GENERAL ASSEMBLY

# INTRO TO DATA SCIENCE
# LECTURE 14:A/B TESTING

## KEY CONCEPTS

- Running a test with 2 ideas, A, and B.

- One idea, A, is better than the other idea, B. **1. How do we know?**

- The longer we run the test the better we are able to quantify how much better A is than B.

BUT

- The longer we run the test the more users who are exposed to the inferior idea.

- **2. How do we know when to stop the test?**

# A or B - Which is Better?

## Key Concepts - Problem Definition - Which is Better

Examples:

- Amazon resellers - who should you buy from?
    - Someone with 20000 reviews and a 90% positive rating, or someone with 10 views and a 100% rating

- App purchases: Will changing the home screen of your app, in a certain way, result in more in-app purchases?

- Advertising banner copy: Will changing the copy of a banner advert increase the web traffic to the seller's website?

## Key Concepts - Classical Experimental Design

- In all examples we are trying to measure some action in response to alteration in the text, copy, or appearance of a website, all other things being equal, with the purpose of deciding the ‘best’ text, copy or appearance in order to maximize web traffic to an linked site

- One way of doing this might be to measure the number of purchases over a time period with a given website.

- Change the website and re-measure the number of purchases for all users over another period of the same length.

- Compare the two measurements and decide which website was better.

## Key Concepts - Classical Experimental Design

- The are a number of problems with doing this:

1. Changing the website might result in less users and less purchases

2. You will be measuring purchases at different times of the year, what are the impacts of this. e.g. change in season, effects of holidays

3. It could take a long time to get enough information from enough users to make a good comparison

4. You cannot monitor significance - classical design requires you run the experiment to completion

## Key Concepts - Classical Experimental Design

- Say we want to be able to detect a difference in conversions at a 1 percentage point level.

- Pick a confidence interval (e.g. 95%), find the appropriate sample size, and run the test.

- At the end of the test we can say, A is better than B, or B is better than A, or A and B are within a percentage point; all with 95% confidence.

## Key Concepts - Bayesian A/B Testing

- Instead of trying to measure one scenario followed by another, Bayesian A/B testing seeks to measure differences simultaneously.
- Take 2 variations of a feature, promotion, advertisement, news headline
- Distribute each of them to unique and separate groups
- Measurements can be collected in real-time
- Criteria are met to allow a decision to be made as to which feature, promotion, advertisement or news headline is the more successful

*and*

- The losing feature, promotion, advertisement, or news headline can be replaced by the winner

## Key Concepts - A/B Testing

- The original feature, promotion, advertisement or news headline is known as the control or Variation A, while the new version is referred to as the test version or Variation B.

- A/B/n testing is an enhancement to allow testing of more than 2 variations

- In addition the experiment need only be run on a subset of the users, as the number of users increases so does the approximation to all users

## Key Concepts - Beta Distribution

- Model using the Beta distribution
- Takes 2 parameters, a, and b
- a = views x CTR, b = views x (1 - CTR)
- Beta(a, b)
- As we collect more evidence our uncertainty decreases

## KEY CONCEPTS - WHY THE BETA DISTRIBUTION?

- Iterative Bayes and the Notion of Conjugacy
- Posterior = Likelihood * Prior
- We are dealing with discrete probability distributions - the CTR is countable
- Someone either clicks through or doesn't
- Two outcomes - The Bernoulli Distribution
- The probability that x = 1, is given by the mean
- Mean = number of 1s divided by the number of trials

## KEY CONCEPTS - WHY THE BETA DISTRIBUTION? - THE LIKELIHOOD

- Example:
  - Coin Tossing, is like the Click-Through-Rate, it is a binary outcome
  - Data = {H, H, H, H}
  - H = 1, T = 0
  - The key parameter is the mean
  - We look at the mean over a number of trials and decide - is the coin biased, is the headline better…
  - Using the data to just measure a likelihood, only results in a point estimate
  - The mean of the Bernoulli distribution where 4 heads in a row are thrown is 1, not overly helpful

## KEY CONCEPTS - WHY THE BETA DISTRIBUTION? - THE BAYESIAN APPROACH - SEEKING THE POSTERIOR

- Use Bayes and get the posterior probability distribution of the mean of the Bernoulli Distribution?
- We can do that by estimating a prior distribution for the mean
- Question: You've flipped 4 heads in a row, what is your prior belief about the fairness of the coin?
- Posterior $\propto$ Likelihood * Prior *or* Posterior = Bernoulli * Prior
- Some distributions have what are called conjugate priors, which greatly, massively, simplifies Bayesian analysis
- The conjugate prior for the Bernoulli Distribution is called the Beta Distribution
- When a distribution has a conjugate it means the posterior distribution arising from Bayes will take on the same algebraic form as the prior.
- Beta = Bernoulli * Beta

## KEY CONCEPTS - THE BETA DISTRIBUTION

- Beta(a, b)
- a, b cannot be zero
- add 1 to a and b in order to ensure non-zero values
- The Beta distribution as a prior:
    - Let’s assume a fair coin, and hence equal priors
    - a = 1, b = 1
    - Likelihood = Bernoulli with a mean of 1 (from 4 heads out of 4)
    - Likelihood * Prior, results in a = 4 (heads) + 1, and b = 0 (tails) + 1
    - Posterior = Beta(4 + 1, 1)
    - mean = 0.83
    - variance (of the estimate of the mean) = 0.14

## KEY CONCEPTS - THE BETA DISTRIBUTION - COMPARING 2 HEADLINES

- Beta Distribution can describe a probability distribution parameterized by counts

- Have a beta distribution for headline A, and one for headline B

- In the case of headlines we assume an equal prior, i.e. we think A and B are equally likely to be the best, so a = 1, b = 1, and Beta(1, 1)

- Gather some data to update a, and b (the likelihood)
    - a = (views * CTR) + 1
    - b = (views * (1.0 - CTR)) + 1

- Estimate the posterior

## KEY CONCEPTS - COMPARING 2 HEADLINES - ITERATIVE BAYES

- The prior and the posterior are the same distributions, with identical mathematical formula

- Hence, we can iterate, making the posterior the new prior

# Key Concepts - Comparing 2 Headlines - Iterative Bayes

Data from Website

Posterior = Likelihood * Prior (initialize with equal priors)

New Prior x Likelihood

Headline A - mean and variance

Headline B - mean and variance

Data from Website

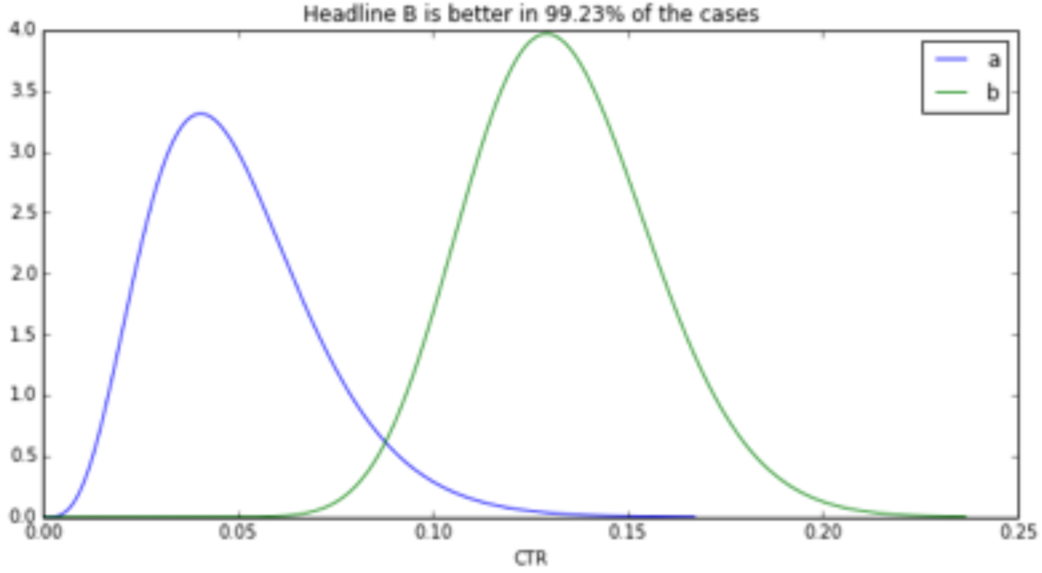
## Key Concepts - Deciding on Which Headline

- The interval in which 95% of the probability density is located decreases exponentially with respect to the number of views.
- By generating random values (i.e., drawing a random sample) from both beta distributions (representing each headline) we can identify which distribution is higher.
- By large sample random sampling we can accurately estimate the probability that B is better than A.
- This probability is the certainty with which we can declare headline B as the true winner.

# Key Concepts - Deciding on Which Headline

```
def percent_better(a_views, b_views, a_ctr, b_ctr, size):
    ra = beta.rvs(a_views*a_ctr, a_views*(1-a_ctr), size=(size))
    rb = beta.rvs(b_views*b_ctr, b_views*(1-b_ctr), size=(size))
    return sum(ra >= rb) / (1.0*size)
```

```
[12]: fig = figure(figsize=(10,5))
demonstrate(100,200, 0.04969, 0.13287, size=1000000)
```



# A or B - Minimizing Regret

## Key Concepts - The Additional Challenges of Instant Headline Testing

- Headlines may be on the front page for a short time, so testing has to be undertaken quickly
- The number of readers varies greatly per front page
- The CTR of a headline depends on front page position
- Front pages are dynamic, so headlines can change position

## Key Concepts - Problem Definition - Minimizing Regret

- When testing 2 headlines you need to decide which performs better, but with the aim of replacing the worst performing of the 2 headlines as quickly as possible
- Performance being measured by the Click Through Rate
- You need to run the experiment for as short a time as you can
- You are loosing traffic while you have a poorly performing headline live
- Regret - represents the loss you experience while a poorly performing headline is still present on your site

*BUT*

- The longer you run the experiment the more confident you can be with your decision of which headline to go with
- As more data comes in the variance of your estimate is dropping

Have 2 headlines live; how long do we have to run the experiment before we can choose a winner and replace the lesser performing headline?

## KEY CONCEPTS - INSTANT HEADLINE TESTING

- Allows editors to improve the quality of a headline after it has made the front page
- Decision making can be done quickly
- Overall they will see an uplift in CTR

# Key Concepts - Headline Testing - Example

## KEY CONCEPTS - HEADLINE TESTING - EXAMPLE

- The following headline was tested:

**Headline A: "What Harbaugh regrets about Super Bowl" (3.06% CTR)**

**Headline B: "John Harbaugh explains Super Bowl tirade" (4.93% CTR)**

- After only 7 minutes headline B was declared the winner!, with a 99.3% certainty
- The winning headline was then served to 100% of the audience for a further hour
- A 61% uplift was achieved, meaning tens of thousands of more viewers

## Key Concepts - Classical Approach

- Perform a statistical test to ascertain whether the CTR for one headline was significantly different from the CTR for the other headline
- This will provide an answer to which one is better
- It provides no answer as to when to stop the test
- Some third party software monitors the significance of the test as it proceeds in an attempt to indicate to the user when they can stop the test
- Real time significance test monitoring

## KEY CONCEPTS - CLASSICAL APPROACH

- The problem is that the statistical power of the test mandates running the test to conclusion
- You need to define ahead of the experiment how many observations will be collected, and stick to this
- After 200 observations a trial may be significant
- After 500 observations is may not be - as more data has arrived
- If you stop the trial after 200 observations then you eliminate the case where additional data provided evidence of a ‘not significant’ result

# Example - Monitoring Significance - What Not to Do!

| | *Scenario 1* | *Scenario 2* | *Scenario 3* | *Scenario 4* |
|---|---|---|---|---|
| *After 200 observations* | *Insignificant* | *Insignificant* | *Significant* | *Significant* |
| *After 500 observations* | *Insignificant* | *Significant* | *Insignificant* | *Significant* |
| *End of experiment* | *Insignificant* | *Significant* | *Insignificant* | *Significant* |

## Example - Monitoring Significance - What Not to Do!

| | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 |
|---|---|---|---|---|
| After 200 observations | Insignificant | Insignificant | Significant | Significant |
| After 500 observations | Insignificant | Significant | test stopped | test stopped |
| End of experiment | Insignificant | Significant | Significant! | Significant |

*Anscombe called this phenomenon, "sampling to reach a foregone conclusion."*

## Key Concepts - Bayesian Approach

- There is a Bayesian approach means not having to wait for a specified number of click throughs to be observed
- Bayes means using the available data as it comes in and making a prediction
- Anscombe described a formula by which a decision can be made, as the experiment proceeds, to whether or not to stop

## KEY CONCEPTS - CLASSICAL VS BAYESIAN/ANSCOMBE

- The two approaches have been widely debated in the context of clinical trials
- In clinical trials you may be providing an inferior treatment during the trial, so there is a significant cost to running the trial (in terms of regret)
- This must be balanced against what you will learn, and therefore can do to help *future* patients (maximizing performance)
- Anscombe developed the Bayesian approach in the 1960s and it is widely used in clinical trials today

## Key Concepts - Bayesian Approach

- Balance the cost of the test vs the cost of making the wrong decision
- i.e. Maximize performance, minimize regret
- Effectively have 2 parameters:

    1. The length of time in which you run the experiment
    2. The length of time you will serve up the winning result

The formula provides a way to determine the stopping point of an experiment.

$$|c_a - c_b| > \Phi^{-1}\left(\frac{n}{k + 2n}\right)\sqrt{(n)}$$

**$|c_a - c_b|$** = absolute difference between clicks for both headlines

**$\phi^{-1}$** = the quantile function of the standard normal

**n** = number of page views so far (length of time you run the experiment)

**k** = number of future views (length of time you will serve up the winner)

## Key Concepts - Stopping Criteria - The Anscombe Boundary

- Keep track of the number of clicks (through the headline)
- When the absolute value between the number of clicks arising from the 2 headlines crosses the Anscombe boundary the headline test is stopped

## Key Concepts - The Anscombe Boundary