Experimental Design

Chi-Square Test for A/B Testing

When implementing an A/B test where the outcome of interest is categorical, we can use a Chi-Square test. Using a Chi-Square test requires us to collect data about each version (A or B) that a particular observation was exposed to and their outcome (eg., click or no click). For example, if wanted to determine if sending a person's name in the subject line of an email significantly increases the probability that the email will be opened, we might collect data that looks something like this:

Version	Outcome
A	Click
A	No Click
A	Click
B	Click
B	No Click
B	No Click

codecademy

Simulating A/B Testing Data

We can simulate data for an A/B test using the `random.choice()` function from NumPy. For example, if we want to simulate data where customers have a 65% chance of opening an email with their name in the subject line and a 50% chance of opening an email that does not have their name in the subject line, we could simulate a dataset of 100 recipients as shown.

```
import numpy as np

# randomly sample observations
sample_control =
np.random.choice(['yes', 'no'], size=50,
p=[.5, .5])
sample_name = np.random.choice(['yes',
'no'], size=50, p=[.65, .35])

# assemble those observations into a
dataframe
group = ['Control']*50 + ['Name']*50
outcome = list(sample_control) +
list(sample_name)
sim_data = {"Version": group, "Outcome":
outcome}
sim_data = pd.DataFrame(sim_data)
```

Significance Threshold

The significance level for any hypothesis test is the false positive rate for the test; therefore, if we simulate data for an A/B test where the true probability of the outcome of interest is the same in both groups (A and B), we'll find that the significance threshold is the proportion of simulations where the p-value is "significant", despite the fact that there is no real difference between the groups.

codecademy

Hypothesis Testing Power

The power of a hypothesis test is the probability of observing a significant result when there is one; therefore, if we simulate data for an A/B test where the true probability of the outcome of interest is different in the two groups (A and B), we'll find that the power is the proportion of simulations where the p-value is "significant", as it should be.

Impact of the Minimum Detectable Effect

The minimum detectable effect of a test ("desired lift" in the context of A/B testing) is the smallest difference between two groups that we care to measure in an experiment. If we decrease the minimum detectable effect (decide that we want to measure a smaller effect size), we'll need a larger sample size to detect that smaller effect.

Increasing Power in A/B Testing

The **power** of a hypothesis test is the probability of correctly finding a significant result. We can increase the power of a hypothesis test by increasing the sample size. We could also increase the power of a hypothesis test by increasing the significance threshold, but this method would also increase the chances of obtaining a false positive result.

codecademy

A/B Tests with Categorical Outcomes

For some A/B tests, the outcome of interest is categorical. For example, if we are testing whether a green background or an orange background leads to more customer purchases on a website, then our outcome (purchase or no purchase) is categorical. A/B tests with categorical results should be conducted using a Chi-Square hypothesis test.

Baseline Conversion Rate

A/B tests usually compare an option that is currently used (eg., the current version of a website) to something we think might be better (eg., a new version of a website) with respect to some metric (eg., the percent of people who click a link). The baseline conversion rate is the estimated value of this metric for the current version and is often based on historical data. For example, the baseline conversion rate might be the percent of people who click a link on the current version of a website.

```
baseline = 25/130*100
print(baseline) #output: 19.2
```

Minimum Detectable Effect

Detecting precise differences in A/B testing requires the use of large sample sizes. In order to determine how large of a sample is necessary, we must first know the smallest difference that we actually care to measure. This “smallest difference” is the minimum detectable effect and it is also referred to as the desired lift (usually in the context of marketing). The minimum detectable effect is often calculated as a percent of the baseline conversion rate. For example, if we want to increase our conversion rate from 6% to 8%, the lift would be calculated as shown.

```
baseline = 6
new = 8
min_detectable_effect = (new - baseline)
/ baseline * 100
print(min_detectable_effect) #output:
33.0
```

codecademy

Significance Threshold

The significance threshold for an A/B test is the false positive rate for the test: the probability of getting a falsely significant p-value. Meanwhile, most sample size calculators estimate the sample size needed to detect a true difference at least 80% of the time — and therefore lead to tests with a false negative rate up to 20%. There is a trade-off between false positives and false negatives (the higher the false positive rate, the lower the false negative rate will be and vice versa). When designing an A/B test, a researcher must determine the significance threshold (false positive rate) that they are comfortable with.

When to Stop an A/B Test

Calculating the sample size before starting A/B testing is important because test data is sensitive to sample size changes. It is crucial that testing ends as soon as the predetermined sample size has been reached. Ending testing early or extending testing beyond the predetermined sample size introduces bias into the test results and increases the chances of an error.

Print Share