MTAT.03.319

Business Data Analytics

Lecture 11

The slides are available under creative common license. The original owner of these slides is the University of Tartu.
Marketing and Sales

A/B Testing
Which ad is better?

First: ?
Second: ?
Third: ?
Which ad is better?

A/B Testing in a nutshell

website variation I

website variation II
A/B Testing in a nutshell

The conversion rate is the percentage of users who take a desired action (eg: bought an item etc.)

Source: vwo.com
Call-To-Action (CTA)

Link to encourage visitors to take action. This is normally the next step (e.g. add to basket on a product page) in your conversion journey and is the link between your default content and a page with a more high value offer on it.

Source: https://www.conversion-uplift.co.uk/glossary-of-conversion-marketing/call-to-action-cta/
Why we need A/B Testing

even small changes can result in significant increases (or decreases) in leads generated, sales and revenue

THE TEST

We used Google’s new split testing tool, Optimize 360, to test the impact of changing our regular buttons on our homepage to ghost buttons. The screenshots below show what was tested.

![We work with some amazing brands](https://conversionxl.com/blog/ghost-buttons/)

Original design with a pink button.

![We work with some amazing brands](https://conversionxl.com/blog/ghost-buttons/)

Variation with a ghost button.

We ran this test over the course of 10,000+ visits to the homepage and measured clicks on any of the CTAs on the homepage.

THE RESULTS

We reached statistical significance for the test, which saw a 20% decrease in clicks on the buttons in the ghost variation.

Source: https://conversionxl.com/blog/ghost-buttons/
“A/B testing helps us gain confidence in the change we’re making. It helps us validate new ideas and guides decision making. Without A/B testing, we’re leaving much of what we do up to chance.”
Customer lifecycle

- Acquisition
- Growth
- Retention
- Winback

Conversion rate
A/B Testing process

1. Study your data. Identify problem areas or areas with the potential improvement

   The homepage has a high bounce rate

2. Use various descriptive tools to form hypotheses

   The homepage does not have an action button on the lending page

3. Construct the hypothesis

   Placing the button on the landing page will increase conversion rates
A/B Testing process

4. Construct A/B test: use control and focus groups. Calculate sample sizes, current conversion rate and the power of the test.

A: a page with the button, B: original page

define the goal of conversion

Source: vwo.com
Terminologies

A focus group is a small, but demographically diverse group of people and whose reactions are studied especially in market research or political analysis in guided or open discussions about a new product or something else to determine the reactions that can be expected from a larger population.

Source: https://en.wikipedia.org/wiki/Focus_group

The control group is composed of participants who do not receive the experimental treatment. When conducting an experiment, these people are randomly selected to be in this group. They also closely resemble the participants who are in the experimental group or the individuals who receive the treatment.

Source: https://www.verywellmind.com/what-is-the-control-group-2794977
Terminologies

The **power** of a **hypothesis test** is the probability of making the correct decision if the alternative **hypothesis** is true. That is, the **power** of a **hypothesis test** is the probability of rejecting the null **hypothesis** $H_0$ when the alternative **hypothesis** $H_A$ is the **hypothesis** that is true.

Source: https://www.verywellmind.com/what-is-the-control-group-2794977
A/B Testing process

How did our ad perform on your network? Great! Your awesomeness metric is up 22%. Your cool factor metric is up 36%. Your smiley-face metric is up 58%. And A.M.W.T.M.U. is up a record 85.4%. What's A.M.W.T.M.U.? Another metric we totally made up.
A/B Testing process

4. **Construct A/B test**: use control and focus groups. Calculate sample sizes, current conversion rate and the power of the test

A: a page with the button, B: original page

5. Perform the test and draw conclusions. If the results inconclusive, go back to 3. and rework your hypothesis.
A/B Testing process

1. Study your data. Identify problem areas or areas with the potential improvement
   The homepage has a high bounce rate

2. Use various descriptive tools to form hypotheses
   The homepage does not have an action button on the lending page

3. Construct the hypothesis
   Placing the button on the landing page will increase conversion rates
Sleeping habits

4 hours of sleep

8 hours of sleep

exam performance
Hypothesis testing

H0: People who sleep 4 hours at night perform the same at exam than those who sleep 8

H1: 4 hours sleepers perform different at exam than those who sleep 8
Hypothesis testing

H0: People who sleep 4 hours at night perform the same at exam than those who sleep 8

H1: 4 hours sleepers perform different at exam than those who sleep 8
Hypothesis testing

H0: People who sleep 4 hours at night perform the same at exam than those who sleep 8

H1: 4 hours sleepers perform different at exam than those who sleep 8

H0: mean(score\text{group1}) = mean(score\text{group2})

H1: mean(score\text{group1}) \neq mean(score\text{group2})
Imagine you want to see the difference between two versions of the website - the current version and the one with the different color scheme. Construct the hypothesis to test this.
Hypothesis testing

Collect the data
Hypothesis testing

Collect the data

When enough is enough?
A/B Testing process

4. Construct **A/B test**: use control and focus groups. Calculate sample sizes, current conversion rate and the power of the test.

- how sure we need to be that we are measuring a real change
- how big the change we expect to see because of the new version, compared to the baseline?

Source: vwo.com

5. Perform the test and draw conclusions. If the results inconclusive, go back to 3. and rework your hypothesis.
how sure we need to be that we are measuring a real change?

- What percentage of the time are we willing to miss a real effect? 
  (power of the test)

- What percentage of the time are we willing to see an effect by random chance? 
  (significance level, or the probability of rejecting the null hypothesis)
how sure we need to be that we are measuring a real change?

- What percentage of the time are we willing to miss a real effect?
  (power of the test)
  typical value: 80%

- What percentage of the time are we willing to see an effect by random chance?
  typical value: 5%
  (significance level, or the probability of rejecting the null hypothesis)
Type I error
(false positive)

You're pregnant

Type II error
(false negative)

You're not pregnant
how big is the change we expect to see?

effect size - expected difference between groups

depends on the business context
How do we measure differences? 

How are sample size, effect size, false positive, and false negative rates related? 

The power of a test \( (P, 1 - \beta) \) is the probability that a test will detect an effect, if an effect is really there. When your power is high, your false negative rate is low. 

The significance level of a test \( (\alpha) \) is the probability that a test will detect an effect, if an effect is really not there. When your significance level is low, your false positive rate is low. 

We would like to not be fooled too often by either false negatives or false positives, so we choose large enough sample sizes for the effect size we expect to see. 

Move the sliders to explore the relationships 

Power threshold 

Significance level 

Baseline conversion rate 

With those parameters, you can measure... 

A relative % change of | With a sample size in each group of 
---|---
40.7% | 1,000 
28.2% | 2,000 
22.8% | 3,000 
19.6% | 4,000 

The sample sizes here are per variation (A and B in an A/B test), not the test as a whole.

https://juliasilge.shinyapps.io/power-app/
How do we measure differences?

How are sample size, effect size, false positive, and false negative rates related?

The power of a test ($P$, $1 - \beta$) is the probability that a test will detect an effect, if an effect is really there. When your power is high, your false negative rate is low.

The significance level of a test ($\alpha$) is the probability that a test will detect an effect, if an effect is really not there. When your significance level is low, your false positive rate is low.

We would like to not be fooled too often by either false negatives or false positives, so we choose large enough sample sizes for the effect size we expect to see.

Move the sliders to explore the relationships

Power threshold

Significance level

Baseline conversion rate

The sample sizes here are per variation (A and B in an A/B test), not the test as a whole.

With those parameters, you can measure...

<table>
<thead>
<tr>
<th>A relative % change of</th>
<th>With a sample size in each group of</th>
</tr>
</thead>
<tbody>
<tr>
<td>40.7%</td>
<td>1,000</td>
</tr>
<tr>
<td>28.2%</td>
<td>2,000</td>
</tr>
<tr>
<td>22.8%</td>
<td>3,000</td>
</tr>
<tr>
<td>19.6%</td>
<td>4,000</td>
</tr>
</tbody>
</table>

https://juliasilge.shinyapps.io/power-app/
quantity before and after the treatment, analyzing the data using a paired t-test. Let \( \mu_D \) and \( \mu_D' \) denote the pre-treatment and post-treatment measures on subject \( i \) respectively. The possible effect of the treatment should be visible in the differences \( D_i = B_i - A_i \), which are assumed to be independently distributed, all with the same expected value and variance.

The effect of the treatment is estimated by comparing the mean of the differences to 0. A test is used to determine whether the mean difference is statistically significant.

\[
T_n = \frac{D_n - \bar{D}}{\sigma_D / \sqrt{n}}
\]

where

\[
D_n = \frac{1}{n} \sum_{i=1}^{n} (B_i - A_i)
\]

\( n \) is the sample size.

If \( n \) is large, one can approximate the t-distribution by a normal distribution and calculate the critical value using the quantile function \( \Phi^{-1} \), the inverse of the cumulative distribution function of the normal distribution. It turns out that the null hypothesis will be rejected if

\[
T_n > 1.64.
\]

Now suppose that the alternative hypothesis is true and \( \mu_D = \theta \). Then, the power is

\[
B(\theta) = \Pr (T_n > 1.64 | \mu_D = \theta)
\]

\[
= \Pr \left( \frac{D_n - \theta}{\sigma_D / \sqrt{n}} > 1.64 | \mu_D = \theta \right)
\]

\[
= \Pr \left( \frac{D_n - \theta + \theta}{\sigma_D / \sqrt{n}} > 1.64 | \mu_D = \theta \right)
\]

\[
= \Pr \left( \frac{D_n - \theta}{\sigma_D / \sqrt{n}} > 1.64 - \frac{\theta}{\sigma_D / \sqrt{n}} | \mu_D = \theta \right)
\]

\[
= 1 - \Pr \left( \frac{D_n - \theta}{\sigma_D / \sqrt{n}} < 1.64 - \frac{\theta}{\sigma_D / \sqrt{n}} | \mu_D = \theta \right)
\]

For large \( n \), \( \frac{D_n - \theta}{\sigma_D / \sqrt{n}} \) approximately follows a standard normal distribution when the alternative hypothesis is true, the approximate power can be calculated as

\[
B(\theta) \approx 1 - \Phi \left( 1.64 - \frac{\theta}{\sigma_D / \sqrt{n}} \right).
\]

According to this formula, the power increases with the values of the parameter \( \theta \). For a specific value of \( \theta \) a higher power may be obtained by increasing the sample size \( n \).

It is not possible to guarantee a sufficient large power for all values of \( \theta \). As \( \theta \) may be very close to 0. The minimum (infimum) value of the power is equal to the size of the test, \( \alpha \), in this example 0.05. However, it is of no importance to distinguish between \( \theta = 0 \) and small positive values. If it is desirable to have enough power, say at least 0.90, to detect values of \( \theta > 1 \), the required sample size can be calculated approximately:

\[
B(1) \approx 1 - \Phi \left( 1.64 - \frac{\sqrt{n}}{\sigma_D} \right) > 0.90,
\]

from which it follows that

\[
\Phi \left( 1.64 - \frac{\sqrt{n}}{\sigma_D} \right) < 0.10.
\]

Hence, using the quantile function,

\[
\frac{\sqrt{n}}{\sigma_D} > 1.64 - z_{\alpha/2} = 1.64 + 1.28 = 2.92 \quad \text{or} \quad n > 8.56\sigma_D^2,
\]

Well...it is a bit complicated. Check this: https://en.wikipedia.org/wiki/Statistical_power
quantity before and after the treatment, analyzing the data using a paired t-test. Let \( A_i \) and \( B_i \) denote the pre-treatment and post-treatment measures on subject \( i \) respectively. The possible effect of the treatment should be visible in the differences \( D_i = B_i - A_i \), which are assumed to be independently distributed, all with the same expected value and variance.

The effect of the treatment can be analyzed using a one-sided t-test. The null hypothesis of no effect will be that the mean difference will be zero, i.e., \( H_0 : \mu_D = 0 \). In this case, the alternative hypothesis states a positive effect, corresponding to \( H_1 : \mu_D > 0 \). The test statistic is:

\[
T_n = \frac{\bar{D}_n - 0}{\hat{\sigma}_D / \sqrt{n}},
\]

where

\[
\bar{D}_n = \frac{1}{n} \sum_{i=1}^{n} D_i,
\]

\( n \) is the sample size and \( \hat{\sigma}_D / \sqrt{n} \) is the standard error. The test statistic under the null hypothesis follows a Student t-distribution. Furthermore, assume that the null hypothesis will be rejected at the significance level of \( \alpha = 0.05 \). Since \( n \) is large, one can approximate the t-distribution by a normal distribution and calculate the critical value using the quantile function \( \Phi^{-1} \), the inverse of the cumulative distribution function of the normal distribution. It turns out that the null hypothesis will be rejected if

\[ T_n > 1.64. \]

Now suppose that the alternative hypothesis is true and \( \mu_D = \theta \). Then, the power is

\[
B(\theta) = \Pr (T_n > 1.64 \mid \mu_D = \theta)
\]

\[
= \Pr \left( \frac{\bar{D}_n - 0}{\hat{\sigma}_D / \sqrt{n}} > 1.64 \mid \mu_D = \theta \right)
\]

\[
= \Pr \left( \frac{\bar{D}_n - \theta}{\hat{\sigma}_D / \sqrt{n}} + \theta > 1.64 - \frac{\theta}{\hat{\sigma}_D / \sqrt{n}} \mid \mu_D = \theta \right)
\]

\[
= 1 - \Pr \left( \frac{\bar{D}_n - \theta}{\hat{\sigma}_D / \sqrt{n}} < 1.64 - \frac{\theta}{\hat{\sigma}_D / \sqrt{n}} \mid \mu_D = \theta \right)
\]

For large \( n \), \( \frac{\bar{D}_n - \theta}{\hat{\sigma}_D / \sqrt{n}} \) approximately follows a standard normal distribution when the alternative hypothesis is true, the approximate power can be calculated as

\[
B(\theta) \approx 1 - \Phi \left( 1.64 - \frac{\theta}{\hat{\sigma}_D / \sqrt{n}} \right).
\]

From which it follows that

\[
\Phi \left( 1.64 - \frac{\sqrt{n}}{\hat{\sigma}_D} \right) < 0.10.
\]

Hence, using the quantile function,

\[
\frac{\sqrt{n}}{\hat{\sigma}_D} > 1.64 - \phi_{0.10} = 1.64 + 1.28 \approx 2.92 \quad \text{or} \quad n > 8.56 \hat{\sigma}_D^2,
\]

Or use calculators:

http://www.evanmiller.org/ab-testing/sample-size.html
Hypothesis testing

H0: People who sleep 4 hours at night perform the same at exam than those who sleep 8

H1: 4 hours sleepers perform different at exam than those who sleep 8

H0: $\text{mean} (\text{score}_{\text{group1}}) = \text{mean} (\text{score}_{\text{group2}})$
H1: $\text{mean} (\text{score}_{\text{group1}}) \neq \text{mean} (\text{score}_{\text{group2}})$
Hypothesis testing

- 24 participants
- randomly divided into one of each group
- after, the same exam is taken and results recorded
Statistical hypothesis testing is a general framework of testing claims.
Statistical hypothesis testing is a general framework of testing claims.

1. state the relevant null and alternative hypotheses

2. decide which test is appropriate for this type of hypothesis

- mean comparison - &gt; t-test
- two sample means -&gt; two-sample t-test
- randomized data, not repeated measurements - &gt; unpaired
- two-sided or one-sided? (≠ vs. <,>)
- equal variances?
Statistical hypothesis testing is a general framework of testing claims.

1. state the relevant null and alternative hypotheses
2. decide which test is appropriate for this type of hypothesis
3. check the statistical assumptions of the test

Normal distribution of both groups?
Statistical hypothesis testing is a general framework of testing claims.

1. state the relevant null and alternative hypotheses
2. decide which test is appropriate for this type of hypothesis
3. check the statistical assumptions of the test
4. state the relevant test statistic $T$

For two-sample, unpaired, with unequal variances:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
Statistical hypothesis testing is a general framework of testing claims.

1. state the relevant null and alternative hypotheses
2. decide which test is appropriate for this type of hypothesis
3. check the statistical assumptions of the test
4. state the relevant test statistic $T$
5. select a significance level $\alpha$ (1% or 5%)

the probability of accepting the H1 hypothesis given that it is false (a type I error)
Statistical hypothesis testing is a general framework of testing claims.

1. state the relevant null and alternative hypotheses
2. decide which test is appropriate for this type of hypothesis
3. check the statistical assumptions of the test
4. state the relevant test statistic $T$
5. select a significance level $\alpha$ (1% or 5%)
6. compute from the observations $t_{\text{obs}}$ of $T$
7. calculate p-value (probability of seeing the same result or more extreme under $H_0$)
8. reject $H_0$ if the p-value is less than $\alpha$
6. compute from the observations $t_{obs}$ of $T$

7. calculate p-value (probability of seeing the same result or more extreme under H0)

8. reject H0 if the p-value is less than $\alpha$

```r
> t.test(group1, group2, alternative='two.sided', paired=FALSE, var.equal=FALSE)

Welch Two Sample t-test

data:  group1 and group2
t = -5.1535, df = 17.718, p-value = 7e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.6765985  0.7047107
sample estimates:
mean of x  mean of y
  6.669761  7.860416
```
Statistical hypothesis testing is a general framework of testing claims.

6. compute from the observations $t_{\text{obs}}$ of T

7. calculate p-value (probability of seeing the same result or more extreme under H0)

8. reject H0 if the p-value is less than a significance level

Do we accept or reject H1?

What can we claim?
Why do we need testing for significance

Statistical significance means highly likely that the differences are real, repeatable, and not due to random chance.
Why do we need testing for significance

Statistical significance means highly likely that the differences are real, repeatable, and not due to random chance.

Infinite monkey theorem

From Wikipedia, the free encyclopedia

The infinite monkey theorem states that a monkey hitting keys at random on a typewriter keyboard for an infinite amount of time will almost surely type a given text, such as the complete works of William Shakespeare.
Common tests

<table>
<thead>
<tr>
<th>Assumed Distribution</th>
<th>Example Case</th>
<th>Standard Test</th>
<th>Alternative Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>Average Revenue Per Paying User</td>
<td>Welch's t-test (Unpaired t-test)</td>
<td>Student's t-test</td>
</tr>
<tr>
<td>Binomial</td>
<td>Click Through Rate</td>
<td>Fisher's exact test</td>
<td>Barnard's test</td>
</tr>
<tr>
<td>Poisson</td>
<td>Transactions Per Paying User</td>
<td>E-test[^7]</td>
<td>C-test</td>
</tr>
<tr>
<td>Multinomial</td>
<td>Number of each product Purchased</td>
<td>Chi-squared test</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>--</td>
<td>Mann–Whitney U test</td>
<td>Gibbs sampling</td>
</tr>
</tbody>
</table>

Source: wikipedia
How to select Tests

<table>
<thead>
<tr>
<th>Select the appropriate Test</th>
<th>Output (Dependant Variable) Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Continuous</td>
</tr>
<tr>
<td>Continuous (Input Independent Variable) X’s</td>
<td>Simple Linear Regression</td>
</tr>
<tr>
<td></td>
<td>Correlation</td>
</tr>
<tr>
<td>Discrete</td>
<td>Normal Data</td>
</tr>
<tr>
<td></td>
<td>T Test (1,2 Sample &amp; Paired)</td>
</tr>
<tr>
<td></td>
<td>ANOVA, F Test, HOV</td>
</tr>
<tr>
<td></td>
<td>Non-Normal Data</td>
</tr>
<tr>
<td></td>
<td>Moods Median, HOV</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: https://www.sixsigma-institute.org/Six_Sigma_DMAIC_Process_Analyze_Phase_Hypothesis_Testing.php
https://conversionxl.com/blog/ab-testing-guide/
Demo time!

https://courses.cs.ut.ee/2018/bda/spring/Main/Practice