Experimentation Simulation Game

OIDD 899: AI for Business

Introduction to the assignment

In Session 10 (9/25), we will have a lecture on reinforcement learning and experimentation. Rather than just tell you about these topics, we have developed a simulation game to give you hands-on experience ahead of the lecture. Between now and Sept. 25th, you should:

- Form a group of 4-5 people
- Familiarize yourself with the simulation game
- Try different strategies for increasing your score and discuss these with your group
- Submit a 1-page write up of the strategies your group has come up with

During the live in-class session on September 25th, all groups will play the simulation game in a competition, with each group being shown an identical version of the simulation. Further details about the simulation and deliverables of the assignment are below.

What is the simulation game about?

This game is a simulation of A/B testing. A/B testing is a kind of randomized experiment in which the performance of different versions (variants) of something are compared. It is frequently used in e-commerce to compare versions of websites. A/B testing is traditionally done manually (as in this simulation), however, the task can be automated using AI (specifically, through reinforcement learning algorithms). We will discuss AI-driven A/B testing later, after doing manual A/B testing in this simulation.

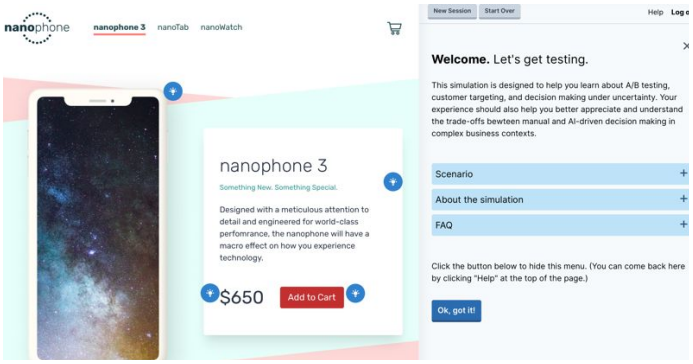
In this simulation, you will be testing different variants of a webpage about the nanophone 3. You will be able to test different variants of several aspects of the webpage, such as different image variants, subtitle variants, feature explanation variants, etc. There are 12 (simulated) weeks in the game, and you will set up the variants you want to test week by week. Each week, you will see the results of your A/B testing, including the conversion rate for each variant and the profit you earned, among other information. Your goal is to maximize profit over the 12 week period.

Getting started with the simulation

1) Start by accessing the simulation app using the information below:

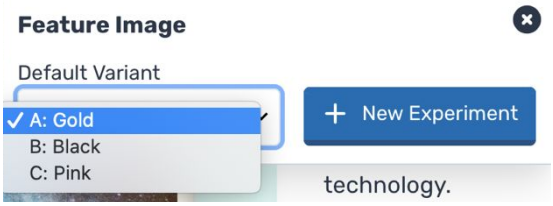
   - Link: https://production.hypothesis.whin.cc/
   - Username: PennKey username
     - E.g. “alexmill”
   - Password: PennKey username + 8-digit Penn student ID number
     - E.g. “alexmill39802834”

1) You will see the page shown in the screenshot to the right. Read through the information on the right side bar for more details about the simulation. Then click “Ok, got it!” to get started. If you experience a very long load time upon first opening the app (5+ seconds), performance should improve as you continue.

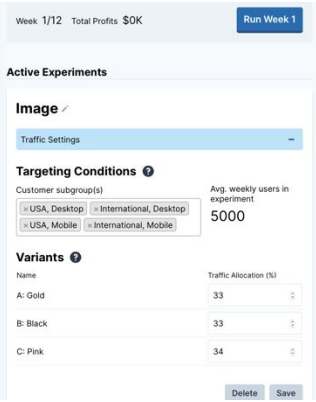


2) Click on the blue lightbulb icon next to the specific aspect of the website you would like to test different variants of. For example, if you would like to test variants of the nanophone image, click the blue lightbulb icon located near the top right corner of the nanophone image (see screenshot to the right).

3) After clicking the blue lightbulb icon, you will see information like what is shown in the screenshot to the right. Clicking on the choices in the "Default Variant" menu will allow you to see what the different variants look like on the website. To run an A/B test on this aspect of the website, click "New Experiment."



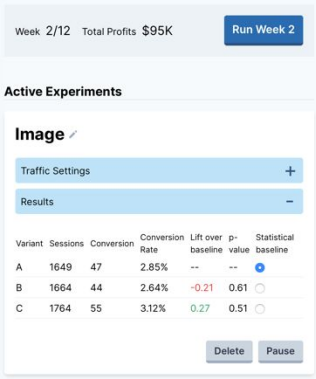
4) After clicking "New Experiment" you will see information in the right side bar that allows you to control the "Targeting Conditions" and "Variants" (see screenshot below).



   a. "Targeting Conditions" allows you to choose which subgroups you are running this particular A/B test on. In this example, we are running the test on all four possible subgroups. However, it is possible that different variants perform differently in different subgroups (some images, explanations, or taglines may be better for US users versus international users, for example). Narrowing your testing to particular subgroups could allow you to pick up on these differences. Please note that you can rename the heading of your test so you can keep track of changes you are making. For example, if you are running a test only on users in the USA, you could rename the heading to "Image, USA."

   b. "Variants" allows you to decide what percentage of traffic you want to allocate to each variant. In this example, we are allocating equal traffic to each variant. You can edit this to reflect the percentage you want to allocate to each variant.

5) To run this simulation, click "Save," followed by "Run Week 1" (in the screenshot above in step 4). You will then be asked to confirm that you would like to run this experiment. Click "Run Week 1 Experiments" to confirm. The app may take a few seconds to load the results of your simulation. You will then see the results of your simulation in the right side bar (see screenshot to the right). For each variant you will see the number of sessions you ran, the number that converted, the conversion rate, the lift over the baseline, and the p-value. The p-value is a measure of statistical significance, which will help you know whether changes you are seeing between variants are due to actual differences or simply to random noise. A p-value less than 0.05 indicates statistical significance (we have bolded the p-value if and when this is the case). You can also change the statistical baseline to change the comparisons between the variants. At the top of the right side bar, you will see the total profits.



6) Now you will move on to Week 2. You can make adjustments to the experiments you ran in the previous week and then run them again, you can run additional experiments, or you can pause the experiments you ran in the previous week (by clicking "Pause" as shown in the above screenshot) and only run new experiments. Please note that you can run multiple experiments on different aspects of the webpage in any given week. In the interest of clarity, these instructions provided an example of running only one experiment in Week 1, however, you can use the same process to run multiple experiments in any given week. Just add all the experiments you would like to run prior to clicking "Run Week X." Please also note that you will see two buttons at the top of your page: "New Session" and "Start Over." "Start Over" will allow you to start fresh while keeping the relative performance of different variants the same (e.g. if variant A of the image is truly better, it will remain better). "New Session" will scramble all data associated with the game- if you click this don't overfit what you learned last time to what will be true this time.

7) At the end of week 12, you will see a page that shows you your final score (as a percent) and your total profit earned. Your percentage score is your score compared to the best possible score. Scores higher than 85% would be extremely impressive! You will also be able to see the expected profits of different possible strategies, additional tips, and the best variant for each element and subgroup.

Deliverables

Come to class on 9/25 with **a 1-page write up of the strategies your group has come up with.** During class on 9/25, all groups will play the simulation in a competition (each group will be shown an identical version of the simulation).

Questions (to be updated as new questions come in)

From Vanessa Folkert... to Everyone:4:53 PM
Is there a multiplicity risk?
From Joshua Gordon to Everyone:4:55 PM
That's a good question Vanessa, you could correct with Bonefrroni to divide essentially set the threshold for P = 0.05/number of tests. That might be aggressive though and at its baseline really its not a yes no its just a smaller confidence interval.
From Vanessa Folkert... to Everyone:4:59 PM
I guess those corrections are built into the growth platform backend

Multiple hypothesis testing is a significant concern in real-world decision making. We've built our app based on standard paradigms in the A/B testing software industry, where standard practice is to report simple p-values for each experiment, unadjusted for any other factors.

From Joshua Gordon to Everyone:4:55 PM
If changing more than one factor simultaneously any way to look at covariance // Collinearity?

For this version of the simulation, there are no interactions between elements. For example, the best price does not vary based on what image your customers see. Running multiple experiments simultaneously may increase the noise of the signal you are able to observe in each experiment, but there should be no systematic interaction between on average.

From Vanessa Folkert
How do you manage data drifts / model drifts when doing large numbers of experiments?

If running an experiment, you could look at how the data varies with time to determine if this is an issue. For algorithmic experiments (e.g., bandits), handling data-drift in an optimal way is cutting edge technology at most A/B testing companies. The standard approach is to give more weight to more recent data, and discount the value of older data.

From NEIL SHAH to Everyone:4:51 PM
Bandit algorithm would do this automatically?

Yes. Stay tuned for more about this relationship in Session 10.