15.9.1 Practice A/B Testing

Jeremy has finally rounded out his basic statistical tests and R practice and is ready to start applying these concepts to his job on the data team. As he begins to take a look at some of the analyses and reports from his colleagues, he notices that these reports keep referring to this concept of A/B Testing. Curious, Jeremy begins to search for more information about what A/B testing is and how he can use it.

Often when performing analysis and testing on a well-established product, website, or software, making changes can be difficult. Well-established products typically have a large consumer base and reliable sales and usage metrics, and are highly valued by their company. As a result, it's too risky to implement changes directly to the product without proper evaluation of the consequences.

To properly evaluate potential product changes, companies can use a technique called A/B testing. **A/B testing** is a randomized controlled experiment that uses a control (unchanged) and experimental (changed) group to test potential changes using a success metric. A/B testing is used to test whether or not the distribution of the success metric increases in the experiment group instead of the control group; we would not want to make changes to the product that would cause a decrease in the success metric.

Although A/B testing has been around for almost a century, giant software and tech companies such as Google and Amazon have popularized the practice by providing in-depth analytic metrics for their Google AdSense and AWS platforms.

Regardless of the industry or product, the process of A/B testing is the same. First, we must decide what changes will be made to the experimental group. Typically, the number of changes will be very limited to ensure comparisons are equal; however, more substantial changes can also be tested using an A/B framework.

Once a consensus has been made on the changes to be made to the experimental group, a success metric should be determined. The success metric can vary widely, depending on what is being tested. For example, a website might use consumer engagement as a success metric (e.g., number of visitors, clicked links, or time spent on a page). Alternatively, an automotive design team might want to know how performance changes after a slight design change to a vehicle's form factor, so the team's success metric might be mpg fuel efficiency.

Once we have decided on our experimental changes and the success metric, we must determine which statistical test is most appropriate. In this course, we'll only concern ourselves with normally distributed data and categorical data, which limits the number of statistical tests we'll need. However, if the A/B test groups are disproportionately uneven, or if the success metric distribution is non-normal, more elaborate statistical analysis may be required.

For our purposes, we can apply the following logic to determine the most appropriate statistical test:

- If the success metric is **numerical** and the **sample size is small**, a **z-score summary statistic** can be sufficient to compare the mean and variability of both groups.
- If the success metric is **numerical** and the **sample size is large**, a **two-sample t-test** should be used to compare the distribution of both groups.
- If the success metric is **categorical**, you may use a **chi-squared test** to compare the distribution of categorical values between both groups.

After determining the testing conditions and statistical test, the next consideration in A/B testing is sample size. It's important to collect a sufficient number of data points for each group to ensure that the A/B test results are meaningful.

There are multiple ways to determine optimal sample size, such as quantitative power analyses, but often a qualitative estimate is sufficient. If the changes made to the experimental group are expected to have a

strong effect on the success metric (often referred to in data science as an **effect size**), fewer data points are necessary for the test. In contrast, if the effect size is small, a larger sample size will be necessary for meaningful statistical findings.

For example, if we were testing purchase rates on an experiment group that receives a pop-up notification when visiting the AutosRUs website, we may use historical purchase rates as an indicator of effect size. If in general people who visit the site are likely to purchase a vehicle, our A/B test sample size can be small. However, if most people who visit the site are not likely to purchase a vehicle, we would need a large number of data points to confirm if the pop-up notifications make a statistical difference.

NOTE

> Using a quantitative power analysis can be helpful to determine sample size when effect size is unknown or resources are limited. Although performing a power analysis is outside the scope of this course, there is robust documentation online regarding **implementation (http://www.statsoft.com/Textbook/Power-Analysis)** and **interpretation (https://www.statisticssolutions.com/statistical-power-analysis/)** .

Due to its simple design and flexible application, the A/B testing framework is quickly becoming a go-to standard in the data science industry and one of the most highly desired data skills for Fortune 500 companies. Regardless, if you have experience in product design or optimization, you can use A/B testing to make informed design changes and confident development decisions.