1-**Data Cleaning**

Read The (ab\_data.csv) included and clean it

1. Null Values is an Error
2. Redundant or Duplication is an Error
3. The Data Shouldn’t include users their Groups is ‘treatment’ and the landing page is ‘old\_page ‘
4. The Data Shouldn’t include users their Groups is ‘control’ and the landing page is ‘new\_page’

**2- Hypothesis Testing (A/B Testing)**

Notice that because of the time stamp associated with each event, you could technically run a hypothesis test continuously as each observation was observed.

However, then the hard question is do you stop as soon as one page is considered significantly better than another or does it need to happen consistently for a certain amount of time? How long do you run to render a decision that neither page is better than another?

These questions are the difficult parts associated with A/B tests in general.

1. For now, consider you need to make the decision just based on all the data provided. If you want to assume that the old page is better unless the new page proves to be definitely better at a Type I error rate of 5%, what should your null and alternative hypotheses be? You can state your hypothesis in terms of words or in terms of and , which are the converted rates for the old and new pages.

**null hypothesis: Pnew<=Pold**

**alternative hypothesis: Pnew>Pold**

2. Assume under the null hypothesis, and both have "true" success rates equal to the **converted** success rate regardless of page - that is and are equal. Furthermore, assume they are equal to the **converted** rate in **ab\_data.csv** regardless of the page.

Use a sample size for each page equal to the ones in **ab\_data.csv**.

Perform the sampling distribution for the difference in **converted** between the two pages over 10,000 iterations of calculating an estimate from the null.