A/B Testing With Python

A/B test

An experiment where you:

- Test two or more variants against each other to evaluate which one performs "best" ,
- A randomized experiment

Control and treatment groups:

Testing two or more ideas against each other:

Control: The current state of your product

Treatment(s): The variant(s) that you want to test

A/B testing process:

- Randomly subset the users and show one set the control and one the treatment
- Monitor the conversion rates of each group to see which is better

Considerations in test design

1. Can our test be run well in practice?

2. Will we be able to derive meaningful results from it?

Good problems for A/B testing

- Users are impacted individually

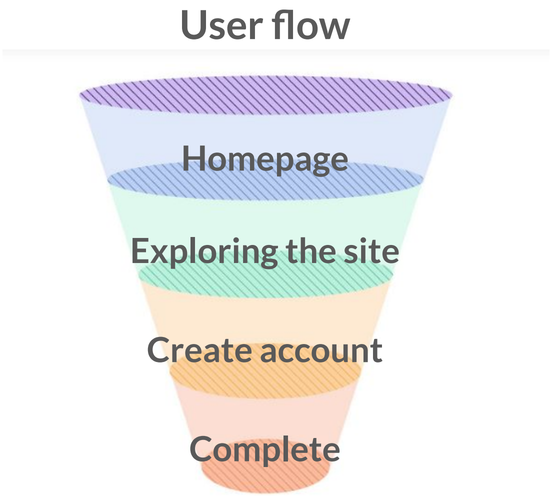
- Testing changes that can directly impact their behavior

- Eg:
    - Improve sales within a mobile application
    - Increase user interactions with a website
    - Identify the impact of a medical treatment
    - Optimize an assembly lines efficiency

Bad problems for A/B testing

- Cases with network effects among users
- Challenging to segment the users into groups
- Difficult to untangle the impact of the test

A/B Test Case: Amazon Prime Example



Hypothesis:

Changing the “TRY PRIME” button from yellow to red will increase how many buyers explore Amazon membership

Metric Choice

- People who have become a member of Amazon Prime ***[Obvious choice]***
- Number of clicks
- Number of clicks/ Number of page views(CTR)
- Unique visitors who click/unique visitors to page(CTP)

Original Hypothesis:

Changing the ***"TRY PRIME"*** button from yellow to red will increase how many buyers explore Amazon membership

Updated hypothesis:

Changing the ***"TRY PRIME"*** button from yellow to red will increase the click through probability of the button

How variable your estimate is likely to be?

Unique Visitors- 1000 (n)

Unique clicks- 100 (x)

1. CTP- ?
2. Which values will you surprised with? 101, 110, 99, 150, 900

Binomial Distribution

Features of Binomial

- 2 types of outcomes
- Independent events
- Identical distribution
    - ‘P’ of success needs to be same for all

Our click action will follow binomial distribution

Confidence Interval

Benefit of knowing it follows Binomial:

- We can use Sample Standard Error(SSE) for the binomial to estimate how variable we expect our prob. of the click to be.
- SSE can be used to find confidence interval at our desired range
    - For eg. ***95% Confidence interval:*** If we repeated the experiment over and over again, we would expect the interval we construct around our sample mean to cover the true value in the population 95% of the time.

Calculating confidence interval

Binomial distribution for large values becomes normal

- To use normal : (n.p`>5)
- Margin of error= z-score of confidence interval * Standard error

$$ME = z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

- Confidence interval= [p`- ME, p`+ME]

Confidence Level Example

x= 100 ; N=1000

p =0.1

$$ME = z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

z= 1.96 (?), SE= Sqr[(0.1*0.9)/ 1000]

ME=? ; CI=?

ME= 0.019

CI= [0.081 , 0.119]

Analysis: If you run this experiment again with 1000 views, you can expect any value between 80 and 120(But no values above or below this range)

Statistically Significant

Hypothesis Testing:

- P(Results due to chance)
  - Pexp - Pcont=0 [Null Hypothesis]
  - Pexp - Pcont !=0 [Alternative Hypothesis]
- Calculate Pexp and Pcont
- Calculate P(Pexp- Pcont| Ho)< 0.05 (Same as 95% confidence interval)

Pooled Standard Error

$$P = \frac{(X_{cont} + X_{exp})}{(N_{cont} + N_{exp})}$$

$$SE = \sqrt{(P * (1 - P) * (1/N_{cont} + 1/N_{exp})}$$

$$\hat{d} = P_{exp} - P_{cont}$$

Practical Significance

- Is the change significant enough to require practical actions?

Design experiment

First is deciding sample size i.e What size of impact is meaningful to detect 1%...? 20%...?

- Smaller changes = more difficult to detect can be hidden by randomness

Parameters for choosing size

Ɑ = Falsely concluding a difference (P(reject null| null true))

B = Falsely failing to draw a conclusion P(fail to reject|null false) [1-B= Sensitivity (80%)]

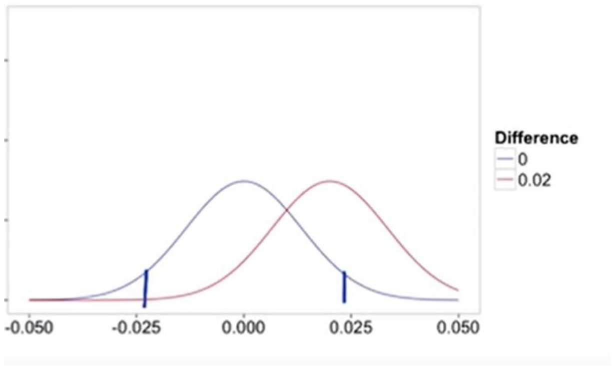
Small Sample -> Low Ɑ: You won’t launch a bad experiment

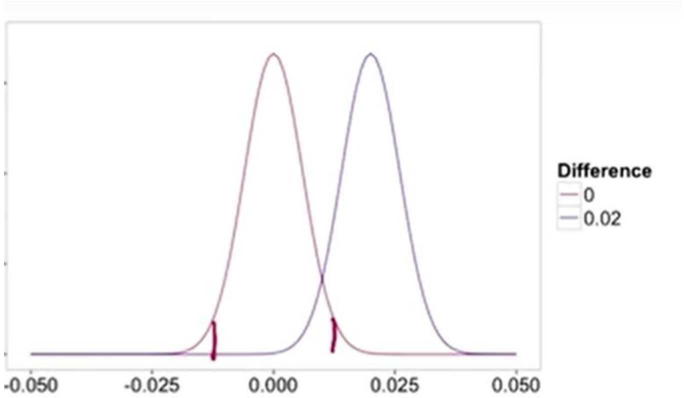
High B: You will fail to launch a good experiment

Large Sample -> Low Ɑ: You won’t launch a bad experiment

Low B: You won’t fail to launch a good experiment

Small sample vs Large Sample

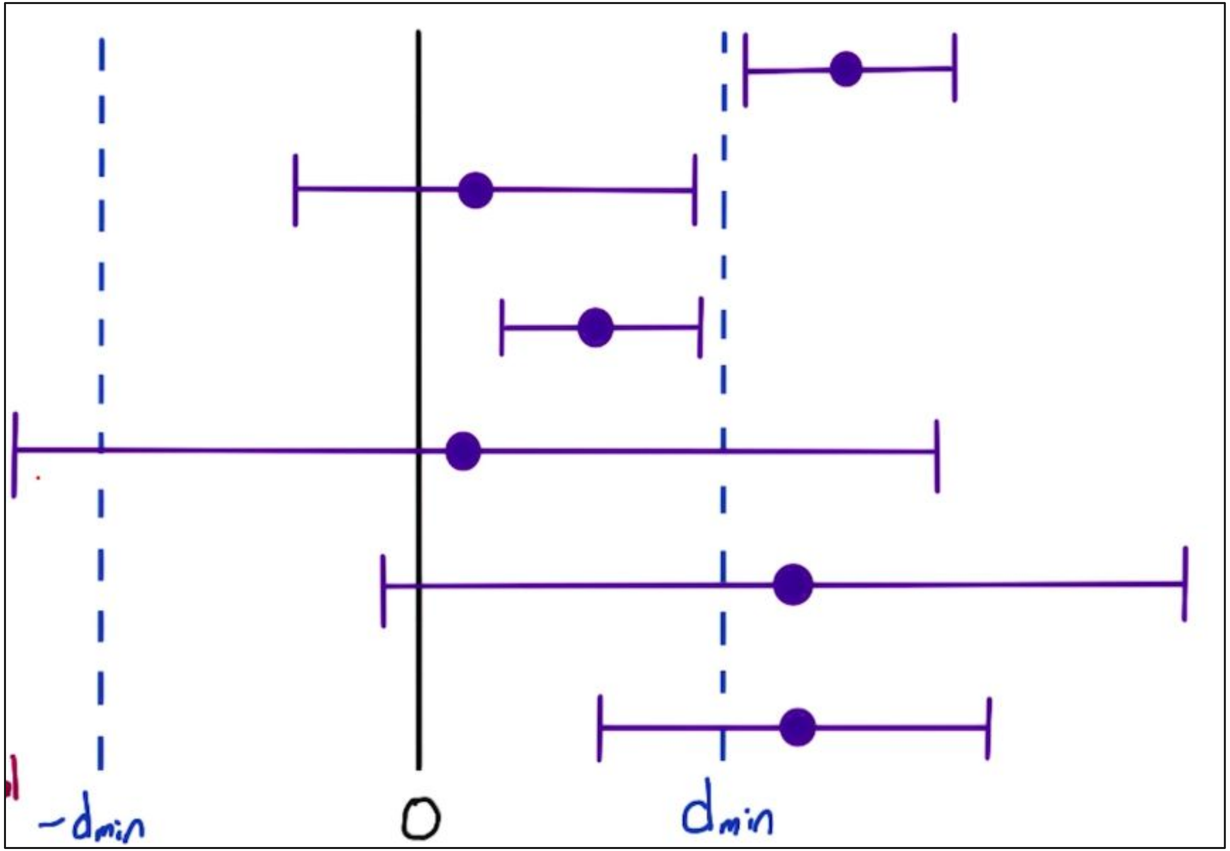




Calculate sample size

Analysing the results

Calculations



Accept or Reject?

TIME FOR THE ACTIVITY

Study Jam Activity

- Conduct A/B testing on “Free Trial" Screens
- Guided Individual activity
- Download the resources from here

Further Reads

- [P-value and False Positives]
- [Confidence Interval Vs P-Value]
- [AB Testing Example]