A/B Tests

1. Does a new supplement help people sleep better?

Partner with a University or a Medical Center to recruit people to participate in a sleep study. Send out announcements to a wide variety of people of different ages, gender, race/ethnicity and sleep patterns to recruit participants, from a variety of geographical areas.

Divide the groups into two. In this double blind study, the first group receives the new supplement, the second group receives a placebo. The participants first fill out a survey about their sleep patterns for two weeks before they begin taking the drugs. Once they begin the experiment they report their sleep patterns each morning over a period of one month. In particular, they record how many hours of sleep they get each night and whether or not they experience sleep interruptions. Sleep interruptions and hours of sleep per night are the primary metrics for the experiment.

We hypothesize that the people receiving the new supplement will sleep better. Meaning their sleep will have less interruption than the group without the sleep aid and they sleep a full 7-8 hours each night.

When comparing the groups we take their initial survey into consideration. Is there a difference between their initial reported sleep patterns prior to the experiment and their sleep patterns during the experiment? Is that change significant? Does region have an effect? For example do people in cold environments in general have different sleep patterns than in warmer environments? During the experiment we must make sure we select a month that is not very stressful, i.e. during the holidays. Age and gender must also be taken into consideration since sleep patterns can differ accordingly.

If people taking the supplement report they slept longer with less interruptions than usual and those taking the placebo report no change in their sleep – and we control for variations according to age, gender, location etc. We conclude that the supplement helps people sleep better.

1. Will new uniforms help a gym’s business?

This experiment will take place over a three-month period of time in March, April and May. These three months are chosen because we want to avoid the rush of gym membership right after the New Year, the inconsistency of gym attendance during the holiday season and the vacations that interrupt people’s gym attendance during the summer months.

For the first 6-weeks the old uniforms will be worn by the employees. The second 6-weeks employees will wear the new uniforms.

We hypothesize that the new uniforms will help the gym’s business. Our metrics are the number of new memberships and the level of attendance by existing gym members. We expect that during the time period when employees are wearing the new uniforms the number of new memberships will increase and the existing members will go to the gym with more frequency than during the time period when the original uniforms are worn. Since most gyms have an automated check-in system we will have access to all of this data through our database.

If we see from our statistics that during the experimental period with the new uniforms there is an increase in the number of new memberships and existing members are going to the gym more frequently we can conclude that the new uniforms have helped the business. Since we are measuring the patterns of all of the members of the gym who we assume are a diverse group of people, demographic bias is not much of a concern for this study. But we can disaggregate the data to understand if gender and age are factors in response to the new uniforms.

1. Will a new homepage improve my online exotic pet rental business?

Similar to the gym uniform experiment we will designate two time periods, one in which the original homepage is evaluated using data collected about rentals of exotic pets and one in which the new homepage is launched and data about rentals is collected. We will avoid conducting the experiment during major holiday periods to avoid bias. We assume that the homepage collects demographic data on its customers including gender and age, which will allow for an interesting comparison to see if there is a difference in how customers respond based on these categorical variables.

We assume that the new homepage will improve the rental business. Our primary metric is number of rentals. But we can also create categories such as customer referrals and see if there is an increase in referrals to friends/family/social networks during the respective time periods during the A/B test. We can also create a customer satisfaction response for customers to write comments about their experience with the business. It would be interesting to see if customers report an increase in satisfaction during the period in which the new homepage is live.

If during the time that the new homepage is up we see an increase in rentals, referrals or improved customer satisfaction we can conclude that the new homepage did improve the online rental business. Through a process of disaggregating the data we can evaluate whether or not there was a differential in impact based upon age and gender.

1. If I put ‘please read’ in the email subject will more people read my emails?

This experiment will be designed in a similar manner to the example provided of the test of green bold text vs. black italic text.

We would create a diverse group of recipients who will make up an email group that will represent our control group. We would randomly send one of the two versions (‘please read’ – or – without ‘please read)’ to this group. We would track the number of people who open the email.

We expect that the email with ‘please read’ will result in more people opening and reading the email. Similar to the example test provided by Thinkful, we can also track how quickly people open the email after receiving it. Since the people are in our group, we should know basic demographic information about the recipients, which we can then use to determine if there is a difference in response based upon these variables.

If the two groups are similar but open rates are higher and faster among those who received emails that say ‘please read’ than the group that received the email without ‘please read’ we could conclude that including ‘please read’ inspired people to read the email with more frequency than with out ‘please read’. As a result we would add ‘please read’ to all future emails.

Drill: Am I biased?

You’re testing advertising emails for a bathing suit company and you test one version of the email in February and the other in May

The design of the study is biased because the ads are tested during two different weather conditions. In February people are much less likely to purchase a bathing suit than in May, at the beginning of the summer season. Therefore there is a clear difference in the sample populations. This will give data that will be skewed and will not reflect the population, which could lead to false conclusions about the advertising campaign. The study should control for this bias by selecting one month and testing both advertising emails version during that month. By testing the email during a particular month and season they will gain more nuanced and accurate information that will enhance their overall advertising campaign.

You open a clinic to treat anxiety and find that he people who visit who a higher rate of anxiety than the general population

The context of the study is problematic because this is not a randomly selected group. The study population already has a higher rate of anxiety than the general population. Therefore will show bias. Would my investment in the success of my anxiety clinic influence my study design? To avoid this, someone knowledgeable about study design should review the experiment design to avoid observer bias. A better study design would sample a random selection of the community where the clinic is located. Ensuring that the sample population is reflective of the population is critical to designing an experiment without bias.

You launch a new ad billboard based campaign and see an increase in website visits in the first week.

There are gaps in the design of this study and limitations in its metrics. This study does not have a control and experimental group. What does the longitudinal data say about the campaign? The study should extend over a period of time that will allow for more data to be collected to understand the impact of the ad campaign. What is the result of these website visits? Are they converting new customers? What are the conditions of the ad campaign? Was it launched during the holidays? A study that does not account for these types of questions most likely will show bias and is not designed well enough to answer a particular question or capture a particular dynamic.

You launch a loyalty program but see no change in visits in the first week

To understand change in visits the study should create two different loyalty programs and test the respective impact on visits. To design the study well and reduce bias an A/A testing could be set up. Divide the visitors to the site in two groups and test the first loyalty program on the both groups. Then evaluate the results to see if they are similar. A third study would design an A/B test in which the study group is sent a loyalty program option and the control group in which the loyalty program is not available. This could be repeated with the comparative loyalty program. This design will provide strong data. What is the role of gender in the study? Perhaps one gender is more likely to respond to one loyalty program over another? Further disaggregation will produce interesting data about visitors and their likelihood to respond to loyalty programs.