothesis, else we fail to reject the null hypothesis.

Alpha: Type I error (False Positive)

Beta: Type II error (False Negative)

Power of the test: 1-beta (rejecting the null hypothesis when it’s false)

If the p-value is less than or equal to the alpha (p< .05), then we reject the null hypothesis, and we say the result is statistically significant. If the p-value is greater than alpha (p > .05), then we fail to reject the null hypothesis, and we say that the result is statistically nonsignificant.

Present the result

State the goal: what you went after

Present the results: whether it is a success or not

Present the evidence: logical

Explain the statistical logic behind the evidence depending on the audience

What the result means to the business and how it will affect the product if the change is shipped?

What questions will I ask?

(the highlighted sentences are the most important questions)

1. What does better lead mean?
2. What’s the metric and what’s the overall evaluation criteria of the metric?
3. We ought to generate more leads to cause what outcome?
4. Let’s assume CTR (Click Through Rate) is the KPM (Key Performance Metric or KPI: Key Performance Indicator). Then is that the best metric to measure the OEC?
5. What’s the causal relationship of this metric with the organizational long-term goals?
6. Are we planning to run this test across all platforms: app on TV, browser on TV, browser on computer, browser on mobile, app on mobile, app on console?
7. Players from which country are we planning to target?
8. Are we planning to target any specific type of customers?
9. Are we targeting browser specific/ country specific/ device specific customers, or customers that are constrained to other specifics