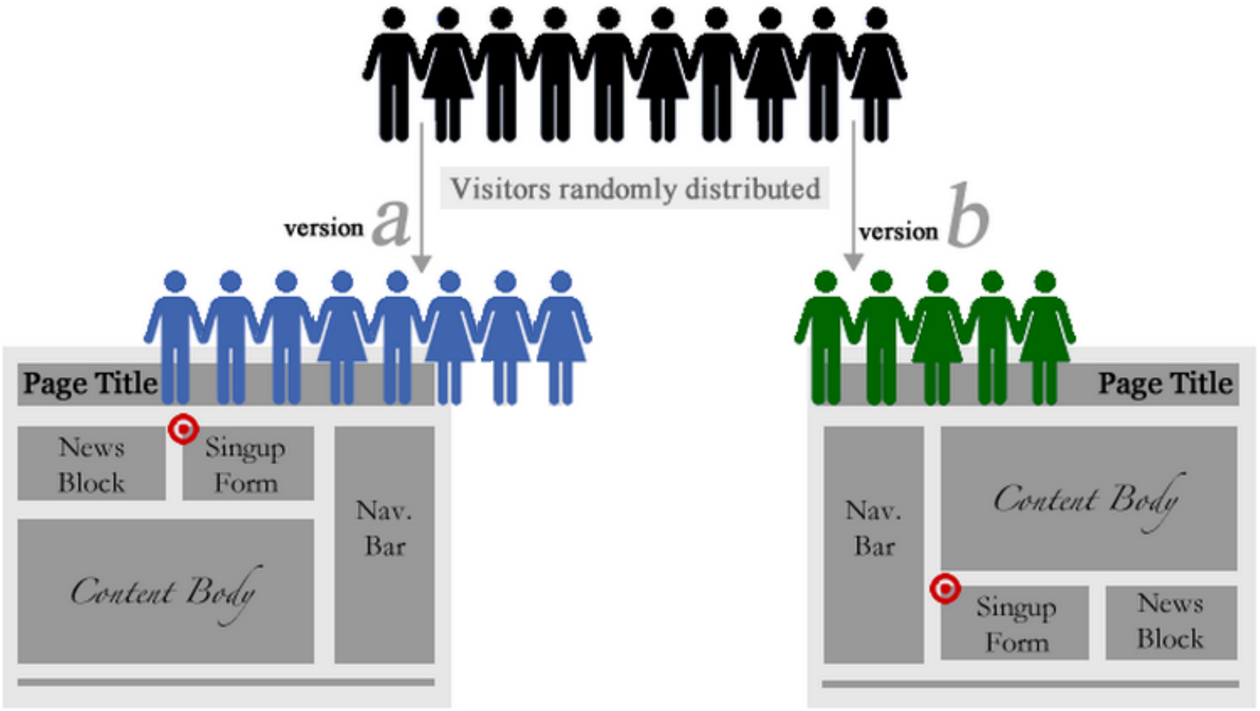
CS 791/591

“Building a Better E-Commerce Website”

Part I: Experimental Design

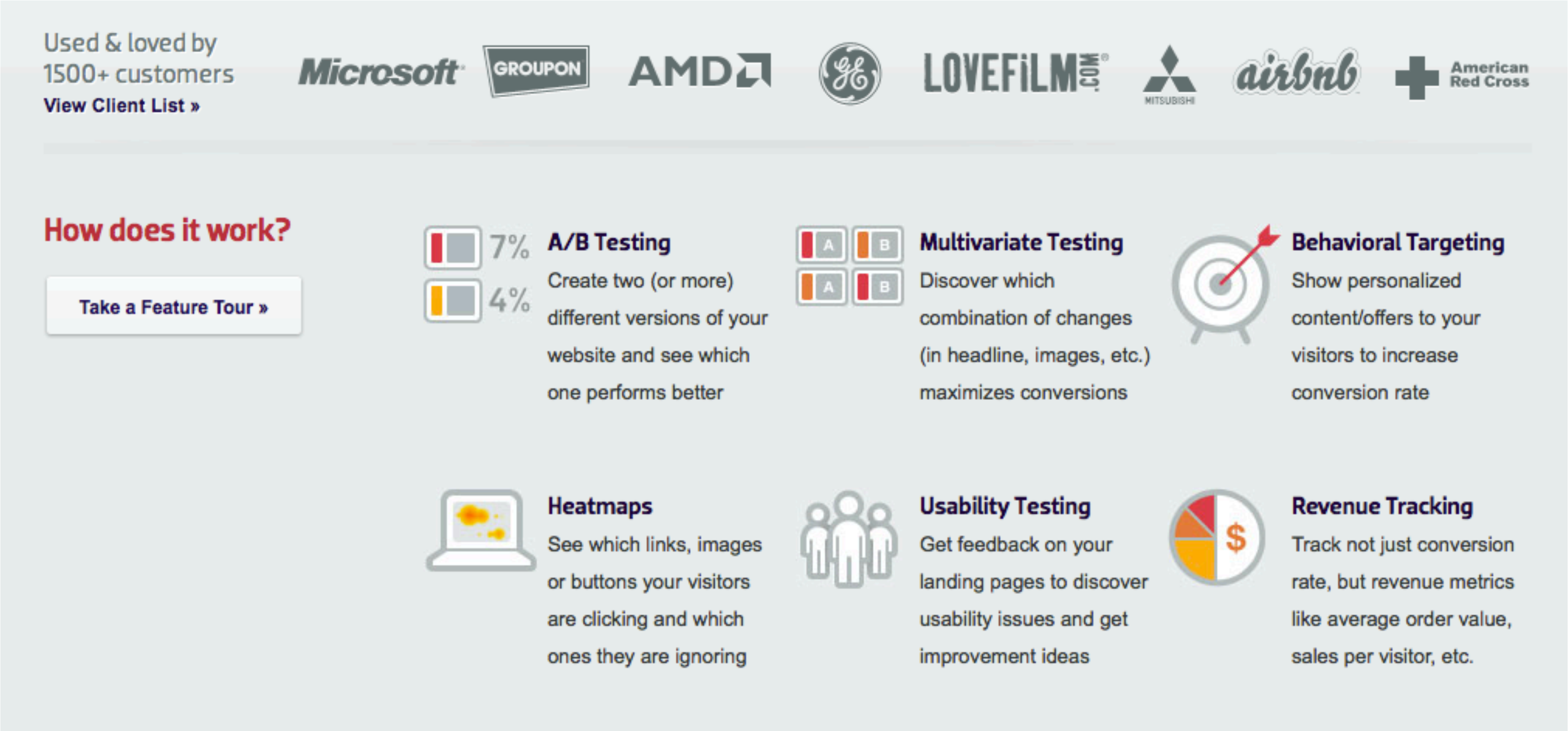
A/B Testing in a nutshell

- While keeping the OLD version online, route a percentage of visitors to the NEW. Bookkeep those users' transactions and compare with statistics from the OLD version



Experiment Design

- Use **A/B test** to test effectiveness (measured in number of users who make a transaction) of new design
- Randomize first-time IP addresses so that half visit the original site and half the modified (Not the best way), this should help normalize any geographical, socio-economical or gender bias.
- Record total number of visitors to original (A) site and number of transactions, and record the same for visitors to the "improved" (B) site. Record data for a span of one week or more to normalize fluctuating buying habits throughout the week.



Used & loved by
1500+ customers
View Client List »

How does it work?

A/B Testing

Create two (or more) different versions of your website and see which one performs better

Multivariate Testing

Discover which combination of changes (in headline, images, etc.) maximizes conversions

Behavioral Targeting

Show personalized content/offers to your visitors to increase conversion rate

Heatmaps

See which links, images or buttons your visitors are clicking and which ones they are ignoring

Usability Testing

Get feedback on your landing pages to discover usability issues and get improvement ideas

Revenue Tracking

Track not just conversion rate, but revenue metrics like average order value, sales per visitor, etc.

Metrics and Testing

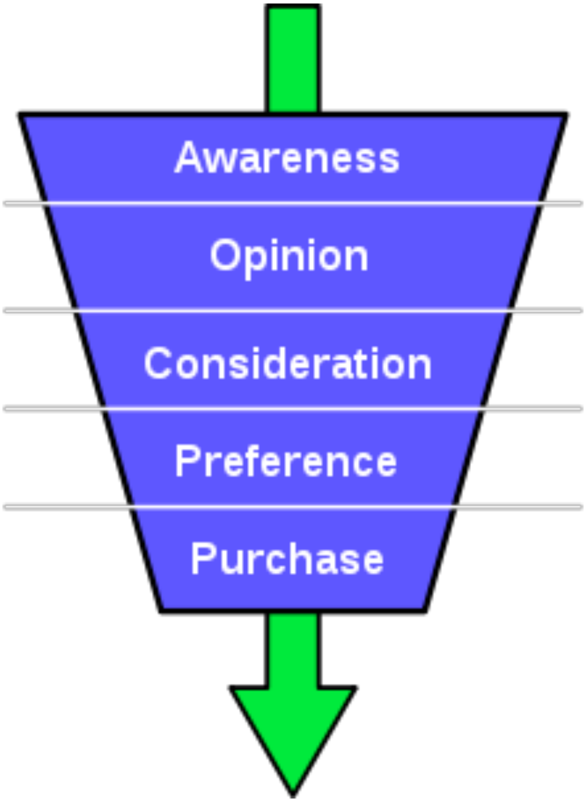
- Conversion rate
- Total revenue
- Other metrics?

- How big does the sample size need to be?

The experiment

- 50% of the live traffic is sent to the original website and 50% to the test website
- Measure the number of users that convert to costumers as: CR = $\frac{\#\ of\ conversions}{\#\ of\ visits}$
- Compare the CR obtained using the two different versions of the website
- Keep the version with better CR

Purchase Funnel



search

click-through (CTR)

page view

conversion

Behavior Metrics

- **Bounce Rate:** The percentage of visitors that enter and then leave the site without any additional interaction.
  - Examples of a “bounce”: closing a tab, timeout, clicking “Back.”
- **Pages Depth:** The number of unique pages viewed during a visit to the site.
  - Metric can be further specified to the **product depth**, the number of product-pages viewed during a visit.
  - Repeat pages not considered
- **Visit Duration:** Total time spent on site.
  - Not particularly interested in the time during checkout / processing.

Methods

- Significance testing of mean conversion rates
  - Users as Bernoulli random trials (0/1)
  - Z-scores
  - T-tests
  - Confidence intervals
- Significance testing of revenue streams
  - Users as real-valued (or integer) values
  - Do above methods all still apply?

Significance Testing

- With the collected data a “conversion rate” can be determined. Which is the percentage of users who made a transaction over all visitors
- A z-test can be done on the difference of the two proportions
- The null hypothesis would be that the proportion of buyers of the B site visitors is the same as the A site visitors.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - d_0}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- We can set the confidence interval to 0.05, so if the p value is less than it, we can say with a 95% certainty that the new design does improve sales

Image Source: http://www.milefoot.com/math/stat/ht-proportions.htm

A/B Testing and Two-Sample t-Testing

- Conduct A/B testing. We direct both old and new users to either site randomly with equal probability.
- $\mu_1 =$ average revenue per visitor generated by old site[1]
- $\mu_2 =$ average revenue per visitor generated by new site
- Null Hypothesis: $H_0 : \ \mu_1 = \mu_2$
- Alternative Hypotheses: $H_a^1 : \ \mu_1 > \mu_2$ and $H_a^2 : \ \mu_1 < \mu_2$
- $\hat{Y}_1$ and $\hat{Y}_2$ are the sample means of the old and new site average revenue per visitor
- N_1 and N_2 number of visitors of old site and new site
- s_1^2 and s_2^2 are the variances

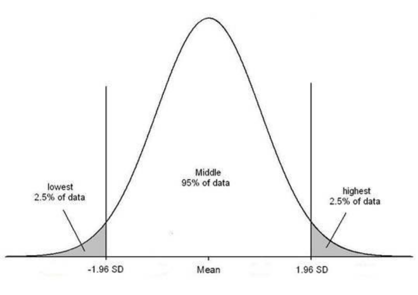
[1] http://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm - Two-Sampel t-Test reference

Testing

- Significance level $\alpha = 10\%, 5\%, 1\%$ depending on how sure we wish to be
- Test Statistic: $T = \frac{\hat{Y}^1 - \hat{Y}^2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$
- Reject H_0 and accept H_a^1 if $T > t_{1-\alpha,v}$
- Reject H_0 and accept H_a^2 if $T < t_{\alpha,v}$
- $t_{1-\alpha,v}$ and $t_{\alpha,v}$ are the critical values of the t distribution

The Solution (I)

- We can simply obtain $\hat{\mu}_1$ and $\hat{\mu}_2$
  - Can we trust these values?
    - Yes! If we use enough number of samples.
    - $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
    - If your intervals do not overlap, you are good to go!



Sample Size

	Cohen's d		
Power	**0.2**	**0.5**	**0.8**
0.25	84	14	6
0.50	193	32	13
0.80	393	64	26
0.90	526	85	34
0.95	651	105	42
0.99	920	148	58

	True H_0	**False H_0**
Fail to reject H_0	Correct $1-\alpha$	Type II error, β
Reject H_0	Type I error, α	Correct $1-\beta$

We determine the sample size *n* using a pre-determined table for a significance level of 0.05.

We calculate Cohen's

$$d = \frac{\bar{Y}_{exist} - \bar{Y}_{new}}{\sqrt{\frac{99}{200}\left(s^2_{exist} + s^2_{new}\right)}}$$

based on a sample of 100 first-time customers to the existing page layout and 100 first-time customers to the new design.

We select the desired statistical power, which is the probability of not committing a Type II error, $1-\beta$.

Note that the sample size varies directly with the power of the statistical test and inversely with Cohen's *d*.

Methods

- Significance testing of mean conversion rates
  - Users as Bernoulli random trials (0/1)
  - Z-scores
  - T-tests
  - Confidence intervals
- Significance testing of revenue streams
  - Users as real-valued (or integer) values
  - Do above methods all still apply?

Testing for Significant Results

- A statistical test for difference of means is performed on the two means of revenue per visitor (m_1 and m_2).
- Confidence Level is set at 99%. That is, we want to be 99% sure that m_1 and m_2 are different before concluding the experiment.
- Once both webpage versions have had 30 visitors, the following z-score is calculated after each additional visitor.

$$z = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The experiment ends when z > 2.575 or z < 2.575. At this time, we are 99% sure of the superior webpage version. The webpage version with higher mean revenue per visit is used permanently.

Interface comparison: short term revenue

- Apply a Student test to verify if there is a statistical significance between interface XT and XC chosen your confident value α.
- Base on those results argue about the improvement (or not) of the new interface on the revenues.

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

Assumptions & Limitations

- It’s possible that version A of the site might increase short term revenue, but also dramatically increase “one and done” shoppers, thereby decreasing profits in the long run. Maybe a weighted formula could be developed that takes into account short term revenue, ad revenue, repeat traffic, etc. (Adam U., Sanaz)
- We must consider the lifetime value of repeat customers (Gaurav).
- Maximizing revenue does not guarantee the greatest possible profit (Eugene, Zheng F.)
- “t-test assumption is that revenue generated from visitors follows a normal distribution” (Eric M.)

Experiment with Short-Term Revenue

- Question:
  - Which version of the website begets greater short-term revenue?

- Experimental Design:
  - Sample fresh users from both versions of the site
  - Take measurements of previously-defined factors
  - Use a formula that assigns weights to the metrics and sums to *one* short-term revenue metric:

$$STR = a_1 x_1^{e_1} + a_2 x_2^{e_2} + \cdots + a_n x_n^{e_n}$$

where:

$a_1 \ldots a_n \in [0,1]$ are weights for metrics $x_1 \ldots x_n$
with exponents $e_1 \ldots e_n \in \mathbb{R}$

Problem

Maximize the short-term profitability of a website by determining whether the new web design, which has just been developed, is more profitable than the old one.

Experimental Goal 1

Determine the preference of new users. That is, compute the probability that new users make a purchase when viewing the old site design, $P(d_1|u_\nu)$ and similarly when viewing the new site design $P(d_2|u_\nu)$.

Experimental Goal 2

Determine the preference of previous users. That is, compute the probability that users will make a purchase using the new design given that they have made a purchase using the old design, $P(d_2|u_o, d_1)$. For comparison compute $P(d_1|u_o, d_1)$.

Carrying out goal 1

To compute $P(d_1|u_\nu)$ and $P(d_2|u_\nu)$.

```
For an arbitrary user $u$:
    If $u$'s cookies indicate that $u$ is a new user:
        randomly and uniformly generate $n \in \{0, 1\}$
        If $n = 1$:
            send $u$ to the new website and record whether $u$ makes a purchase
        If $n = 0$:
            send $u$ to the old website and record whether $u$ makes a purchase
```

Carrying out goal 2

To compute $P(d_2|u_o, d_1)$.

```
    For an arbitrary user $u$:
        If $u$'s cookies indicate that $u$ is not a new user:
            send $u$ to the new website and record whether $u$ makes a purchase
```

To compute $P(d_1|u_o, d_1)$ use company records to obtain the number of users who made a second purchase after their first, $freq(p_1 \wedge p_2)$ and the users who have not made a second purchase after the first $freq(p_1 \wedge p_2)$ – only counting those that satisfy $t_{1+} > \mu_{t_2 - t_1}$ where t_{1+} is the elapsed time after the first purchase and $\mu_{t_2 - t_1}$ is the mean time between the first and second purchases.

maximizing decision rule

If the expected number of new users in the short term is x_ν and the number of current users is x_o, then the web design that maximizes short term profits is:

$$\arg\max_{d_1,d_2}\{x_\nu P(d_1|u_\nu) + x_o P(d_1|u_o,d_1),\, x_\nu P(d_2|u_\nu) + x_o P(d_2|u_o,d_1)\}$$

number of samples

Since we can view $P(d_1|u_o, d_1)$ and $P(d_2|u_o, d_1)$ as a binomial distribution over current users and $P(d_1|u_\nu)$ and $P(d_2|u_\nu)$ as a binomial distribution over new users we can use the Wilson score interval for both to determine the numbers of samples needed for the desired confidence level. The equation is fairly long so here is a link:

http://www.mwsug.org/proceedings/2008/pharma/MWSUG-2008-P08.pdf (see equation 7)

Assumptions

I made the following simplifying assumptions: (1) It doesn't matter how many times someone is a repeat customer. (2) Cookies will remain fairly constant, i.e. people won't be clearing their cookies so frequently that our sample of purportedly new customers, will, in fact, be new customers. Management wrongly assumed that the web design that maximizes short-term profits will also maximize long - term profits. For example, the new web design might allow for fast purchases without registration. This would presumably have the effect of increasing short-term profits but also decrease long - term profits and the number of repeat customers as it would lack any sort of recommendation system based on previous purchases.